Genetic Control of Mosquitoes

A modular approach to improving transgene design for complex CRISPR population control strategies

Thesis submitted in accordance with the requirements of the
Liverpool School of Tropical Medicine for the degree of Doctor in Philosophy
by Jessica Colleen Purcell

30th November 2021

Supervised by Prof. Luke Alphey & Dr. Gareth Lycett

of

The Pirbright Institute and Liverpool School of Tropical Medicine

For the Dr. Purcell that came before me,
and for Riley,
who will be whatever kind of Purcell she wants to be.

This has been a hard road and I have learnt a lot; my thanks to the following, without whom I would not have succeeded:

Adrian Zagrajek, Adriana Diaz, Alex Ballinger, Alex Jones, Ambi Batra, Amelia Collins, Amy-lee Levey, Angela Durin, Anne and Keith Johnson, Anthony Wilson, Anto Rawlinson, Anya Green, Barbara Holzer, Barry Atkinson, Becca McLean, Becca Philip, Becca Ireland, Becca Heywood, Becks Cadwallader, Ben Hu, Caitlin Rea, Christina Cosma, Claire Purcell, Dana Perry, David Navarro, Deborah Greer, Deepak Purusothaman, Dr. Eastwood, Ed Poore, Ela Krzywinska, Elise McDonald, Ellen Lockheart, Emma Howson, Emma Dunn, Emma Bartram, Estela Gonzalez, Fran Low, Fran Evans, Gabriella and Robin Poore, Gareth Shimmon, Gemma Watson, Genevive Masters, Gigi Fateh, Grace Logan, Hannah Walker, Harriet Stevenson, Heather Offord, Holly Turner, Ian Wrightson, Ilona Flis, Issy Lewis, Izi Hutchison, James Henderson, Jamie Casswell, Jarek Krzywinski, Jen Burry, Jenny Eyre, Jessica Swanson, Jessica Mavica, Jessica Stokes, Jessica Allen, Jessica Powell, Jessica Clay, Jill Tombs, Jim Barber, Jo Stoner, Joe Oddy, Jonnie Clowes, Josh Ang, Julia Smith, Julie McIvor, Kat Nevard, Kat Benjamin, Kate Dulwich, Kate Lennon, Kate Seabourne, Kathy Hoffmeyer, Katie Potter, Kaye Goulding, Kelly Keenan, Kim Stirk, Kirsty Poore, Kiya Human, Lara Harrup, Laura Dunn, Laura Tugwell, Lesley Bell-Sakyi, Lewis Shackleford, Liv O'Malley, Liv Stanford, Lizzie Leith, Lottie Leith, Lucy Balicki, Luke Alphey, Luke Johnson, Lynda Moore, Lyra Poore, Maira Bana, Maria Parkes, Marion's Marvellous Midges, Martha Watson, Martin Donnelly, Matt Edmans, Matt Edgington, Matt Shannon, Michelle Anderson, Mike Pun, Milly Powers, Nathan Williams, Naz Thakur, Nicki Mollison, Nicki Pollock, Nirja Joshi, Nora Kettleborough, Paul Foster, Phil Leftwich, Pippa Hatchett, Priscilla Tng, Rachel Drummond, Rachel Firth, Radiographer, Rennos Fragkoudis, Richard Bourne, Rikki Lovett, Rob Carey, Ryan Overton, Sam Stokes, Sanjay Basu, Sarah Deakin, Sarah Dick, Saskia Bennett, Sebald Verkuijl, Simon Carpenter, Sophia Fochler, Stacey Human, Tali Poore, Tammy Husseini, Teddy Middlebrook, Tim Harvey-Samuel, Tom Nicholson, Victoria Topham, Victoria Sy, William Tuohy.

With greatest thanks to my family, who have supported me throughout this journey.

Abstract

Genetic Control of Mosquitoes – A modular approach to improving transgene design for complex CRISPR population control strategies

J.C. Purcell, Liverpool School of Tropical Medicine, 2021

Mosquito borne diseases present an ongoing public health burden in many parts of the world. There are several, successful strategies for reducing disease burden (often by targeting the vector population) but there remains a need for additional strategies that can improve efficacy, reduce cost and reduce undesired side effects (such as the ecological impact of spraying chemical insecticides). One such group of strategies are those using genetic modification of the mosquito to control the wild vector population and reduce disease transmission. Such strategies have been increasing in complexity in recent years with the wide availability of programmable gene editing through the CRISPR/Cas system. Even as genetic control systems are developed and tested, there remains a need for optimisation in the design and function of different elements within the transgenes they are based on.

This thesis presents practical tools for the field of mosquito transgenics, particularly in the design and implementation of complex, CRISPR gene drive strategies. Using cell culture as a model, this work describes the validation of several, specific components of transgene design. The cell culture format was exploited to test a large number of variants of each component, more than would be practical *in vivo*. These findings are presented as an empirical resource for aid in design of mosquito genetic modification constructs.

Alternative translation initiation sequences (TIS) were characterised as a mode of modulating expression efficiency of transgenic proteins such as toxic effectors or fluorescent markers. TIS and 3'UTR sequences were identified that can be used independently or in tandem to induce 2 – 20 fold increases in transgene expression.

A CRISPRa assay was then developed and used in a pipeline for identification and validation of alternative U6 and 7SK RNA polymerase III promoter sequences to drive guide RNA expression in mosquito species of interest. Availability of multiple, sequence-divergent promoters increases capacity to express multiple guide RNAs within a single transgenic individual, a requisite for may complex CRISPR/Cas vector control strategies. This work expands on the state of the field through examining the use of exogenous promoters from closely related species. It was then demonstrated that truncations of the U6 promoters retain considerable activity, offering a route to reduce transgene cassette size.

Table of Contents

List of Abbreviations

Individuals

JP	Jessica C. Purcell
MA	Michelle A. E. Anderson
PL	Philip T. Leftwich
SB	Sanjay Basu
SV	Sebald A. N. Verkuijl
THS	Tim Harvey-Samuel
VN	Victoria C. Norman

Abbreviations

°C	Degrees centigrade
µl	Microlitre
3'UTR	3' untranslated region
A	Adenosine
AcNPV	Autographa californica nuclear polyhedrosis virus
adh	*alcohol dehydrogenase* gene, D. melanogaster
Ae.	*Aedes*
AGG#	Unique identifier for plasmids
ALU	Arbitrary light units
An.	*Anopheles*
Approx.	Approximate
Arbovirus(es)	Arthropod-borne virus(es)
B.	*Bombyx*
BLAST	Basic local alignment search tool
BLASTn	BLAST, nucleotide
bp	Base pair (of DNA or RNA)
C	Cytosine
C.	*Culex*
Cas	CRISPR associated protein(s)
Cas9	CRISPR associated protein 9
CDC	Centre for Disease Control, USA
CDS	Coding sequence
CI	Confidence interval
cm	Centimetre
CMV	Cytomegalovirus
CO_2	Carbon dioxide
CRISPR	Clustered regularly interspaced short palindromic repeats
CRISPRa	CRISPR activation
crRNA	CRISPR RNA
D.	*Drosophila*
dCas9	(enzymatically) deactivated Cas9
dCas9-VPR	Conjugation of dCas9 and VPR
DEPC	Diethyl pyrocarbonate
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dsDNA	Double stranded DNA
DSE	Distal sequence element
e.g.	Exempli gratia
ECDC	European Centre for Disease Control
EDTA	Ethylenediaminetetraacetic acid
EMCA	European Mosquito Control Association, WHO
EoNPV	Ectropis obliqua nucleopolyhedrovirus
FF	Firefly luciferase
FF/RL	Ratio of FF to RL, calculated by FF ÷ RL
g	Gravity
G	Guanine

gDNA	Genomic DNA
gel	Agarose gel
(e)GFP	(enhanced) Green fluorescent protein
GM	Genetically modified
hrs	Hours
i.e.	Id est
iv	*in vitro* transcribed
K10	Female sterile (1) K10, D. melanogaster
Kb(p)	Kilo base (pair)
L-15	Leibovitz's L-15 media
LA#	Unique identifier for primers/oligos
LB	Lysogeny broth
M.	*Manduca*
MCS	Multiple cloning site
min	Minutes
min(Hsp70)	Minimal (promoter)
miRNA	micro RNA
ml	Millilitre
MNPV	Multicapsid nuclear polyhedrosis virus
mRNA	Messenger RNA
N	Number of samples
ncRNA	Non-coding RNA
NEB	New England Biosciences, UK
ng	Nanogram
ns	Not significant
nt	Nucleotide(s)
oligo(s)	Oligonucleotide(s)
P	P value
P.	*Plutella*
P65	Nuclear factor κB 65kDa subunit
PAM	Protospacer adjacent motif
pBac	PiggyBac transposon
PBS^-	Phosphate buffered saline, ion-free
PCR	Polymerase chain reaction
Pen/strep	Penicillin-Streptomycin
PLB	Passive lysis buffer
pol	Polymerase
PSE	Proximal sequence element
qPCR	Quantitative PCR
RL	Renilla luciferase
RNA	Ribonucleic acid
RO	Reverse osmosis
Rta	Epstein-Barr virus R trans-activator
s	Second(s)
S.	*Spodoptera*
Schneider's	Schneider's Drosophila medium
SD	Standard deviation
Serum	Fetal bovine serum
sgRNA	Single guide RNA
SIT	Sterile insect technique
snRNA	Small nuclear RNA
sp.	Species (plural)
T	Thymine
TAE	Tris-acetate-EDTA
Temp.	Temperature
Tet	Tetracycline
TetO	Tetracycline operator
TetR	Tetracycline repressor protein
TIS	Translation initiation sequence
TMV	Tobacco mosaic virus
TPB	Tryptose phosphate broth
tracrRNA	Tracer RNA
TRE	Tet response element

tRNA	Transfer RNA
tTA	Tetracycline controlled trans-activator
tTAV3	Conjugate tetracycline trans activator – VP16
U	Units
µg	Microgram
UK	United Kingdom
µl	Microlitre
US$	United States Dollar
USA	United States of America
VP16	Herpes simplex viral protein 16
VP64	Tetrameric repeat adaptation of the herpes simplex virus VP16
VPR	Conjugation of VP64-P65-Rta
WHO	World Health Organisation
WT	Wild type
ZsG	ZsGreen fluorescent protein

Chapter 1: General Introduction

Pest and nuisance insects

Insect species (or local populations of a species in a given area) can be considered a 'pest' or 'nuisance' for health or economic reasons, though these are often intertwined. A clear example of an economic pest could be biting midges (*Culicoides sp.*) that do not typically cause health issues to humans, but do disrupt outdoor labour and tourism (Logan et al., 2009). An economic pest that overlaps with areas of health impact could include agricultural pests such as the Diamond back moth (*Plutella xylostella*), whose larvae eat brassica plants and are thought to have amounted at least US$1 billion in damages since 1993 (Zalucki et al., 2012). Economic losses, in particular agricultural losses, can contribute to negative health outcomes in less economically stable demographics or nations. The converse is also true, where poor health is a major contributor to the maintenance of poverty (World_Bank, 2014). Insect species responsible for animal injury and illness can include those that directly cause injury through biting, hematophagy and carnivorous activity. This mode of damage predominantly affects non-human animals. For juveniles, high densities of insect can cause morbidity and mortality through biting and hematophagy (consumption of blood) (e.g. fleas (*Ctenocephalides sp.*) (Traversa, 2013). Hematophagy is less dangerous for adult animals, but severe morbidity can occur as a result of carnivorous insect activity. A predominant example of this is the New World screw-worm fly (*Cochliomyia hominivorax*), which lays eggs in the flesh of a living animal and whose subsequent larvae consume that flesh (Spickler, 2016).

Serious morbidity and mortality are more frequently caused by insects through their activity as vectors of other pathogens. All manner of pathogens can be transmitted by insects and infection can be through mechanical or biological routes. In mechanical transmission, the tissue damage of the insect bite presents an opportunity for infection with pathogens present on the skin of the animal or on the insect (Foil and Gorham, 2000); there is growing evidence that transmission can also occur where feeding activity includes regurgitation – for example transmission of *Chlamydia trachomatis* by flies (*Musca sp.*) (Brewer et al., 2021). Biological transmission instead includes infection of the insect itself, with part of the pathogen's life cycle taking place in the insect. Bacteria, viruses and other parasites (uni- and multi-cellular) can be transmitted through this route and account for a global disease burden of animals and humans (reviewed by (Sciences, 2016, WHO, 2017, WHO, 2021a, Rozendaal, 1997)). Although pathogens can be transmitted this way by a variety of flies too, mosquitoes are

recognised as a major vector for diseases, including Dengue fever and Malaria (WHO, 2020, Bhatt et al., 2013), and are the focus of this body of work.

Mosquitoes

Mosquitoes (*Culicidae*) are a family of small, dipteran insects that are found on every continent bar Antarctica; of thousands of mosquito species, only a small number (circa 100) are known to transmit human pathogens (Rozendaal, 1997). This vector activity arises in species with hematophagous females, who consume nutrients from a host animal's blood in order to develop successful eggs. Within the focus of 'mosquitoes', this work primarily investigates the Culicine mosquito species, *Aedes aegypti.* Exploiting the cell culture nature of the experimental work, some resources are also extended to work in other Culicine mosquito species (*Ae. albopictus* and *Culex quinquefasciatus*) and latterly the Anopheline mosquito *Anopheles gambiae*.

Mosquito-borne diseases of humans

For mosquito-borne diseases of humans, the mosquitoes are an obligate vector and disease distribution is therefore tightly linked to the geographic distribution of the mosquitoes (Rozendaal, 1997, Sciences, 2016, Smith et al., 2014, WHO, 2020, Girard et al., 2020). Anthropophilic mosquitoes occur in two subfamilies – Anophelinae and Culicinae. Since the mass transport of goods and people became commonplace (circa 1940s) many mosquito species have become increasingly prevalent as invasive species, particularly those of the *Aedes* genera. The distribution of anthropophilic mosquitoes is expected to increase as global climate change continues to mature (Reiter, 2001, Girard et al., 2020).

Aedes aegypti

Aedes sp. mosquitoes *Ae. aegypti* and *Ae. albopictus* are predominant vectors of endemic arthropod-borne viruses (arboviruses) (e.g. Dengue viruses in South America) and (re-)emerging epidemics such as Zika and Chikungunya (Vega-Rua et al., 2014, Simmons et al., 2012, Gratz, 1999, Diagne et al., 2015, Bhatt et al., 2013, Girard et al., 2020). *Ae. aegypti* mosquitoes are well adapted to urban environments, feeding during the day, resting indoors and laying eggs in small – often manmade - containers of water (Rozendaal, 1997, ECDC., 2017). Following the spread of goods and people, *Ae. aegypti* mosquitoes are found globally in tropical and sub-tropical regions. Where *Ae. aegypti* populations are present, mosquito-borne viruses can follow (Girard et al., 2020, Kraemer et al., 2015, Charrel et al., 2014).

Ae. aegypti mosquitoes have been demonstrated to vector many different arboviruses that affect humans. Dengue fever virus is the greatest of these (by annual incidence and number

of people at risk), causing pathology ranging from non-specific febrile illness to haemorrhagic fever and (less frequently) death (Girard et al., 2020, Sciences, 2016, Simmons et al., 2012). One analysis suggests that 390 million people are at risk of contracting Dengue fever annually, of which 96 million are symptomatic (Bhatt et al., 2013). There is no specific treatment for Dengue fever and incidence of the virus is increased in urban areas of over-crowding and poor sanitation – conditions that favour the vector *Ae. aegypti (Simmons et al., 2012)*. Reported deaths were in excess of four thousand in 2015, biased towards younger age groups (WHO, 2022b). A dengue vaccine (Dengvaxia, Sanofi Pasteur) was licensed in 2015 and is now approved in approximately 20 countries, but has a limited use profile (WHO, 2022b, Thomas and Yoon, 2019). In high risk areas, it is recommended to children with a confirmed previous exposure to Dengue (children who are seropositive). This recommendation is made in response to indications that vaccination of seronegative individuals increases risk of severe dengue if they are subsequently infected for the first time (WHO, 2019a).

Other arboviruses vectored by *Ae. aegypti* mosquitoes include (but are not limited to) Chikungunya virus, Zika virus, West Nile virus and Yellow Fever Virus. Of these, only Yellow Fever virus has a broadly used vaccine (Kraemer et al., 2015).

Aedes albopictus

Ae. albopictus mosquitoes have an overlapping range of vector competencies with *Ae. aegypti* and have been demonstrated to transmit Chikunguya virus, Dengue virus and Yellow fever virus (amongst others) (Vega-Rua et al., 2014, Kraemer et al., 2015, Gubler, 2002, Girard et al., 2020, Amraoui et al., 2016). *Ae. albopictus* can also have an overlapping geographic range and ecological niche with *Ae. aegypti*, though they are typically less urban and less anthropophilic, feeding opportunistically from humans and animals (ECDC., 2017, Rozendaal, 1997). There are indications that decreasing populations of one *Aedes* species can lead to increasing numbers of the other (in a given geographical area)(Kraemer et al., 2015, O'Meara et al., 1995). This highlights a benefit to considering both species when implementing vector control strategies and led to the inclusion of *Ae. albopictus* as a second species of interest in this body of work.

Culex quinquefasciatus

Moving away from the *Aedes* genera, *Culex* mosquitoes are also of the subfamily Culicinae. *Culex quinquefasciatus* mosquitoes are found in tropical to temperate climates, occupying both urban and suburban areas. They are opportunistic blood feeders, with human and avian hosts, creating pathways for zoonotic transmission of a range of pathogens (ECDC., 2020,

Hamer et al., 2009, Lura et al., 2012). In addition to arboviruses such as West Nile virus, *C. quinquefasciatus* mosquitoes are vectors of filarial parasites, avian malaria and avian pox viruses (Farajollahi et al., 2011).

C. quinquefasciatus was chosen as a third species of interest in this body of work, due to its importance as a disease vector and due to increasing availability of genetic resources (genome sequence assembly) and a cell culture line.

Anopheles gambiae

In the latter parts of this body of work (RNA polymerase III promoters), *An. gambiae* cell lines were included as representation of a further species of interest. This arose opportunistically, as *An. gambiae* gene editing tools were explored in an *Ae. aegypti* context. Experimental design did not preclude the use of *Anopheles*-origin cell lines, and so these were included as further work, developing on ideas demonstrated in Culicine-origin cell lines.

An. gambiae mosquitoes are a predominant vector of human malaria, which can cause symptoms ranging from febrile illness to severe anaemia, cerebral malaria and respiratory distress. WHO (2021c) report 241 million clinical cases of malaria in 2020, with 627,000 deaths (mostly in children). Investment in Malaria programmes and research has reached approximately US$ 3 billion each year, 2018 – 2020, and this sustained effort has resulted in well-developed genetic resources for vector (*Anopheles*) species (WHO, 2021c).

An. gambiae (sensu stricto) mosquitoes are anthropophilic, typically taking a blood meal at night and then resting indoors (Rozendaal, 1997). This behaviour is exploited by vector control methods such as bed-nets and by indoor residual spraying of insecticides (Bhatt et al., 2015, Pryce et al., 2022). Malarial disease is caused by infection with *Plasmodium sp.*, whose life cycle requires both an arthropod host and a vertebrate host. *An. gambiae* is the predominant vector of *Plasmodium falciparum*, which causes severe Malaria in humans (Snow et al., 2005, Tolle, 2009).

Interventions for control of mosquito-borne diseases

Mosquito-borne pathogens have complex life-history adaptations that enable them to succeed across both an invertebrate and a vertebrate host. Most arboviral infections have no specific pharmaceutical treatment, though vaccines have been licensed for a handful of viruses (Yellow Fever virus, Japanese encephalitis virus, Dengue virus, West Nile Virus (equine only) and tick-borne encephalitis virus)(FDA).

The Yellow Fever virus vaccine is cheap (not patent-protected) and effective, estimated to have reduced infections and deaths by 27% in Africa between 2006 and 2012, up to 82% in

countries with targeted vaccination campaigns (Garske et al., 2014, Gotuzzo et al., 2013, Staples et al., 2010). Japanese encephalitis virus vaccines have also been in production for decades and are considered safe and effective (WHO, 2019b, Vannice et al., 2021). The one licensed Dengue vaccine (Dengvaxia, Sanofi Pasteur) is much newer (first licensed in 2015) and has a been marred by inconsistent safety and efficacy profiles (Thomas and Yoon, 2019, WHO, 2022b).

Unlike arboviral infections, there are a number of antimalarial drugs available. Treatment recommendations are updated in response to drug-resistance in the pathogens, and there are reports that the current gold-standard treatment – artemisinin combination therapy – is encountering drug-resistance (WHO, 2022a, Ashley et al., 2014, Saito et al., 2020, Nosten and White, 2007, Ashley and Phyo, 2018). Excitingly, WHO (2021c) marks the occasion of the first malaria vaccine to be recommended by World Health Organisation, indicated by a pilot study (Chandramohan et al., 2021) in addition to clinical trials (Rts, 2015) that show 30% efficacy in reducing severe malaria disease.

Although these vaccines provide options for prevention of some mosquito-borne diseases, there remains a need for further measures to prevent disease transmission. Interventions targeting the invertebrate vector, rather than the pathogen itself, can be very successful in reducing disease burden and are recommended by national and international public health bodies (EMCA., 2013, ECDC., 2017, WHO, 2017, WHO, 2021b).

Mechanical interventions

Targeting the mosquito-human interaction at the point of the blood-meal can be accomplished by preventing mosquitoes from biting humans. For mosquitoes that typically seek blood-meals at night, bed-nets are an effective, and cheap, solution (WHO, 2021b, Rozendaal, 1997). Chemically treated bed-nets often offer increased protection, combining a physical barrier with a repellent or an insecticide. Studies have reported that sufficient use of insecticide treated bed-nets can decrease the local population of mosquitoes, reducing the likelihood of anyone being bitten, not just the person sleeping with the net (Bhatt et al., 2015, Pryce et al., 2022, Pryce et al., 2018).

Aedes mosquitoes are more active in the day and so are less affected by bed-nets. They are vulnerable, however, to targeted reductions in larval habitats - small pools of standing water (ECDC., 2017, WHO, 2020, Rozendaal, 1997). Urban mosquitoes in particular use man-made objects that collect standing water for egg laying – discarded tyres are a classical example. Educating people to not allow pools of water to collect and stand for extended periods of

time (days or weeks) can produce a measurable reduction in breeding sites and a corresponding mosquito population decrease (Ledogar et al., 2017).

Mechanical approaches such as these can be cheap and effective at reducing human-biting and can reduce disease transmission in a given area (ECDC., 2017, EMCA., 2013, WHO, 2021b)). They do, however, require an extremely high level of adherence and coverage as only one transmission event (an infected mosquito feeding from a susceptible host) is needed to infect a human, or vice versa.

Chemical interventions

Chemical insecticides and repellents offer a more aggressive approach to reducing the population of mosquitoes that overlap with susceptible humans. These can be applied directly to humans (repellents) or to objects (such as bed-nets or indoor residual spraying, where an insecticide is persistently active on a surface (Pryce et al., 2022, Pryce et al., 2018)), as well as generally to an environment (spraying of neighbourhoods) (Esu et al., 2010)). Although these methods can be efficacious and cost effective, there are issues with insecticide resistance spreading amongst mosquito populations (van den Berg et al., 2021, Ranson and Lissenden, 2016, Moyes et al., 2017).

Biological interventions

Conventional biological interventions such as introducing predators to a niche can be effective in limited scenarios (ECDC., 2017, WHO, 2021b, Ledogar et al., 2017). Insectivorous fish or reptiles can suppress juvenile life-stages of mosquitoes in a body of water, but are limited to the scale of habitat that can in turn support predators (Ledogar et al., 2017). Mosquitoes, particularly *Aedes sp*., can successfully reproduce in extremely small pools of water, such as that which collects (from rainfall) in an unused bucket or the rim of a discarded tyre - environments that are not suitable for their predators.

Instead of introducing predators, there are more sophisticated biological interventions that can also aim to reduce vector population size or can instead aim to directly reduce vector competency of the mosquito population. These exploit male mosquitoes as a ‘harmless’ delivery system to introduce biological characteristics to a pest or nuisance mosquito population, in a species-specific manner (McLean and Jacobs-Lorena, 2016, Hill et al., 2005, Iturbe-Ormaetxe et al., 2011, Burt, 2003, Bian et al., 2010, Alphey, 2002, Alphey, 2014, Alphey et al., 2010). The ‘harmlessness’ of male mosquitoes is predicated on hematophagy being an exclusively female behaviour, necessary for the development of viable eggs.

Population suppression

A longstanding example of such a biological control strategy is the Sterile Insect Technique (SIT) that has been used to control agricultural pest populations of *C. hominivorax* and *Ceratitus capitata* (Mediterranean fruit fly, aka Medfly), amongst others (Alphey et al., 2010, Dunn and Follett, 2017). The SIT principle is to induce sterility in male-only populations of the target species, for example through irradiation. This sterile population is then released where the males will mate with wild females; if females do not mate with fertile males, they will be removed from the breeding population. If sufficient females are removed from a breeding population, overall size of the population is reduced or crashed. This approach relies on a sufficient incidence of sterile males and on sufficient mating competitiveness of those males (Alphey et al., 2010, Dunn and Follett, 2017).

For many insects, including mosquitoes, it is difficult to obtain sufficient male-sterility through irradiation, without unduly compromising their fitness (Alphey, 2002, Yamada et al., 2019, Helinski et al., 2009). The use of genetic modifications to induce sterility can circumvent that issue, as well as providing benefits for sex-sorting of laboratory raised insects (Alphey, 2002). A British company, Oxitec, have demonstrated a female-flightless system for *Ae. aegypti* control. This system utilises a female-specific isoform of Actin 4 that is necessary for flight. Male mosquitoes retain a suitable fitness (and flight) but will produce only flightless female offspring, a sterile/lethal phenotype (Fu et al., 2010). Oxitec has also used another female-lethal ("OX5034") in field trials, reportedly with good success (personal correspondence, Luke Alphey, 2021). The development of this modified mosquito line has furthermore yielded improved strategies for sex-sorting immature mosquitoes and an inducible/repressible gene system to enable development of a line with a female lethal/sterile phenotype (Alphey, 2002, Fu et al., 2010)).

Population replacement

Looking away from population suppression as a method for disease control, there are avenues to directly impact a mosquito's vector competence, therefore reducing transmission by reducing the mosquito's ability to be infected or to infect in turn. This could be achieved through genetic editing systems, similar to engineered sterility, or through infection with the microbial agent, *Wolbachia* (Gantz et al., 2015, McLean and Jacobs-Lorena, 2016, Ogunlade et al., 2021, Iturbe-Ormaetxe et al., 2011).

Wolbachia sp. are intracellular bacteria found in a wide range of arthropod hosts and have been demonstrated in mosquitoes to reduce an individual's susceptibility to further infection with pathogens of human concern (Bian et al., 2013, Bian et al., 2010, Glaser and Meola,

2010, Zug and Hammerstein, 2012, Hilgenboecker et al., 2008). *Wolbachia* infections are maternally transmitted and can have a cytoplasmic incompatibility phenotype in mosquitoes. This phenotype generates an incompatibility between the sperm of *Wolbachia* infected males and the eggs of uninfected females, and can be unidirectional: the eggs of *Wolbachia* infected females remain compatible with the sperm of uninfected males (Sinkins, 2004, Dobson et al., 2001, Kambhampati et al., 1993, Turelli, 2010)). Unidirectional cytoplasmic incompatibility disadvantages uninfected females in a mixed-infection-status population, increasing the prevalence of *Wolbachia sp.* generation by generation and the reduced vector competence phenotype along with it (Dorigatti et al., 2018, Ferguson et al., 2015, Turelli, 2010). Cytoplasmic incompatibility is not seen in all *Wolbachia sp.*, but infected mosquitoes have been demonstrated to reduce mosquito-borne disease transmission in cage and field-trials (Hoffmann et al., 2014, O'Neill et al., 2018, Walker et al., 2011).

Limitations of extant systems

While there are a range of mosquito control strategies available and there has been success in reducing disease incidence (and therefore burden) in human populations, mosquito-borne diseases remain a significant concern to millions of people (WHO, 2021b, ECDC., 2017). With global climate change thought likely to increase the geographic distribution of vector mosquito species, and interventions targeting the mosquito still the most effective to reduce disease incidence, there is a call for additional tools with which to control mosquito populations.

Increasingly complex genetic modification vector control strategies are beginning to be realised; this can be attributed to improvements in molecular biology tools in the last decade (inc. DNA sequencing and CRISPR gene editing). These strategies are typically founded on having a degree of ‘persistence’ amongst a wild population, where a desirable phenotype acts beyond the immediate offspring of released individuals. Such systems are the focus of this work.

Complex gene editing strategies for biological population control

Complex gene editing strategies for vector control have both an effector phenotype and a mechanism for persistence. Each of these is built in a species specific fashion, but there is often overlap in parts of a transgene system that can be directly used in another species, as well as areas where an existing transgene can be replicated in another species of interest through use of homologous genes.

Effector phenotypes remain consistent with those already discussed – those intending population suppression (e.g. sterile and lethal phenotypes) and those intending population replacement (i.e. reduced or removed vector competence). The mechanism for an effector phenotype can be integrated into the mechanism for persistence (e.g. inserting a transgene into and disrupting an essential gene) or it can be separate from the mechanism for persistence (e.g. a transgene expressing a lethal effector, inserted into a non-essential area of the genome). Effector phenotypes are not further explored in this work.

Mechanism for persistence

Population control strategies based on SIT are limited in their effect to offspring (or lack thereof) of the released males and repeated releases must be made to maintain their effect on population size; they are immediately self-limiting. The unidirectional cytoplasmic incompatibility aspect of some *Wolbachia* strains offers a mechanism for persistence where sufficient incidence of the *Wolbachia* infected mosquitoes will cause the phenotype (infection) to overtake the wild-type and reach fixation in the population (reviewed by Alphey (2014)). This persistence is predicated on many variables, including fitness cost of the phenotype, release quantity of mosquitoes, frequency of release and geographic migration of individuals into and out of the population.

Increased persistence can be engineered through genetic modification and the most invasive systems are described as ‘gene drives’ (Alphey et al., 2020). One major class of gene drives, so called ‘homing drives’, are based on the use of site-specific selfish genetic elements. In an individual that is heterozygous for that site-specific genetic element (which can be considered a genetically modified ‘allele’), the genetic element is reproduced in the corresponding wild-type allele – resulting in a cell that is homozygous for the genetic element (Burt, 2003) (Figure 1). Where this genetic element can self-replicate in gamete producing cells, such a system is inherited at super-Mendelian rates (more than 50% of offspring). If an effector mechanism can be tied to this system, it too can therefore be persistent in a population.

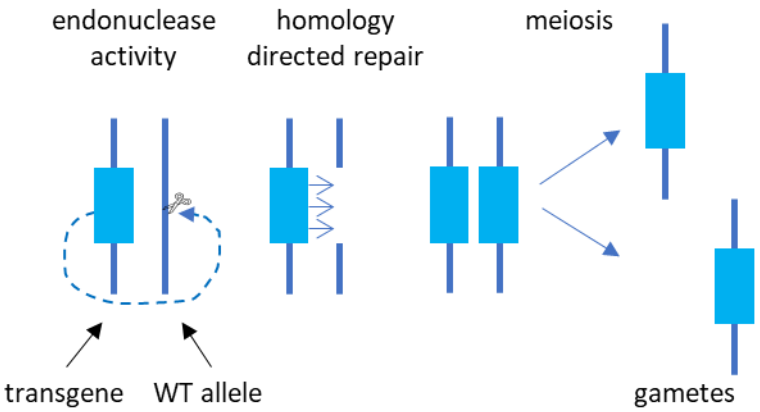


Figure 1: Cartoon representation of an autonomous homing drive. The box represents a transgene insertion that encodes a site-specific endonuclease. In this scenario, the endonuclease is specific for the wild type (WT) allele corresponding to the locus of the transgene insertion. Once this endonuclease is expressed, it cuts the WT allele. The DNA break is repaired by homology directed repair, duplicating the sequence of the transgene. The cell goes from heterozygous to (transgene) homozygous. If this cell undergoes meiosis, every daughter cell (gamete) will have a copy of the transgene.

Gene drives of this design that can replicate the entire drive system (mechanism of persistence plus effector mechanism) by homing, so called 'autonomous' homing drives, potentially have extremely low invasion thresholds in target populations. Correspondingly they may, in the absence of mutation or resistance, be able to invade all populations of a species linked by non-zero migration rates. Such drives are sometimes called 'global gene drive', representing a system's potential to affect an entire species with sufficient time. Such an occurrence is of ethical concern and much care should be taken to prevent accidental construction and release of such genetically modified (GM) individuals (Annas et al., 2021, Benedict et al., 2018, WHO, 2021a).

Nonetheless, there remains a call for vector control systems with the cost-effectiveness of persistence, balanced with a constraint that prevents the spread of a system past its intended borders (temporally or spatially) (Alphey, 2014). The advent of CRISPR gene editing technology has dramatically lowered the technological barriers to building such systems.

CRISPR technology

CRISPR gene editing technology has been adapted from the discovery and description of a bacterial adaptive immune system.

CRISPR - Adaptive immunity

Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated proteins (Cas) are an adaptive immune system in bacteria. The CRISPR gene regions express

two small, nuclear RNAs (snRNAs), tracer RNA (tracrRNA) and CRISPR RNA (crRNA), which together form a duplex structure known as guide RNA. The tracrRNA is a conserved 'backbone' sequence that contributes the secondary structure of the guide RNA needed for (Cas) protein interactions; the crRNA is adapted from the DNA sequence of an invading organism, such as a virus, and enables the guide RNA to bind to the corresponding invading sequence if it is encountered again (Jinek et al., 2012, Garneau et al., 2010, Bhaya et al., 2011, Horvath and Barrangou, 2010). Cas proteins bind to guide RNA and in turn to corresponding DNA, where they have an effector action. For Cas9, this binding is specific to double-stranded DNA (dsDNA) and the effector action is double-stranded endonuclease activity.

This bacterial immune system has been demonstrated to be programmable and to be functional in a wide range of vertebrate and invertebrate organisms. The crRNA sequence can be adapted to correspond to any DNA sequence, so long as there is a corresponding protospacer adjacent motif present (e.g. 5'-NGG-3' for *Streptococcus pyogenes* Cas9) (reviewed by Ran et al. (2013)). A single guide RNA (sgRNA) structure has been demonstrated to function in place of the tracrRNA/crRNA duplex, and the resulting 'CRISPR/Cas9' system has enormous versatility as a gene editing tool (Jinek et al., 2012).

CRISPR - Gene editing

CRISPR-targeted use of Cas9 endonuclease activity to create double-stranded breaks in genomic DNA can result in different repair outcomes, depending on the molecular context (whether there is a repair template available to cell machinery) and on the propensities of a species (homology-directed repair is more common in some species than others) (reviewed by Ran et al. (2013)). Healthy cells do not tolerate DNA damage and will seek to repair the break or else engage apoptosis if damage cannot be repaired (e.g. if there are too many breaks).

End-joining DNA repair

Repair pathways for dsDNA breaks can be described in two groups, end-joining and homology-directed repair (Sansbury et al., 2019, Vitor et al., 2020, Lieber, 2010)). End-joining seeks to resolve the break by re-joining the available 5' and 3' ends. This can be done seamlessly with no loss or gain of nucleotides. Frequently, however, there is loss of nucleotides on one or both ends and the repair therefore alters the DNA sequence (Lieber, 2010). This loss of nucleotides can be as small as one base pair or as big as several thousand base-pairs. Furthermore, end-joining can result in the addition of nucleotides to one or both ends; where there is no homologous template used, this addition is typically small. Considering the wider context of a double-strand break in coding DNA, any change in

sequence can result in changes to protein sequence and therefore function. 'Silent' mutations can arise from end-joining DNA repair where a synonymous sequence change occurs or there is an in-frame deletion (or addition) that does not affect the function of the resulting protein. A frame-shift mutation at the 3' end of a coding sequence can result in a functional, though truncated, protein; not all protein changes result in obvious phenotypes. End-joining mutations typically result in loss-of-function mutant phenotypes, though they can cause gain-of-function phenotypes in some genes (Lieber, 2010, Vitor et al., 2020).

Homology-directed DNA repair

Homology-directed repair requires the presence of a DNA template that is homologous to the site of the dsDNA break. For the cell, a sister chromatid is an ideal template that allows perfect repair of the dsDNA break, without introducing mutations. Imperfect, homologous templates can also be used to repair the dsDNA break, but can result in gain or loss of nucleotides for the damaged site (according to the sequence of the homologous template). The homologous template can be endogenous (a homologous chromosome) or exogenous (e.g. plasmid DNA).

In homology directed repair, the damaged DNA strands are cut back at each 5' end to generate single-stranded DNA 3' overhangs. One overhang is bound in a protein complex that facilitates 'strand invasion' of the homologous dsDNA template, generating a displacement loop. DNA polymerase activity then occurs for both 3' overhangs, using the 'invaded' template DNA strands to determine the nucleotide sequence of the repair. This activity forms junctions between the newly repaired DNA strands and the template, which are resolved by DNA nickase activity, to separate the two double helices (West et al., 2015, Sansbury et al., 2019). The repaired DNA is now an identical sequence match to the template DNA, bookended by the homology regions that were used to identify the template during the repair.

Homology-directed repair can be harnessed to generate sequence insertions by providing an exogenous DNA template with 5' and 3' homology sequences corresponding to the targeted DNA damage (Bibikova et al., 2003, Bollag et al., 1989, Rouet et al., 1994). This method can be used to generate loss-of-function mutations by introducing a stop codon or by adding stretches of nonsense code that disrupt protein secondary and tertiary structure. Similarly, homology-directed repair can cause gain-of-function mutations, adding entire transgenes to a cell's genome. If successfully applied to a germ cell, this method can generate a stable transgenic line (reviewed by Ran et al. (2013)).

CRISPR/Cas9 gene drives

With CRISPR/Cas9 as site-specific gene editing technology, very few components are needed to create a genomic insertion of a transgene: Cas9 protein, a target sequence specific sgRNA and sequence specific homology regions 5' and 3' of the transgene, provided as a DNA template. This system has been used to generate transgenic insects in various mosquito species (including *Ae. aegypti (*Dong *et al., 2015)* and *An. gambiae (*Hammond *et al., 2016)*) and is the underpinning technology for two gene drive systems.

Recessive female-sterile CRISPR gene drive

An example gene drive is a recessive female-sterile system demonstrated by Hammond et al. (2016). The drive is based on integrating a Cas9 - sgRNA transgene into a gene required for female fertility, in laboratory mosquito strain *An. gambiae* (G3). The sgRNA targets the wild-type allele of the same gene, creating a circumstance where the disrupted allele (containing Cas9 - sgRNA) is both the origin of endonuclease activity and the template for homology driven repair (a 'homing drive'). Where the Cas9 - sgRNA transgene is expressed and cuts the wild-type allele, repair of the damaged allele by homology-directed repair results in it containing the Cas9 - sgRNA transgene; heterozygous alleles become homozygous where the transgene is expressed. Hammond et al. (2016) demonstrate that this activity can be induced in gamete producing cells, leading to a super-Mendelian inheritance of the transgene; cage trials report that a 1:1 introductory ratio of transgenic to wild-type individuals escalated to a 75% presence of the transgene (heterozygotes) after four generations.

The phenotype of this gene drive is predicated on the presence of the Cas9 - sgRNA transgene knocking out the function of the disrupted gene, and of that mutation being recessive for female-sterility. As the frequency of the transgenic allele increases in the population, so too will the frequency of individuals homozygous for the transgene/disrupted gene (as in Figure 1). If these individuals are male (and the phenotype is specifically female-sterile), they will continue to contribute the transgene to the next generation. If the homozygous individual is female, it is phenotypically sterile and therefore removed from the reproductive capacity of the population; at sufficient frequency, this could reduce or even crash a population's size.

Although there is a theoretical risk of such a drive taking on 'global gene drive' status, the inherent fitness cost of the transgene (even in heterozygotes) introduces a minimum introduction threshold for the homing drive. Without sufficient emigration of the transgene to neighbouring populations, the drive will not spread. A further constraint, the emergence of resistant alleles, is discussed later in this Chapter (Burt, 2003, Hammond et al., 2016).

Daisy-chain gene drive

The 'Daisy-chain' gene drive system proposed by Noble et al. (2019) is an example of a constrained gene drive and is based on multiple components interacting with one-another. Although this brings complexity, it increases the range of phenotypes that such a system could convey (it is not limited to lethal/sterile).

The system as it is discussed here is tripartite (Figure *2*) and three individual transgenes propagate each other in turn. Although the system could function with as few as two parts (a 'split drive'), using a third element confers additional persistence. A fourth element could be added for further enhanced persistence, but must be balanced against risk of transgene instability and fitness cost to the insect.

An initial "A" element (transgene) is designed as an allele conveying a desired phenotype; there is no CRISPR/Cas feature in this element. A "B" element expresses Cas9 protein and sgRNAs that target the wild-type (WT) allele of the "A" element - where "A" and "B" are present in the same cell, "A" will be duplicated and its WT allele made absent. In the context of a gametes, this set-up directs inheritance of the "A" allele at above-Mendelian ratios and "B" at normal Mendelian ratios. The corresponding "C" element expresses an sgRNA that targets the WT allele of the "B" transgene. Where "B" and "C" are present, the "B" allele will replace its WT counterpart and be inherited at above-Mendelian ratios, whereas "C" cannot affect its own WT counterpart and has a Mendelian inheritance pattern. This is summarised in Figure *2*, along with a graphical representation of the persistence of each element.

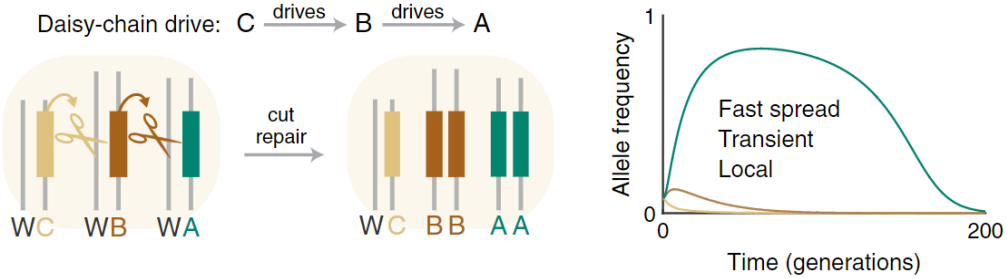


Figure 2: Daisy-chain gene drive illustration, reproduced directly from Noble et al. (2019). "W" indicates the wild-type allele and "C", "B" and "A" indicate the three transgenes that make up the Daisy-chain system. As can be seen at the far left, each element is required for the site-specific cutting of the WT allele of the upstream element. The final element, "C", is not replicated in this system and quickly drops out of the population according to Mendelian inheritance, which is shown along the bottom of the right-hand graph. Upstream elements will persist as long as their downstream element is present, leading to a staggered effect where the loss of "C" ceases the maintenance of "B" and the same again happens to "A" when "B" begins to drop out of the population. This design limits the persistence of the gene drive.

In this system, the "C" element is limited by Mendelian inheritance and will dissipate in a population if it is not re-introduced frequently enough (each transgene is assumed to carry a fitness cost, the magnitude of which will affect dissipation rate). The same dissipation would occur to the "B" element in the absence of "C", and to "A" in the absence of "B". Where all three elements are present, "B" and "A" take on an amplified frequency in the population, hence conferring the desired phenotype caused by the "A" transgene. The system is therefore self-propagating, in a fashion that is limited by the frequency of "C".

Of note, measures are needed to mitigate the possibility of undesired recombination between elements, which could result in an unconstrained global drive. For example if the sgRNA expressed by "C" is acquired by the "A" element, then "A" and "B" together in a population can sustain each other without limit. Noble et al. (2019) term this a 'daisy-necklace' and propose that sufficiently eliminating sequence homology between elements could suitably mitigate this risk (this would need to be proved in controlled trials).

Building a Daisy-chain gene drive in *Ae. aegypti* mosquitoes was a focus of the host lab for this project and the work described in this thesis largely arose from this context.

Current limitations of CRISPR gene drives

CRISPR gene editing technology simplifies the process of creating a gene drive for population control, bringing increased opportunity to ameliorate endemic vector borne diseases in a targeted fashion. There are, however, remaining design challenges to overcome before a system can be fully realised. Hammond et al. (2016) duly note that there is a collateral fitness cost of their transgene, even in individuals that were designed to be 'carriers'. This is thought to result from off-target expression of the Cas9 - sgRNA in somatic cells. Fitness cost of the transgenic individual is an important factor in the efficacy of a gene drive system (Burt, 2003) and this could be improved through identification and validation of promoters with sufficient germline activity and further reduced off-target (somatic) activity (Hammond et al., 2016, Hammond et al., 2021).

These systems furthermore rely on identification of genes and alleles that can deliver the desired phenotype. Identification of such genes relies on homology of function and sequence between species of interest and model species, and additionally on the availability of high quality genome sequence assemblies for the species of interest. Although important, these areas are not directly within the scope of this project.

Where a system calls for dsDNA endonuclease activity, it is noted that many non-homologous repair events (end-joining) are expected to result in sequence changes to the repaired allele. For both CRISPR gene drives described here, there is an assumption that the wild-type allele

of a disrupted transgene will be a sequence match for the sgRNA(s) expressed to target it. This assumption can be false if there is natural allelic variation in a wild population; furthermore, such variation can be induced by non-homologous imperfect dsDNA repair following successful endonuclease activity (i.e. the transgene is not copied during the repair and the repaired sequence is now different from the wild-type sequence) (Burt, 2003, Noble et al., 2017, Unckless et al., 2017).

These alleles that cannot be cut by the CRISPR transgene(s) are collectively referred to as 'resistant' alleles. It is assumed that the transgene carries an intrinsic fitness cost and resistant alleles are therefore detrimental to gene drive activity, as they will typically confer a fitness benefit to the transgene-heterozygous individual. If this fitness benefit is greater than the cost of the mutation (the sequence changes to the allele) then it will be selected for by Mendelian inheritance and a gene drive will correspondingly decrease in frequency in the population. If the resistant allele is driven to fixation, the population will now be unaffected by the release of further transgenic individuals. This effect is described for the recessive female-sterile gene drive reported by Hammond et al. (2017).

The emergence of resistant alleles in a CRISPR-based gene drive is almost a foregone conclusion. Mitigation then, rather than outright prevention, is the goal to enable these gene drives to persist and act as intended in a population. One design solution is to target a sequence location where any imperfect non-homologous repair is likely to confer a fitness disadvantage (Burt, 2003, Noble et al., 2017, Unckless et al., 2017). One example of this is at the 5' end of a haplo-insufficient gene. Mutations at the 5' end of a gene are most likely to disrupt gene product function if there is a frame-shift mutation and the haplo-insufficient quality confers a significant (if not lethal) fitness cost to heterozygotes, preventing the mutation from persisting in the population. Another example with reported success in cage-trials targets the *doublesex* gene of *An. gambiae*, where resistant alleles disrupt gene function and convey a significant fitness cost (Kyrou et al., 2018).

A simpler design solution would be to express multiple sgRNAs each targeting different loci in the same gene allele (Noble et al., 2017, Marshall et al., 2017). Noble et al. (2019) estimate that four sgRNAs targeting each allele would suffice to comfortably mitigate against the persistence of resistant alleles in a Daisy-chain gene drive. These loci can be chosen based on availability of population genome resources showing alleles already present in wild populations, targeting areas that do not show natural variation. The multiple nature of the sgRNAs (and therefore target loci) should ensure that resistant sequences produced by repair of one cut site are still vulnerable to then being cut by the other sgRNAs.

Although this design solution is simpler, it relies on a validated capacity to express multiple sgRNAs from a single individual and then upon those stretches of homology within a transgene not inducing homologous recombination amongst itself - or between different transgenes each expressing multiple sgRNAs if there is a system such as Daisy-chain gene drive in place. This limitation is explored in more depth in Chapters 4 and 5.

Focus: Building a Daisy-chain gene drive in *Ae. aegypti* mosquitoes

The circumstance of mosquito-borne diseases and the concept of a CRISPR/Cas based Daisy-chain gene drive for control of mosquito populations, particularly *Ae. aegypti*, form the wide-angle context of this research project. From this, three granular targets were selected:

1) Assess the effect on transgene expression in *Ae. aegypti* of translational modifiers reported in *D. melanogaster* and *B. mori*, amongst others.

> This work aimed to establish methodological techniques and to fulfil a technological gap in our ability to modulate transgenic protein expression, which is important for many gene editing studies.

2) Establish a robust CRISPRa assay for differential measurement of sgRNA abundance and use that assay to validate novel RNA polymerase III promoters for expression of sgRNAs in *Ae. aegypti* and additional species of interest.

> This work aimed to recreate a published assay (CRISPRa) as a way of improving the process for measuring sgRNA abundance. The established assay was then used to measure the RNA polymerase III promoter activity of a panel of novel promoter sequences. This contributed towards the construction of a Daisy-chain style gene drive in *Ae. aegypti* (not within the scope of this project).

3) Validate methods for improving transgene design for the expression of multiple sgRNAs, as would be required by a Daisy-chain gene drive in *Ae. aegypti*.

> Following preliminary experiments, decreased lengths of RNA polymerase III promoters were found to show no decrease in sgRNA expression activity. This was confirmed for a panel of promoters in cell culture and was done to improve our capacity to build a stable transgenic insect, containing multiple transgenes and expressing many different sgRNAs. This built on the work of chapter 2 and also contributed towards the construction of a Daisy-chain style gene drive.

In addition to *Ae. aegypti* cell line Aag2, cell lines representing *Ae. albopictus* and *C. quinquefasciatus* were included in several experiments. With the facilities in place to conduct

these experiments, it was decided that repeat use of the same plasmid constructs and assay equipment in additional species of interest was a worthwhile endeavour. As well as demonstrating cross-species activity of some factors, this approach generated increased statistical significance for panel experiments. Moth cell line Sf9, *Drosophila* cell line S2 and *An. gambiae* cell lines Sua5.1 and 4a_2 were used variously for positive control experiments, and for a piece of further work with *Anopheles sp.* RNA polymerase III promoters.

Methodology

To maximise the ratio of time and resource cost with results generation, this work utilises cell culture models of insect species of interest and is done entirely *in vitro* without the use of *in vivo* resources. For each area of work, this methodology is in line with preceding reports, such as Pfeiffer et al. (2012) and Tatematsu et al. (2014) (Chapter 3), Chavez et al. (2015) (CRISPRa assay), Konet et al. (2007) (Chapter 4) and Noble et al. (2019) (Chapter 5).

Immortalised cell lines used here do not directly represent any specific tissue, organ or life-stage. They do, however, represent the genome and cellular machinery typical of their species of origin, which is often sufficient for experimental purposes. Many insect cell lines are robust and have a low resource demand, requiring mainly adequate nutrition through appropriate media, adequate temperatures for growth and sterile conditions to prevent unintended infection. They can be grown as adherent monolayers and can grow quickly, often reaching 100% confluence in 72 - 96 hours after splitting. Cell lines can be grown in parallel, with less effort and cost for additional cell lines than for the first.

For comparison, mosquito husbandry is far more intensive, requiring almost daily care, climate controlled rooms and maintenance at each life stage (eggs, larvae, pupae, adults). If an additional species has similar climate control requirements, than it could be maintained in the same room; this is generally the only area of reduced cost for additional species as opposed to the first species (unless there are already full-time technicians available). Conducting genetic experiments with these insects is a further undertaking, requiring specialised skills and equipment for embryo injections. There is then additional resource cost for identification of successfully transformed individuals and maintenance of them for experimental purposes. The process takes weeks or months, as opposed to the days or week required for a cell culture experiment.

Insect cell cultures can be successfully transfected with commercial lipid-based reagents, enabling introduction of exogenous nucleic acids and proteins to the intracellular environment. Throughout these projects, transfection is used to introduce various plasmids

expressing experimental transgenes into cells and reporter protein systems are used to measure effects. Although there are clear limits to the applicability of cell culture results as compared to those generated in a whole adult insect, the cell culture format enables a scale of work that can very quickly generate statistical significance. The limitations of a cell culture approach to each research question is examined more closely in each results chapter.

Chapter 2: General Methods

Cell Culture

Eight distinct cell lines were used in the course of these experiments. Cell lines Aag2, Hsu, C6.36 and U4.4 represent Culicinae mosquitoes; Sua5.1 and 4a_2 represent *Anopheles gambiae* mosquitoes. Cell line Sf9 is of species *Spodoptera frugiperda* and was used as a representation of lepidopteran species in this project. *Drosophila melanogaster* cell line S2 was used as a positive control, where experimental constructs or designs originated in *D. melanogaster* publications. Table 1 indicates the cell lines used, their species of origin, their publication history and the source of each cell line used here.

Table 1: Cell lines used in this project

Cell line	**Species (tissue)**	**Published origin**	**Source** (Kind gift of)
Aag2	*Aedes aegypti* (embryo)	(Peleg, 1968a, Peleg, 1968b)	Rennos Fragkoudis
Hsu	*Culex quinquefasciatus* (ovary)	(Hsu et al., 1970)	Alain Kohl
C6.36	*Aedes albopictus* (larvae)	(Igarashi, 1978)	Rennos Fragkoudis
U4.4	*Aedes albopictus* (larvae)	(Singh and Pavri, 1967)	Rennos Fragkoudis
Sua5.1	*Anopheles gambiae* (larvae)	(Catteruccia et al., 2000, Muller et al., 1999)	Leon Mugenzi
4a_2	*Anopheles gambiae* (larvae)	(Catteruccia et al., 2000, Muller et al., 1999)	Leon Mugenzi
Sf9	*Spodoptera frugiperda* (pupa)	(Vaughn et al., 1977)	Rennos Fragkoudis
S2	*Drosophila melanogaster* (embryo)	(Schneider, 1972)	Rennos Fragkoudis

Cell line validation

The cell lines used in this project are designated as models of their species of origin, and experimental results are interpreted, at times, with reference to their species. In recognition of this, samples were taken from each cell line and PCR species barcoding was used to validate species origin of each sample. This experiment is detailed in full in Appendix A (page 130) and all cell lines tested were found to be concurrent with their described species of origin (against a reference database and against reference insects where available).

Cell line maintenance

Materials

All insect cell lines were maintained at 28 °C in benchtop incubators without CO_2 or humidity control, as adherent cultures. Corning cell culture flasks with angled neck and plug seal cap style (Fisher Scientific, UK), were used as standard for cell culture maintenance; this was of particular importance for the health of Hsu cultures. Corning Falcon cell scrapers (Fisher Scientific, UK) were used for mechanical disruption to suspend cells in their supernatant; this material was chosen for operator convenience and was not noted to have an effect on culture health.

Cell harvesting (techniques that include lysis) was done under laboratory conditions (not in sterile conditions) and all other cell handling was done in a class 2 microbiological safety cabinet, following standard sterile technique protocols. Virkon reagent (2%) was used to decontaminate all liquids for disposal and solids were autoclaved for disposal, in accordance with the containment level 2 standards of the host laboratory.

Cell cultures were maintained in sterile media, supplemented as noted in Table 2. Manufacturer's recommendations were followed for storage and shelf-life of reagents. Material specifics are noted in Appendix B (page 142).

Table 2: Media and supplements for maintaining each cell line

Cell line	Media	Supplement	Antibiotic	Serum
Aag2	L-15	TPB, 8%	Pen/strep, 1%	10%
Hsu	Schneider's	n/a	Pen/strep, 1%	10%
C6.36	L-15	TPB, 8%	Pen/strep, 1%	10%
U4.4	L-15	TPB, 8%	Pen/strep, 1%	10%
Sua5.1	Schneider's	n/a	Pen/strep, 1%	10%
4a_2	Schneider's	n/a	Pen/strep, 1%	10%

Cell line	Media	Supplement	Antibiotic	Serum
Sf9	Insect Xpress	n/a	Pen/strep, 1%	10%
S2	Schneider's	n/a	Pen/strep, 1%	10%

Sub-culturing cell lines

Culture health for each cell line was maintained by regular sub-culturing, to encourage cells to remain in a growth phase and to prevent over-crowding. A range of flask sizes were used depending on the number of cells required week by week and so sub-culturing was governed by confluence of each flask. When a culture reached 80% confluence (+/- 10% as needed), it was sub-cultured into a new flask at a ratio of 1:3 to 1:6 for each cell line apart from S2, which grew faster and was sub-cultured at 1:10 to 1:15. These ratios maintained cultures at two sub-cultures per week and could be adjusted depending on culture health and cell requirements.

To sub-culture, existing media was aspirated off and discarded. Fresh media sufficient to cover the culture surface of the flask was added gently to the flask. A (disposable) cell scraper was used to gently but thoroughly scrape the entire growth surface of the flask, paying particular attention to corners and edges. The fresh media in the flask was gently tilted across the scraped surface to suspend cells and a 10ml serological pipette was used to gently aspirate suspended cells up and down, rinsing the growth surface of the flask. This step was important to reduce cell clumping; over-mixing could kill cells through shearing.

For sub-culturing, the total volume of suspended cells was measured by serological pipette and the desired volume (e.g. 1/3rd, 1/5th) was transferred to a new flask and the rest discarded. Fresh media was added to bring the new flask to a suitable volume for culture maintenance and the flask was labelled and put away in the incubator.

Cell line freezing and thawing

To establish a working stock of each cell line, aliquots could be cryo-preserved for future use. This was achieved through use of a freezing medium, made up of 10% dimethyl sulfoxide (DMSO), 70% un-supplemented growth medium and 20% fetal bovine serum (serum). Cultures for cryo-preservation were grown as normal, then suspended in media (scraped from the flask surface) as for sub-culturing. A known volume of fresh media (usually 10ml) was used during scraping and a small volume of cell suspension was set aside for cell counting.

Cell counting

Cell counting was done using a standard trypan blue method. One part cell suspension was mixed with one part 0.4% trypan blue and pipetted onto a re-usable glass cytometer, under a microscope slide. Cells in the reticule were counted using 10X magnification with a light microscope, then multiplied by the dilution factors to obtain the estimated cell density of the suspension.

Freezing

For cryo-preservation, the total number of cells was calculated and a suitable freezing density chosen in the range of 5 x 10^6 to 1 x 10^8 (ideally ~1 x 10^7) cells/ml. The entire cell suspension was transferred to a 50ml tube and centrifuged at 500g at 4°C for 5 minutes to pellet cells. The supernatant was discarded and cells resuspended to the desired volume in freezing medium. DMSO is cytotoxic at room temperature and so was added last to the cell suspension, once cryovials were prepared and labelled. Cells were then transferred to cryovials at 1ml/vial and tightly sealed (screw top) to prevent liquid nitrogen ingress. Cryovials were then placed in a controlled-rate-of-temperature-decrease device (Corning CoolCell, Fisher Scientific, UK) and stored at -80°C for 24 hours. After 24 hours, frozen cryovials were transferred (on dry ice) to long term storage in liquid nitrogen.

Resuscitation

As needed, cryo-preserved cells could be resuscitated to begin a fresh culture. Cryovials were removed from liquid nitrogen storage and transferred on dry ice to the laboratory space. 25cm^2 flasks were prepared with fresh media (specific to the cell line in question) and then cryovials were defrosted in a dry heat bath at 37°C for the minimum time required to melt the suspension. The cell suspension (in cryo-preservation media) was transferred by 1ml pipette to the prepared flask, then stored in the incubator for 24 hours to allow cells to adhere to the flask. The supernatant was gently removed and replaced with fresh media. Cells were then sub-cultured as normal (typically by moving to a larger flask size).

Cell seeding in plates

Cell transfections were carried out predominantly in 96-well plate format, using Nunc 96 well TC-treated microplates (Fisher Scientific, UK).

Healthy cell cultures were maintained as described (page 21). Once cells were suspended in fresh media, they were counted as described (page 23). Cell density was then used to calculate the total available cells and the dilution needed to achieve seeding density for the cell line in question. A small range of seeding densities were initially tested for each cell line

and those shown in Table *3* were selected based on confluence of cells at 24 hours post-seeding. At the time of transfection (24 hours post-seeding), cells needed to be firmly adhered to the bottom of the well and at a density high enough to maintain the growth phase for a further 48 hours post-transfection.

Cells were diluted in fresh media to the required density, then seeded at 100µl per well in the 96-well plate(s). A multi-channel pipette, filter tips and a sterile reservoir were used for this process. Seeded plates were labelled, had the lid put on and were stored (often stacked) in the incubator.

Table 3: Seeding densities for each cell line (96-well plate)

Cell line	**Cells/well**
Aag2	$6.00x10^4$
Hsu	$5.50x10^4$
S2	$6.25x10^4$
Sf9	$5.00x10^4$
U4.4	$5.00x10^4$
C6.36	$6.25x10^4$
Sua5.1	$1.20x10^5$
4a_2	$1.20x10^5$

24 hours post-seeding, media was gently removed from each well using a multi-channel aspirator and replaced with antibiotic- and serum-free media for transfection (90µl/well).

Cell transfection

All cell transfections were carried out with lipid-based transfection reagents and Opti-MEM media (Gibco, Fisher Scientific, UK). Lipofectamine 2000 (Invitrogen, Fisher Scientific, UK) was initially used, then replaced with TransIT Pro transfection reagent (Mirrus Bio, Gene Flow, UK), which conveyed better transfection efficiency in all cell lines tested. Each transfection reagent was used according to manufacturer's recommendation, with the

following specifications used for reagents with a suggested range (for one well of a 96-well plate):

Lipofectamine 2000: 0.2µl reagent in a 10µl reaction volume
Incubated 20min before adding to cells, then 3-4hrs before removing from cells and replacing with maintenance media

TransIT Pro: 0.2µl reagent with 0.1µl boost reagent in a 10µl reaction volume
Incubated 15min before adding to cells, then 3-4hs before removing from cells and replacing with maintenance media

Transfected cells were incubated in the transfection mix for four hours (for each transfection reagent) before removing it by aspiration and replacing it with 100µl/well of maintenance (supplemented) culture medium. Time between transfection and cell harvesting was typically 48 hours, reduced to 24 hours for some experiments.

Dual luciferase assay

Materials

Promega brand Dual Luciferase Assay technology was used throughout. Sample handling was carried out in laboratory conditions, not using a microbiological safety cabinet, in accordance with local codes of practice. Waste liquids were decontaminated with Virkon (2%) and solid waste that had contacted live cells was autoclaved for disposal (in accordance with the containment level 2 rating of the facility).

Promega Dual Luciferase Assay kit and supplemental 5x Passive Lysis Buffer (Promega, UK) were used throughout. Cell culture grade (sterile before use) ion-free phosphate buffered saline (PBS^-) was kindly provided by an on-site service team. Opaque, white optical plates were obtained from Thermo Scientific (UK). Reverse osmosis (RO) water was generated at point of use and molecular grade water was Millipore brand (Merck, Germany).

Access to a GloMax multi+ (Promega, UK) luminometer with automatic dual injectors was kindly provided by Rennos Fragkoudis. This luminometer was maintained to the manufacturer's recommended service schedule and was cleaned with the recommended wash program at the end of each session. Data output was in a .CSV format.

Cell lysis

Passive lysis buffer (PLB) was made up to 1x using RO water at sufficient volume for 28µl per well. Transfected cells were prepared for lysis by aspirating off supernatant from each well

and then rinsing twice with 100µl/well PBS^-. This wash step was introduced to prevent sample contamination with residual media that could affect downstream assay chemistry. Passive lysis buffer (1x) was then applied to the cells in each well at 28µl/well. Lysed plates were incubated at room temperature if a high volume of plates was processed at once. All plates were then stored at -80°C for a minimum of 1 hour to ensure that every sample was exposed to at least one freeze-thaw cycle. Stored plates were sealed with parafilm and labelled to include the number of freeze-thaw cycles that a sample had been exposed to, as this could affect luciferase activity.

Preparation of samples for dual luciferase assay

Acknowledging the sensitivity of luciferase products to freeze-thaw, plates of samples were stored at -80°C until time of use. Samples were defrosted at room temperature and allowed to equilibrate to room temperature before use, to avoid temperature effects on luciferase activity. Cell lysate samples were stored in their tissue-culture plate, with a volume of each sample transferred to an opaque, white optical plate for the dual luciferase assay. Up to 7µl of cell lysate was measured by dual luciferase assay in one event, maintaining sufficient sample for a duplicate assay if needed and additionally for preliminary testing to calibrate the amount of sample used for the dual luciferase assay for that experiment (discussed further in Chapter 3).

Each sample was made up to 10µl using molecular grade water for the dual luciferase assay. Multichannel pipette and single-use reservoirs were used for liquid handling.

Preparation of dual luciferase assay reagents

Following a local protocol, dual luciferase assay reagents were used at 1 in 10 dilution. Although this configuration presented limitations of the assay chemistry (discussed in Chapter 3), it enabled a meaningful increase in sample quantity without affecting resource cost. Dual luciferase assay reagents were stored, aliquoted and made up according to manufacturer's recommendations. Where a 1x solution is advised, a 0.1x solution was made by diluting with RO water. Particular care was given to the freeze-thaw sensitivity of Stop & Glo Substrate as well as the light-sensitivity of the same (Promega, 2015).

Preparation of luminometer

The GloMax Multi+ luminometer (Promega, UK) with dual injectors was prepared and used according to manufacturer's recommendations. Default settings were used for the injectors apart from using a 70µl injection volume. Priming protocols were used for every session. At

the end of each session (1 to 5 consecutive sample plates), the recommended rinse cycle was used (alternating RO water and 70% ethanol) to clean the injector needles and tubing.

At the beginning of each session, three blank wells were measured to confirm that background activity of the assay (reagents and optical plate) was below 100 arbitrary light units (ALU). At the end of each plate read, sample plates were visually inspected to ensure that the total volume appeared correct for every sample well. These quality assurance checks arose from a series of troubleshooting incidents and where any plate of samples needed to be repeated, all samples from that experiment were re-read, ensuring that they had all undergone the same number of freeze-thaw cycles.

Nucleic acid techniques

The DNA and RNA materials used in cell transfection experiments were generated through a combination of standard molecular biology techniques and commercial synthesis. Plasmids with transgenes expressing a protein product (typically luciferase) were generated in house, whereas those with small RNA gene products were typically shorter (<1kbp) and were commercially synthesised. The specifics of each plasmid or *in vitro* transcribed sgRNA are discussed in the relevant results chapter and where colleagues kindly contributed to the production of materials, they are cited alongside the material description.

Materials

Plastic consumables (e.g. pipette tips and 1.5ml tubes) were sourced predominantly from StarLabs (UK). DNA preparation kits were Machery Nagel brand (Germany) and were used according to manufacturer's recommendations. Enzymes for DNA techniques were purchased from New England Bioscience (NEB) (UK). RNA reagents and materials were Ambion brand (Fisher Scientific, UK). Molecular grade water was used for all DNA work, DEPC-treated water was used for RNA handling.

Standard equipment (e.g. centrifuge, thermocyclers, heat blocks) were used as appropriate and Thermo Scientific Owl EasyCast gel tanks were used for agarose gel electrophoresis. Agarose gels were made up with Agarose from Sigma-Aldrich (Merck, Germany) and Tris-Acetate-EDTA (TAE) buffer was made up on site using RO water and 50x TAE (Fisher Scientific, UK). All nucleic acid quantification was done using Nanodrop 2000 (Thermo Fisher Scientific, UK), according to manufacturer's recommendation.

Plasmid cloning

PCR primers

Primers for polymerase chain reactions (PCRs) were designed using Benchling [Biology Software] (https://benchling.com). Due consideration was made for off-target effects of any primer design and any specific design considerations (e.g. to integrate additional sequence 5' or 3' of the PCR product) is discussed in the pertinent results chapter. Custom oligonucleotides were purchased from and synthesised by Sigma-Aldrich (Merck, Germany).

PCR

PCRs were carried out using two specifications of polymerase enzyme, Q5 hot-start High Fidelity (NEB, UK) for plasmid sequences and DreamTaq (ThermoFisher Scientific, UK) for diagnostic screening where high sequence fidelity was not necessary. PCR reactions – both reagent mix and thermocycles – were designed according to manufacturer's recommendations and specifics for each reaction are detailed in each results chapter. Purification of PCR products ('PCR clean-ups') were carried out using spin column kits (Machery Nagel, Germany).

Restriction enzyme digests

Restriction enzyme digests ('digests') were carried out in accordance to manufacturer's recommendations, with adjustments to the length of incubation and quantity (units) of enzyme made in order to maintain a rule-of-thumb rate of 5U per 1µg DNA per hour of incubation. Dephosphorylation was carried out concurrent with digests where the two reactions used a common buffer, in accordance with manufacturer's recommendations. Where dephosphorylation followed a digest, a column clean-up was used to change the buffer to that suitable for dephosphorylation. Dephosphorylation was used to prepare DNA products for ligation to a phosphorylated, second DNA product.

Agarose gel electrophoresis

Agarose gel electrophoresis was routinely used to separate nucleic acid products by size, both to visualise size-grouped products of a reaction (PCR or digest) and to isolate and extract a particular size of product. Agarose gels were made to a percentage of 0.8% - 1.2% depending on the size of the desired product. The running conditions (time and voltage) were determined for each gel based on size of the gel, agarose percentage of the gel and the size of desired nucleic acid product. Hyperladder 1kb DNA ladder (Bioline Reagents, USA) was used for the majority of products and GeneRuler 50bp (ThermoFisher Scientific, UK) was used for nucleic acid products smaller than 300bp. Gel loading dye (6x, NEB, UK) was used

according to manufacturer's recommendation. Gel imaging was carried out using a blue light box (B-BOX Blue Light LED Epi-illuminator, SMOBIO (Cambio Ltd, UK)).

Ligation

T4 DNA ligase (NEB, UK) was used for ligation of double stranded DNA products. Ligations were carried out according to manufacturer's recommendations.

Transformation of competent bacteria

Ultra-competent XL-10 Gold cells (Agilent, UK) were used to transform plasmids. This was done according to manufacturer's recommendations, but with one tenth of the recommended number of cells. Transformed cells were plated on to LB-agar plates with selection antibiotics, 100µg/ml Ampicillin or 50µg/ml Kanamycin, according to the resistance factor expressed by the plasmid. Ligation-negative controls were transformed alongside experimental groups to determine the rate of non-specific colonies that would grow. Successful, individual colonies (clonal) were picked from LB-agar plates and used to grow up clonal cultures of transformed cells.

Plasmid purification

Clonal cultures of transformed cells were processed by plasmid purification kit (Machery Nagel, spin column), with appropriate kits used for the volume of cells to be processed. These protocols were carried out according to manufacturer's recommendations and purified plasmids were eluted in the elution buffer supplied with the kits. Plasmid concentration was measured using Nanodrop.

Working stocks of plasmid were stored at -20°C. Clonal cultures of sequence confirmed plasmids were stored in a 25% glycerol solution (diluted in one part water and two parts LB media) at -80°C for long term preservation of plasmid stocks. A fresh culture could be obtained from this glycerol stock by spreading a small scraping of preserved cells across an LB-agar selection (ampicillin or kanamycin) plate and selecting a single clone to grow up to the desired culture volume in LB-media.

Sanger sequencing

Routine Sanger sequencing was used to verify all completed plasmids. This service was procured through Source Bioscience (UK). Sequence data was uploaded to Benchling [Biology Software] (https://benchling.com) for trace and sequence analysis. The specific primers used for Sanger sequencing are noted alongside each plasmid's design.

Custom synthesised sequences

For short (< 1kbp) transgene sequences, nucleotide sequences were synthesised by Twist Bioscience (USA). These sequences were procured as linear fragments that could be blunt end cloned into commercial expression vectors (e.g. pJet, ThermoFisher Scientific, UK) or as entire plasmids in the Twist Bioscience proprietary expression vector, pTwist.

In some instances, entire plasmids were synthesised where large, complex transgenes were desired, for example insect codon-optimised dCas9 expression plasmid. This service was procured from Genewiz, USA.

sgRNA synthesis

in vitro transcribed sgRNAs were generated for some experiments. This work was kindly carried out by Michelle Anderson and Victoria Norman, using custom synthesised oligonucleotide primers and a MEGAscript T7 kit (short transcripts) (Ambion, FisherScientific, UK) according to manufacturer's recommendations.

Chapter 3: Modulation of transgene expression through translational modification

Introduction

Expression of transgenic proteins in insect systems

The ability to express genes of interest in a species of interest has underpinned genetic engineering and synthetic biology from their conception. Insect cell systems have been used to express recombinant proteins since the 1980s (Smith et al., 1983, van Oers et al., 2015, Christian and Andreas, 2013).

Expressing transgenic proteins in a temporally (developmental stage) or spatially (tissue) specific way is possible in many insect species, typically with limited control of the quantity of the protein produced. Finding a way to address this gap in our technical capability is of interest to several fields of insect research:

- Commercial production of recombinant proteins
- Fundamental research, typically in model species
- Production of genetic modification systems for vector control

While some work has been done to address this shortcoming in model insect species (*D. melanogaster*) and commercially relevant species (*Bombyx mori*), there is a lack of robust, empirical evidence for predictable modulation of protein expression levels in non-model, nuisance and pest species (Pfeiffer et al., 2012, Tatematsu et al., 2014). Genetic systems of population control can depend on control of protein expression levels, so this lack of information is a key gap in our technical capability to fulfil such designs.

The case for being able to increase expression of a transgenic protein is relatively straightforward: increased expression of a transgenic marker can make a target tissue or a modified insect easier to identify, by eye or by machine sorting (Pfeiffer et al., 2012). Similarly, increased expression of a toxic effector protein can improve efficacy of a lethal-phenotype – especially if it is constrained to a particular trigger, such as blood-meal induced expression (Haghighat-Khah et al., 2019). The converse, however, is also valuable. The ability

to down-regulate expression of a transgenic protein is an option for reducing unwanted toxicity during ubiquitous expression or potentially to reduce effects of 'leaky' expression of effector proteins in the 'wrong' tissues or at the 'wrong' developmental stage. The fitness cost of unwanted Cas9 expression noted by Hammond et al. (2016) is an example of this.

Control of amount of protein expression

Eukaryotic gene expression is a multi-part process (Figure 3) that has several points of endogenous control and therefore several potential points of intervention for designed control of protein expression.

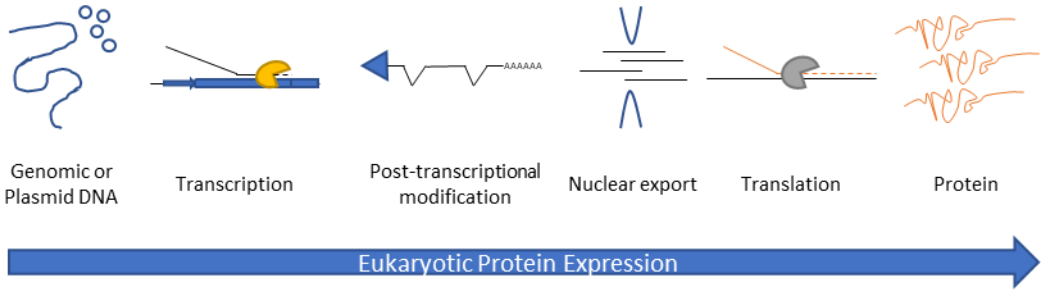


Figure 3: Representation of Eukaryotic gene expression. Transgenic gene expression follows the same process as endogenous gene expression. There are several points of intervention that can be targeted to effect protein expression as part of transgene design.

DNA

One solution to increasing transgene expression is to increase the 'copy number' of the transgene – either through multiple insertions into the genome or by providing more plasmids with the same transgene. In this process the entire gene (promoter, coding sequence (CDS) and terminator sequence) is multiply present and more protein is therefore produced. There is a limit on how much protein a cell can produce, a ceiling at which further copies of a gene will not increase protein expression. In practical terms there are more limiting factors, particularly for transgenes inserted into the genome. Each insertion runs the risk of being in a transcriptionally inactive region, of being lost across generations and of homologous recombination between identical insertions in different locations (causing unintended mutations). In short, multiple insertions are more difficult to achieve and carry more risks than a single insertion.

Extra-genomic DNA (e.g. plasmid DNA) is more tolerant of increased gene 'copy number' and more likely to be limited by availability of transcription factors and other protein expression machinery. This is convenient for *in vitro* experiments and can be harnessed for increased production of recombinant proteins. For generation of stably transgenic organisms, genome

insertion is key and increased transgene copy number or additional plasmids is therefore of limited use.

Transcription

Transcription is the synthesis of an RNA strand from complementary DNA. Chromatin structure of the DNA can affect the transcription rate of a genome-integrated transgene but is often dictated by chance (non-specific integration) or by other design factors (e.g. targeted gene knock-out by insertion of a transgene). The most effective control of transcription is achieved through selection of an RNA polymerase II promoter sequence (promoter). The promoter sequence of a gene dictates the amount of transcription activity, timing of transcription activity and spatial specificity of transcription activity (i.e. in a specific cell or tissue type), though this may be further modified by chromatin context, sometimes quite substantially.

There are a limited (but significant) number of cis-acting promoter sequences that have been identified and validated for use in transgenic mosquitoes (for example in *Ae. aegypti* (Anderson et al., 2010, Akbari et al., 2014, Webster and Scott, 2021). The availability of validated promoters is positively linked with the popularity of an organism or species in research; non-model organisms have a narrower pool of promoters to choose from. Promoters with temporal or spatial specificity typically cannot cross the species barrier (synthetic promoter 3xP3 is a notable exception (Berghammer et al., 1999)) and must be identified in the species of interest. There is not usually an option to select promoter by strength of activity if specificity is a pre-requisite.

Without such constraints, ubiquitous promoters can be selected based on strength of activity. Ubiquitous promoters can be of endogenous (e.g. *Ae. aegypti* promoter of poly-ubiquitin (Anderson et al., 2010)) or viral origin. Insect viral promoters are well described and can have a wide activity across species and even across orders (e.g. promoter HR5-IE1 is active in Diptera and Lepidoptera (Huynh and Zieler, 1999, Ren et al., 2011, Fu et al., 2010)).

Choice of promoter is a suitable way to modulate protein expression if a ubiquitous expression pattern is acceptable. RNA polymerase II promoters are typically in the range of 500-2500bp, which is a consideration when a complex genetic control strategy calls for multiple transgenes in the same individual.

Post-transcriptional modification

Post-transcriptional modification is the maturation of a transcript into a mature messenger RNA (mRNA), which includes addition of a 5’ cap and a 3’ poly-adenylation ‘tail’ as well as

removal of intron sequences (splicing). The addition of 5' cap and 3' tail is mediated by the transcription complex and is not a target for control of protein expression. Splicing is mediated by the intron sequence of the transcript and is shown to enhance transgene expression in mammalian cell culture; there is evidence to corroborate this effect in insects (Brinster et al., 1988, Buchman and Berg, 1988, Duncker et al., 1997, Huang and Gorman, 1990, Pfeiffer et al., 2010, Zieler and Huynh, 2002).

The magnitude of this effect has been shown to depend on the promoter sequence and the orientation of the intron, but not the intron sequence used (Pfeiffer et al., 2010, Zieler and Huynh, 2002). Because of the evidence that protein expression is enhanced by inclusion of an intron but not by the specific intron used, it was not selected for further exploration as an experimental variable in this project.

Nuclear export

mRNA is transcribed and modified in the nucleus, then exported to the cytoplasm for translation. Nuclear export of mRNA is controlled by a suite of export factors, linked to both transcription and splicing machinery (Stewart, 2010, Katahira, 2015). This does not represent a viable target for modulation of transgene expression.

Translation

Translation is a multi-phase process whereby mRNA in the cytoplasm is used to generate amino acid chains (peptides) that are the primary structure of a protein. This process is mediated by a complex of translation factors (ribosomes and proteins) and is cyclic – a single mRNA can be used to produce multiple (identical) polypeptides. Each phase of translation (initiation, elongation, termination and recycling) can be targeted to modulate (usually enhance) translation efficiency and therefore protein output (Pestova and Hellen, 2001, Jackson et al., 2010, Zhou et al., 2016, Schuller and Green, 2018). This can be done through the untranslated regions of mRNA (5' and 3'UTR) or through the translated (coding) sequence.

There are untranslated regions (UTR) at both the 5' and 3' end of mRNA. These regions persist from DNA to cytoplasmic mRNA, but do not appear in the expressed protein. The UTRs modulate translation initiation, termination and are thought to effect recycling of translational machinery. They also contribute to mRNA stability (how long an mRNA exists in the cytoplasm before it is degraded). An aspect (translation initiation sequence) of the 5' UTR is the primary focus of this chapter. Some work is also done to characterise different 3' UTR sequences.

Translation Initiation

Translation initiation is the process by which the translation complex is associated with the mRNA (5' cap) and recognises the start codon (5'UTR) so that elongation can begin. The translation initiation sequence, immediately 5' of the start codon, has been demonstrated to modulate translational efficiency in vertebrates and invertebrates (Kozak, 1987a, Kozak, 1986, Cavener, 1987, Cavener and Ray, 1991); it is thought to mediate translation complex recognition of the start codon.

The translation initiation sequence (TIS) was initially identified as the mechanism for recognition of the appropriate AUG codon by the translation complex (Kozak, 1987a, Kozak, 1986). More recent work has demonstrated that intentional design of the TIS can mediate predictable changes in protein expression from transgenes (Pfeiffer et al., 2012, Sano et al., 2002, Tatematsu et al., 2014, Horstick et al., 2015). This has not yet been shown in non-model species. The TIS is very short (~10nt) and does not affect the activity of the promoter or the sequence of the transgenic protein. For these reasons, and its seeming conservation across vertebrates, Diptera and Lepidoptera, TIS was selected for further exploration in this project as a tool to modulate protein expression.

The phenomenon of multiple start codons in the 5'UTR of a gene is thought to be an aspect of endogenous control of gene expression. Translation complexes have been observed to scan mRNA from 5' to 3' and to not always initiate translation from the 5' most start codon. This 'leaky scanning' theory is suggested as the reason for the importance of the TIS, which demarcates the genuine start codon to the translation complex. Under this theory, the presence of multiple start codons would reduce efficiency of translation initiation, slowing down the entire translation cycle for that mRNA. This is thought to be an endogenous feature for control of gene expression. Without an existing research base for using this as a tool in modulating transgene expression, it was decided not to proceed with the presence of additional start codons as an experimental variable in this project.

Translation termination and recycling

The 3'UTR is more closely associated with mediating the termination of translation and the efficiency of recycling the translation complex so that further translation can occur from the same mRNA. Termination begins with recognition of a stop codon by the translation complex and is modulated by presence and proximity of 3'UTR binding proteins (Schuller and Green, 2018). 3'UTR sequences immediately 3' of the stop codon have been demonstrated to directly impact stop codon recognition, but 3'UTRs have also been found to have regulatory effects at almost every stage of protein expression (transcription, mRNA maturation, nuclear

export, translation, post-translational modification, mRNA stability) (Schuller and Green, 2018, Mayr, 2019).

This complexity and specificity of the 3'UTR suggests that it might be wise to use the 3'UTR of an endogenous gene with the features of interest (e.g. temporal or spatial specificity) when such features are important for a transgene. This project, however, focuses on ubiquitous transgene expression.

When ubiquitous transgene expression is desired, similar to selection of a promoter, virus derived sequences are often used. Simian virus 3'UTR SV40 is commonly used in design of insect transgenes (Brand and Perrimon, 1993, Pfeiffer et al., 2012). Work comparing 3' UTR sequences has noted, however, that SV40 is out performed in lepidopteran cells by a baculovirus 3' UTR (P10) from *Autographa californica* nucleopolyhedrovirus (AcNPV) (van Oers et al., 1999). Pfeiffer et al. (2012) corroborated this finding in *D. melanogaster* and additionally found that there was an enhancing effect of using 3'UTR P10 with an altered translation initiation sequence (TIS).

Empirical data for comparisons of commonly used 3'UTR sequences in insects is sparse, and absent for nuisance and pest species. A comparison of three common 3'UTR sequences is explored in this project, and the relationship between these sequences and specific TIS sequences is investigated.

Translation elongation

Looking now at the coding sequence, it is known that making synonymous[1] changes can affect the efficiency of translation elongation and potentially affect the failure rate of translation, each altering the number of viable peptides resulting from an mRNA (Zhou et al., 2016, Schuller and Green, 2018). The way in which this is addressed presently is through use of 'codon optimisation' programs, which take into account a species' particular codon bias[2] to refine an exogenous coding sequence for expression in the species of interest (Zhou et al., 2016). Of note, this process is more recently thought to affect transcription efficiency as well as translation (Zhou et al., 2016). Codon optimisation was not selected for further exploration

[1] Synonymous changes alter the nucleotide sequence without affecting the amino acids that are ultimately coded for. This takes advantage of codon degeneracy: there are more 3 nucleotide codon combinations than amino acids.

[2] Building on codon degeneracy, it has been found that different species make use of specific codons with different frequencies. This profile is called the 'codon bias'.

as an experimental variable in this project as there are already suitable tools to make use of this effect in transgene design.

Protein stability

Once a protein has been expressed, the amount of effective protein can be reduced if completed proteins are degraded. The rate of degradation will be variable, based largely on the protein sequence that has been expressed. Based on the dependence on protein sequence, protein stability has not been selected as a variable for further exploration in this project.

Translation initiation sequence and 3' UTR as modulators of protein expression

The translation initiation sequence (TIS) and its comparative activity in insect species of interest is the main focus of this project. The comparative activity of common 3'UTR sequences, and their interaction with TIS, is the auxiliary focus of this project. It is aimed to identify sequences that can be used to alter the magnitude of transgene expression, independent of (in order of priority) promoter, coding sequence and species.

Translation initiation sequence (TIS)

Kozak (1986) and Kozak (1987b) described the importance of the translation initiation sequence (the nucleotides immediately 5' and 3' of the start codon) for translation efficiency. Kozak (1987a) went on to describe an eponymous consensus nucleotide sequence for that nucleotide position (5' of the start codon), derived from vertebrate mRNA sequences. Cavener (1987) described a consensus translation initiation sequence for *D. melanogaster* in the same way and later a general non-vertebrate consensus sequence (Cavener and Ray, 1991). These sequences are all described in Table 4. Use of these sequences in transgene design is well established and efforts have been made to further enhance transgene expression by identifying TIS consensus sequences for individual species of interest: *Manduca sexta* (Chang, 1999), *B. mori* (Chang, 1999, Tatematsu, 2014), *S. frugiperda* (Sano, 2002) (all Lepidoptera). Similar work has been done looking at the TIS of viruses used in transgene expression from insect cells (Chang, 1999; Sano, 2002; Pfeiffer, 2012).

Table 4: Translation initiation sequences (TIS) from the literature

Sequence name	**Nucleotide position (relative to ATG)**														**Source**
	-10	**-9**	**-8**	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**+1**	**+2**	**+3**	**+4**	
Kozak sequence		G	C	C	G	C	C	A G	C	C	A	T	G	G	Kozak (1987a)
D. melanogaster consensus							C A	A	A	A C	A	T	G		Cavener (1987)
Non-vertebrate consensus	A	T	A	A	A	T	A C	A	A	C	A	T	G	A G	Cavener and Ray (1991)
M. sexta consensus							C	A	A	A	A	T	G	N	Chang et al. (1999)
B. mori consensus					A	N	C	A	A	A	A	T	G	N	Chang et al. (1999)
D. melanogaster consensus	A	A	N	A	A	N	C A	A	A	A C	A	T	G		Tatematsu et al. (2014)
S. frugiperda consensus	A G	N	C	C T	N	A C	C	A	A C	C G	A	T	G	G A	Sano et al. (2002)
B. mori consensus	A T	A	N	A T	A	T	C	A	A	A	A	T	G	N	Tatematsu et al. (2014)
B. mori least common motif	C	C	N	C G	C	G	N	C T G	G C T	T G	A	T	G	C	Tatematsu et al. (2014)
Abridged Syn21 sequence	A	A	A	A	A	T	C	A	A	A	A	T	G		Pfeiffer et al. (2012)

Sequence name	Nucleotide position (relative to ATG)													Source
	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	
Abridged AcNPV gene consensus (with P10)	A	T	A	T	A	A	C	A	A	A	A	T	G		Pfeiffer et al. (2012)
Abridged initiation codon of EoNPV gene	C	T	A	C	A	A	T	A	C	T	A	T	G		Pfeiffer et al. (2012)
Abridged Omega element from TMV	T	A	C	A	A	T	T	A	C	A	A	T	G		Pfeiffer et al. (2012)

The literature consistently reports that the nucleotide sequences used at the position immediately 5' of the start codon can independently and consistently enhance or diminish transgene expression *in vitro* and *in vivo* (Kozak, 1986, Cavener, 1987, Kozak, 1987a, Chang et al., 1999, Sano et al., 2002, Suzuki et al., 2006, Pfeiffer et al., 2012, Tatematsu et al., 2014). Such work has not been published in pest or nusiance insect species (including local species of interest: culicine mosquitoes and *P. xylostella* diamondback moth). This offers a compelling opportunity to develop and validate TIS from the literature as a tool for modulation of transgene expression in species of interest.

3' Untranslated Region (3'UTR)

The relative efficacy of 3'UTRs commonly used in insect transgenesis is not well characterised, particularly for our species of interest. The experimental format for measuring the effect of different TIS on transgene expression is well suited to characterising relative efficacy of 3'UTR sequences. Furthermore, such experiments would allow the interaction of TIS and 3'UTR noted in *D. melanogaster* by Pfeiffer et al. (2012) to be corroborated (or not) in our species of interest. It was therefore decided to include characterisation of alternate 3'UTR sequences as an auxiliary aim for this project.

Experimental design

Concept

It was decided to conduct these experiments *in vitro*, using cell lines as models of our species of interest. This approach is less resource intensive than *in vivo* work and increased the number of experimental variables that could be considered. The main limitation of cell culture experiments is the unknown fidelity between *in vitro* results and those that would be found in whole organisms. Experiments with TIS in *B. mori* (*in vitro* and *in vivo*) and in *D. melanogaster* (*in vivo*) demonstrate that the principles identified in cell culture are likely to have an acceptable fidelity to what might be found in whole organisms (Tatematsu et al., 2014, Pfeiffer et al., 2012).

Working with embryo injections was considered as it is thought to be a more representative model of whole insects. As experiments would be limited by the extra-genomic nature of the transgene in either scenario, embryo injection was discarded as more resource intensive than cell culture.

In this project, transgene expression is assessed by measuring the quantity of a reporter protein present at a set interval after introduction of the transgene to the cells (transfection). A plasmid vector is used for the transgene and a different plasmid is generated with each TIS and 3'UTR. By using transgene components (e.g. promoter) with cross-species activity, the same set of plasmids can be used in cell lines from multiple species. The relative rate of protein expression between different transgenes (plasmids) can therefore be examined in multiple species and attributed to the TIS and 3'UTR used.

A dual luciferase reporter assay was selected as it is quantitative, reliable, and well established in the community. Using fluorescent reporter proteins was discarded as being less quantitative with the available equipment, though it was used for *in vivo* experiments in the literature (Pfeiffer et al., 2012, Tatematsu et al., 2014).

It was decided that a scope of five TIS and three 3'UTR would be sufficient to give an insight to their effect on transgene expression *in vitro*. This decision was informed by the resource requirement of cloning and testing multiple plasmids concurrently, but also by the high level of sequence duplication noted in the TIS described in the literature (Table 4).

Cell lines

Five cell lines were selected to represent three species of Culicine mosquito and to include a lepidopteran cell line (Table 5). The lepidopteran cell line could be considered as a positive

control, due to the focus of the literature on moths. Multiple cell lines were included to identify species-specific effects of TIS and 3'UTR on transgene expression and to serve as biological replicates where effects are independent of cell line. This scale was made possible by use of transgenes with cross-species activity (no plasmid was designed for one cell line in particular). Although *Anopheles* cell lines are explored in later work, these experiments focused on Culicine mosquitoes.

Table 5: Cell lines used in Chapter 3

Cell line	**Species (tissue)**	**Published origin**
Aag2	*Aedes aegypti* (embryo)	(Peleg, 1968a, Peleg, 1968b)
Hsu	*Culex quinquefasciatus* (ovary)	(Hsu et al., 1970)
C6.36	*Aedes albopictus* (larvae)	(Igarashi, 1978)
U4.4	*Aedes albopictus* (larvae)	(Singh and Pavri, 1967)
Sf9	*Spodoptera frugiperda* (pupa)	(Vaughn et al., 1977)

Plasmids

Transgenes were designed to mimic existing *Aedes* mosquito transgenes as closely as possible, without using components inactive in lepidoptera.

Plasmid backbone – pGL3-Basic (AGG1183, Promega, UK) was selected as a simple, readily available construct with no other eukaryotic expression cassettes.

Promoters – Tissue-specific promoters, such as 3XP3 and carboxy peptidase A promoter, were not considered for these experiments as the cell lines involved do not have a defined tissue type (they are spontaneously immortalised cells originating from a crushed specimen such as an embryo or a larva). Of the available ubiquitous promoter sequences, those derived from viruses (e.g. HR5-IE1) were selected over those of insect origin (e.g. poly-ubiquitin) as they are active in both Diptera and Lepidoptera.

"HR5-IE1" promoter is the *Autographa californica* nuclear polyhedrosis virus (AcNPV) immediate-early 1 (IE1) promoter, with enhancer homologous region 5 (HR5) (Gong et al., 2005, Jarvis et al., 1996, Huynh and Zieler, 1999, Pullen and Friesen, 1995, Guarino et al., 1986, Douris et al., 2006). It is a versatile promoter for insect transfection and transgenesis (Fu et al., 2010, Grossman et al., 2001, Haghighat-Khah et al., 2015, Harvey-Samuel et al., 2020, Martins et al., 2012, Meredith et al., 2013, Wilke et al., 2013). HR5-IE1 was selected as the promoter for the experimental constructs.

"OpIE2" promoter from *Orgyia pseudotsugata* multicapsid nuclear polyhedrosis virus (MNPV) immediate-early gene 2 is a similarly versatile, ubiquitous promoter in insects (Haghighat-Khah et al., 2015, Li et al., 2017, Volohonsky et al., 2015, Hegedus et al., 1998, Pfeifer et al., 1997, Theilmann and Stewart, 1992). It was selected as the promoter for the control luciferase construct.

Intron - It was decided to include an intron sequence as these are typically included in insect transgenes (as they have been demonstrated to increase expression). A locally used intron from *D. melanogaster* gene *alcohol dehydrogenase* was included as "adh".

Translation initiation sequences – Five TIS were designed and selected based on those described in the literature (Table 4). The process is described below and the sequences carried forward are summarised in Table 6.

Kozak sequence (Kozak, 1987a) is the first described TIS and is so ubiquitous as to be considered synonymous with the nucleotide position, 5' of the start codon. It was selected for inclusion, despite its vertebrate origins, as it remains commonly used in insect transgenic constructs (Fu et al., 2010, Martins et al., 2012, Morrison et al., 2012). The nucleotide ambiguity at nucleotide position -3 (5' of the start codon) (Table *4*) was resolved based on the sequences used in other transgenic *Ae. aegypti* and *P. xylostella* at The Pirbright Institute (personal communication, Tim Harvey-Samuel, 2016).

"*B. mori* least common motif" (Tatematsu et al., 2014) is the only described 'down-regulating' TIS in Table 4 and was selected for inclusion as "*B. mori* low". "*B. mori* consensus" from the same paper (Tatematsu et al., 2014) was included as a counterpart; it was resolved using the highly similar sequences: "non-vertebrate consensus" (Cavener and Ray, 1991), "*M. sexta* consensus" (Chang et al., 1999) and each "*D. melanogaster* consensus" (Cavener, 1987, Tatematsu et al., 2014). This sequence was named "*B. mori* high".

The sequences used to resolve "*B. mori* consensus" were then considered to be represented by "*B. mori* consensus" (now "*B. mori* high"). "*S. frugiperda* consensus" (Sano et al., 2002) is the only remaining TIS in Table 4 derived from a non-viral consensus and was selected as the fourth sequence to carry forwards. Ambiguous nucleotides were resolved using "*B. mori* consensus" (Tatematsu et al., 2014) as the closest related sequence. This sequence was named "Lepidopteran".

The virus-derived TIS studied by Pfeiffer et al. (2012) were each described as having equivalent transgene activity. Pfeiffer et al. (2012) chose to continue only with "Syn21", as

the shortest of their experimental sequences and therefore the most practical for oligonucleotide synthesis. “Syn21” was thus selected as the fifth TIS.

A G nucleotide directly 3’ of ATG is known to be critical for optimal expression with “Kozak sequence” (Kozak, 1987a), but is not reported as such in insect TIS (Kozak, 1986). In *D. melanogaster*, *S. frugiperda*, invertebrates and AcNPV, a co-consensus of G and A is reported (Cavener and Ray, 1991, Sano et al., 2002). Tatematsu et al. (2014) found no consensus at this position in *B. mori* and further investigated the effect on expression of different nucleotides at this position, finding no significant difference. It was therefore decided to maintain G as the nucleotide directly 3’ of each TIS in this project.

CDS – A firefly luciferase coding sequence, *luc+* from pGL3-basic (AGG1183, Promega, UK), was selected as the reporter luciferase and remained unchanged throughout the project. Renilla luciferase (RL) was selected as the control luciferase.

Table 6: Five translation initiation sequences used in these experiments

Name	**Sequence**
Kozak (Koz)	GCCGCCACC**ATG**G
Lepidopteran (Lep)	AACCAACAAC**ATG**G
B. mori High (BmHi)	AAAAATCAAA**ATG**G
B. mori Low (BmLo)	CCGCCGGCGT**ATG**G
Syn21 (Syn21)	AACTTAAAAAAAAAAATCAAA**ATG**G

3’ Untranslated region (3’UTR) – Three 3’UTR sequences were selected based on the ubiquity of simian virus 40 (SV40) 3’UTR in the literature and on a hypothesis, based on the literature, of each other 3’UTR giving higher (P10 3’UTR) or lower (K10 3’UTR) transgene expression than SV40.

Simian virus 40 (SV40) early and late polyadenylation signals have been widely used in baculovirus expression systems, as well as plasmid vector expression systems and in insect transgenesis (Vlak et al., 1990, Brand and Perrimon, 1993, Morrison et al., 2012, Pfeiffer et al., 2012, Fu et al., 2010, Gong et al., 2005, Tamura et al., 2000, Grossman et al., 2001). The 3'UTR of AcNPV gene *p10* (p10 3'UTR) has been described as enhancing protein expression (from SV40) in baculovirus expression systems (van Oers et al., 1999, Liu et al., 2015). This was confirmed in *D. melanogaster* by Pfeiffer et al. (2012), particularly in conjunction with their Syn21 TIS.

D. melanogaster female sterile (1) K10 (henceforth referred to as *K10*) encodes an oocyte protein that is localised to the nucleus and originates from a nurse cell (Prost et al., 1988, Serano et al., 1994, Rorth, 1998). This 3'UTR is widely used, particularly when expression in the female germline or early embryo is desired and is thought to confer lower transgene expression than SV40 3'UTR (personal communication, Luke Alphey) (Rorth, 1998, Koch et al., 2009, Fu et al., 2010, Martins et al., 2012, Morrison et al., 2012).

Methods

Cloning

Control luciferase plasmid

In addition to the experimental plasmids expressing reporter protein Firefly luciferase (FF), a second luciferase plasmid (expressing Renilla luciferase (RL)) is included in every experimental condition. Activity from the second luciferase (RL) is measured alongside the luciferase of interest (FF) and then acts as an internal control for variation introduced in the multi-step experimental process (e.g. number of cells in a sample, pipetting error in sample handling). This is known as a dual luciferase assay (Promega, 2015).

Preliminary experiments with pRL-CMV (AGG1079) (Promega, UK) confirmed that the human cytomegalovirus (CMV) promoter is not sufficiently active in insect cell culture and the assay could not be adjusted to read RL activity and FF activity (driven by an insect-active promoter) in the same sample. pRL-OpIE2 (AGG1080) was built from pRL-CMV (Anderson et al., 2020). The baculovirus-derived promoter OpIE2 gave suitable RL activity in preliminary experiments with the plasmid and it was used in all subsequent experiments (Figure 4).

Figure 4: Control luciferase plasmid (pRL-OpIE2, AGG1080). OpIE2 driving Renilla luciferase expression.

Firefly luciferase foundation plasmid (MCS)

A foundation plasmid (pIE1-FF-SV40 (AGG1185)) was devised to facilitate cloning of multiple FF plasmids with altered TIS sequences (Figure 5, Table 9). The foundation plasmid was based on pGL3-Basic (AGG1183) (Promega, UK), which is a promoter-less construct with the expression sequence for Firefly luciferase (*luc+*) and SV40 3'UTR (GenBank Accession Number U47295; https://www.ncbi.nlm.nih.gov/nuccore/U47295.2) (Clark et al., 2016) (Figure 5).

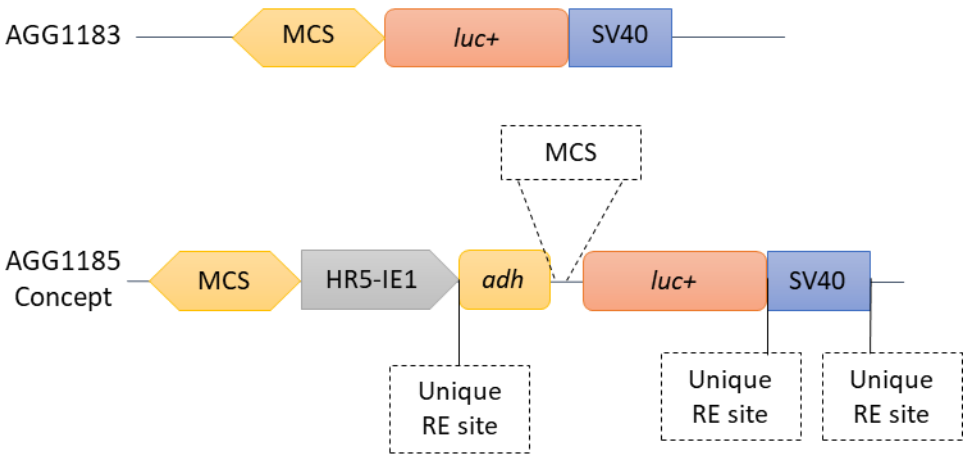


Figure 5: Foundation FF plasmid and pGL3-Basic. pGL3-Basic (AGG1183) is a commercial plasmid with no promoter, expressing *luc+* (FF). AGG1185 was conceived as a foundation plasmid (starting from pGL3-Basic, AGG1183) for cloning FF plasmids with varied TIS and 3'UTR sequences.

pGL3-Basic (AGG1183) has a multiple cloning site (MCS) 5' of *luc+* (FF), which is designed for insertion of a promoter sequence (Figure 5). Unfortunately, it could not be used directly to create foundation plasmid AGG1185 as it contains an NcoI recognition site across the ATG start codon of *luc+*. The NcoI recognition sequence (below) includes a cytosine (C) immediately 5' of the start codon, which is not compatible with all of the chosen TIS (Table 6).

NcoI recognition sequence: $5' \ldots C \mid C\mathbf{ATG}G \ldots 3'$

To resolve this issue, a new MCS was designed using type IIS 'shifted cleavage' restriction enzyme BsmBI to facilitate 'seamless' cloning results in downstream plasmids. The existing MCS of pGL3-Basic (AGG1183) was removed through restriction enzyme digest with HindIII-HF and NcoI-HF (NEB, UK), according to standard protocols. The new MCS – BsmBI-NsiI-AatII-BsmBI-NcoI – was synthesised as a 58bp oligonucleotide (oligo) (LA246). LA246 was designed as a pair with oligo LA233, to PCR amplify HR5-IE1-*adh* from plasmid template AGG1032

designed for constitutive expression of DsRed fluorophore in insect transgenesis (Tim Harvey-Samuel)) (Figure 6).

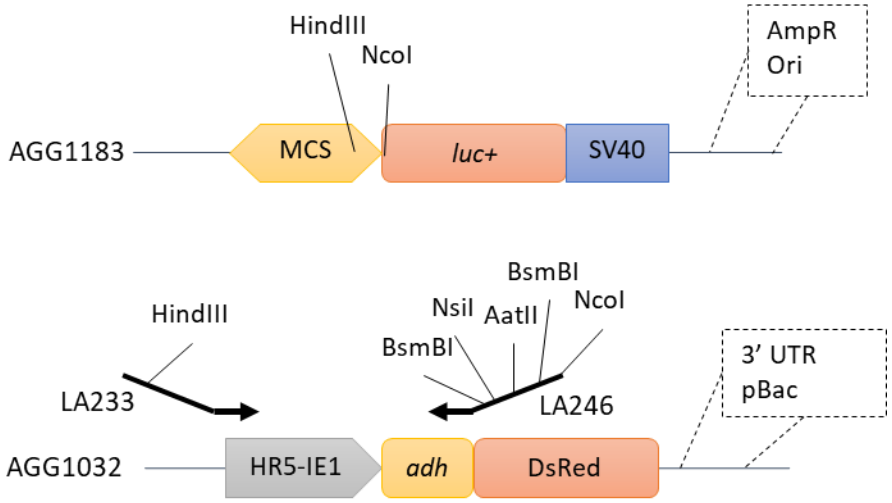


Figure 6: pGL3-Basic (AGG1183), HindIII and NcoI restriction recognition sites indicated. PCR template AGG1032 with primers LA233 and LA246 indicated. HindIII and new MCS (BsmBI-NsiI-AatII-BsmBI-NcoI) indicated as non-complimentary parts of the primers (to be incorporated in the PCR product).

PCR product LA233-LA246/AGG1032 was produced using a variant of touchdown-PCR, with a lowered annealing temperature for the first 10 cycles (accounting for the short portion of oligo LA246 that is complementary to the template) and a higher annealing temperature for the subsequent 20 cycles (once product LA233-LA246/AGG1032 exists as a template in the reaction) (

Table *7*). The PCR product was visually confirmed by agarose gel, then purified and prepared for ligation by digestion with HindIII-HF and NcoI-HF.

Table 7: PCR thermocycle for LA233-LA246/AGG1032

Step	Cycles	Temp. (°C)	Time (s)
Initial denaturation	1	98	60
Denaturation	10	98	10
Annealing		59	20
Extension		72	60
Denaturation	20	98	10
Annealing		67	30
Extension		72	60
Final Extension	1	72	300

Linearised vector pGL3-Basic/HindIII_NcoI was ligated with insert LA233-LA246/AGG1032/HindIII_NcoI and transformed into competent cells using standard methods. PCR colony screening was carried out to identify colonies carrying the desired plasmid (AGG1185, pIE1-FF-SV40), as identified by a 249bp product that spanned the junction of insert and vector (primers LA227 and LA228) (Figure 7).

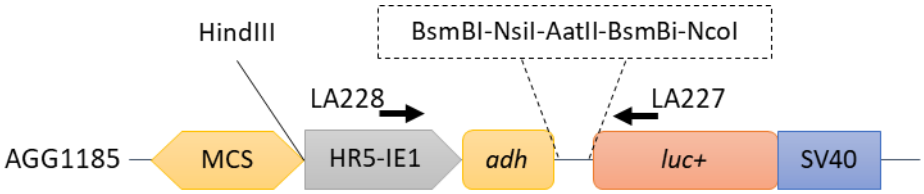


Figure 7: pIE1-FF-SV40 (AGG1185). Colony screening primers LA227 and LA228 indicated.

Two colonies were grown up overnight. Purified plasmid DNA from each colony was verified by Sanger sequencing (primers LA227 and LA228) and a confirmed colony was grown to a larger volume to purify a stock of pIE1-FF-SV40 (AGG1185).

Experimental plasmids: alternative TIS

Foundation plasmid pIE1-FF-SV40 (AGG1185) was designed to allow insertion of a translation initiation sequence (TIS) as an annealed oligo, based on restriction digest with BsmBI.

pIE1-FF-SV40 (AGG1185) was prepared by restriction digest with BsmBI (incubated overnight to achieve reaction saturation), then purified by gel electrophoresis. Dephosphorylation was omitted, in recognition of the insert oligos not being phosphorylated.

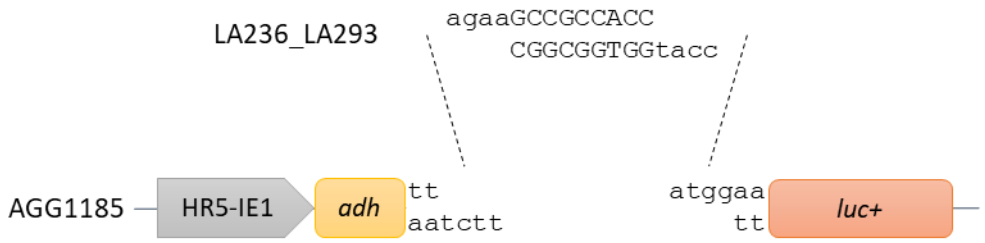


Figure 8: pIE1-FF-SV40 (AGG1185) digested with BsmBI and annealed oligo LA236_LA293 with single strand nucleotide overhangs corresponding to those of AGG1185. Annealed oligos with matching overhangs can be ligated into the prepared vector, AGG1185/BsmBI.

Insert oligos were designed as complementary pairs that could be annealed to produce the desired TIS, with single-stranded DNA overhangs complementary to the BsmBI digested vector (AGG1185/BsmBI). In this way, the method can be repeated to create any number of experimental plasmids from a single linearised vector (Figure 8).

Each oligo pair was annealed by mixing equal (equimolar) amounts of each strand in 1x T4 DNA ligase buffer (NEB, UK), then heating the mixture to 95°C and cooling slowly to room temperature over two hours.

Table 8: Oligos designed to create TIS inserts once annealed

Primer	**Description**	**Sequence (5' to 3')**	**length (nt)**
LA236	Kozak sequence ("Kozak")	AGAAGCCGCCACC	13
LA293		CCATGGTGGCGGC	13
LA237	*S. frugiperda* consensus sequence ("Lep")	AGAAAACCAACAAC	14
LA294		CCATGTTGTTGGTT	14
LA238	*B. mori* consensus sequence ("BmHi")	AGAAAAAAATCAAA	14
LA295		CCATTTTGATTTTT	14
LA239	*B. mori* least common motif ("BmLo")	AGAACCGCCGGCGT	14
LA296		CCATACGCCGGCGG	14

Primer	Description	Sequence (5' to 3')	length (nt)
LA240	Synthetic sequence 'Syn21' ("Syn21")	AGAAAACTTAAAAAAAAAAATCAAA	25
LA297		CCATTTTGATTTTTTTTTTTAAGTT	25

Linearised vector pIE1-FF-SV40/BsmBI was ligated with each insert (Table 8) according to standard methods. To mitigate potential inefficiencies of the BsmBI digestion and to reduce presence of viable (circular) parent plasmid before transformation, a post-ligation digest was done with NsiI-HF (NEB, UK), which is uniquely present in the MCS region removed from pIE1-FF-SV40 by digesting with BsmBI. Each ligation reaction was diluted to 20µl in CutSmart buffer (NEB, UK) and 1U of NsiI-HF was added. The digestion reaction was incubated at 37°C for 1hr before 4µl was carried forward to transform competent cells without further purification.

Colonies were grown up and plasmid prepped before screening by restriction enzyme digest to distinguish between desired (daughter) plasmids and pIE1-FF-SV40 (parent plasmid). Preparations of the daughter plasmids were verified by Sanger sequencing using primers LA227 and LA228.

This method was used to create pIE1-Koz-FF-SV40 (AGG1186) (Genbank accession no. MT119956 (Tng et al., 2020)) (Figure 9), pIE1-Lep-FF-SV40 (AGG1187), pIE1-BmHi-FF-SV40 (AGG1188), pIE1-BmLo-FF-SV40 (AGG1189) and pIE1-Syn21-FF-SV40 (AGG1190).

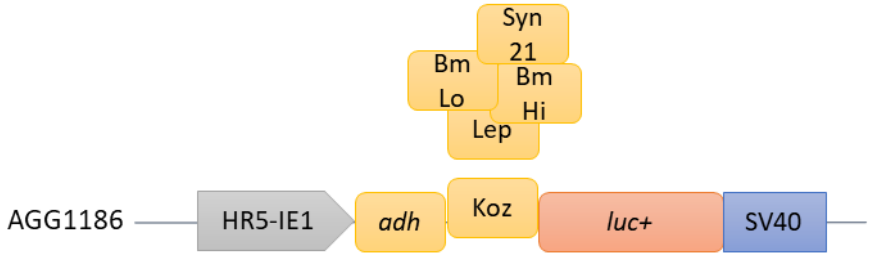


Figure 9: FF plasmid pIE1-Koz-FF-SV40 (AGG1186). Alternate translation initiation sequences are illustrated as blocks above the relevant section of the construct.

Experimental plasmids: alternative 3' UTR

Each FF plasmid has unique restriction enzyme recognition sites 5' (XbaI) and 3' (SalI) of the SV40 3'UTR sequence (inherited from pGL3-Basic); this configuration is then present in daughter plasmids, including AGG1186 (Figure 10). XbaI and SalI-HF were used to create linearised vectors of each daughter plasmid (AGG1186 – 1190), using standard methods.

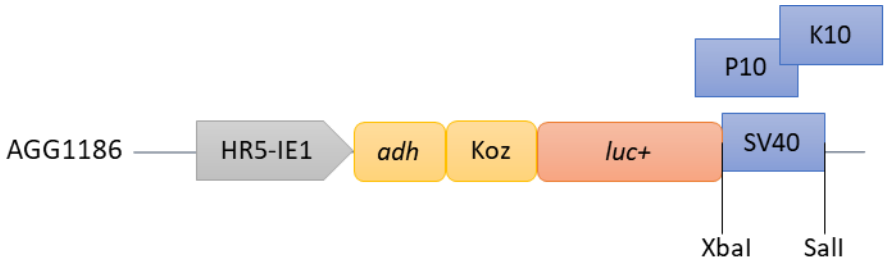


Figure 10: FF plasmid pIE1-Koz-FF-SV40 (AGG1186) with restriction sites XbaI and SalI indicated. Alternative 3'UTR are illustrated as blocks above the relevant section of the construct.

3'UTR 'Inserts' were created by PCR amplification of the desired 3'UTR sequence from a plasmid template, using the primers to incorporate XbaI and SalI recognition sites (Figure 10). Once digested with XbaI and SalI-HF, 'inserts' could be ligated to each linearised vector, creating a panel of 15 experimental plasmids (Table 9).

K10 3'UTR was amplified from plasmid template pBac-ZsG-A4-tTAV3 (AGG1019, Sanjay Basu), using primers LA444 and LA445. P10 3'UTR was amplified from plasmid template pT7-P10 (Promega, UK). Primers LA392 and LA393 were used.

Table 9: Panel of experimental plasmids with promoter HR5-IE1

		3'UTR sequence		
		SV40	**P10**	**fs(1)K10**
Translation initiation sequence	**Kozak**	AGG1186	AGG1196	AGG1191
	Lepidopteran	AGG1187	AGG1197	AGG1192
	B. mori High	AGG1188	AGG1198	AGG1193
	B. mori Low	AGG1189	AGG1199	AGG1194
	Syn21	AGG1190	AGG1200	AGG1195

Control plasmid: luciferase-null (HR5-IE1-ZsGreen)

pHR5-IE1-ZsGreen (AGG1201) (Figure 11) was designed as a not-luciferase-expressing control for preliminary experiments, to standardise the nucleic acid content of transfection mixtures (e.g. to replace the molecular weight of the Renilla luciferase expressing plasmid if an experimental condition called for transfection of only Firefly luciferase expressing plasmid). It was additionally used in experiments assessing efficacy of different transfection reagents, and as a visible transfection efficiency marker for quality control checks while protocols were being iterated and improved (Genbank accession no. MT119955, (Tng et al., 2020)).

Figure 11: Luciferase-null, control plasmid HR5-IE1-ZsGreen (AGG1201). Plasmid with fluorophore ZsGreen (ZsG) expression under promoter HR5-IE1, no luciferase expression.

pHR5-IE1-ZsGreen (AGG1201) was cloned from pBac-ZsGreen-tTAV3 (AGG1024), a pre-existing moth transformation marker construct. Unique restriction enzyme sites for SacII and PacI were used to remove the tTAV3 cassette, due to concerns of cytotoxicity of tTAV3. The linearised vector was blunted and then blunt end ligated using CloneJET PCR cloning kit, to manufacturer's recommendations. pHR5-IE1-ZsGreen was otherwise cloned using standard methods.

Transfection

Transfections were carried out according to standard methods, in a 96-well plate format and using a transfection master-mix wherever possible. Repeats were generated by using the same master-mix to transfect 8 wells (1 column) of cells with the same transfection condition (12 transfection conditions per plate). Each cell line is represented on independent plates, to minimise risk of error (e.g. cross contamination, labelling error or confusion), but master-mix was used across cell lines where possible. A single 'experiment' is defined as transfections that happened on the same day, regardless of transfection conditions or cell lines. Different 'experiments' therefore have different flasks of cells, cell seeding on different days, different transfection master-mixes and transfections on different days.

The amount (ng) of plasmid transfected was measured as ng/well (of cells) and varied by cell line (which have different transfection efficiencies and promoter efficiencies) (Table *10*).

Table 10: Transfection amounts for each plasmid in each cell line

	Aag2	**C6.36**	**Hsu**	**Sf9**	**U4.4**
FF plasmid (ng/well)	1	1	1	5	1

	Aag2	C6.36	Hsu	Sf9	U4.4
RL plasmid (ng/well)	50	5	5	1	5
TransIT Pro reagent (ul/well)	0.2				
TransIT Boost reagent (ul/well)	0.1				

Dual luciferase assay

Transfected cells were harvested and read by dual luciferase assay, according to standard protocols. The volume of cell lysate processed by dual luciferase assay could be varied for each cell line (Table 11), to confirm that samples were measured within the dynamic range of the assay. The preliminary work to determine sample volumes is discussed in Appendix C.

Table 11: Sample lysate volumes used for dual luciferase assay for each cell line

	Lysate volume (µl)				
	Aag2	C6.36	Hsu	Sf9	U4.4
Experiment 3	1	1	1	1	7

Analysis

Dual luciferase assay output is FF and RL activity for each sample, given in arbitrary light units (ALU), in a Microsoft Excel compatible format. GraphPad Prism was used in preliminary experiments to graph and analyse control transfection conditions for each experiment and the RL background threshold was calculated in Microsoft Excel (discussed in Appendix C). Data quality control included screening sample luciferase activity against the RL background threshold and the FF quenching threshold using the "highlight cells" function in Microsoft Excel. Any aberrant FF or RL activity led to the sample being excluded from the data set, and such values are not shown in Figure *12*.

For statistical analysis, a generalised linear model was kindly designed and developed by Phil Leftwich, who produced the graphs and tables shown from Figure 12, using R Studio (RStudio Team 2019; v1.4.1106) in R version 3.6.2 (R Development Core Team).

The full terms of the model are included in Appendix C. In brief, the generalised linear model has a log-link function and a Gamma family distribution to account for increasing variance with the mean. The TIS, 3'UTR and cell line are included as categorical predictors, along with all two-way interactions (there were insufficient repeats in the data set to allow for analysis of three-way interactions). Because of the underlying non-normal distribution, a DHARMa package and a simulation-based approach were used to produce interpretable results. Figure 12 and Figure 13 were generated using 'ggplot2' (Wickham, 2016) and 'patchwork'

(Pedersen, 2020). Data sets were summarized using 'tidyverse' (Wickham et al., 2019) and 'emmeans' (Lenth, 2020).

Results for different cell lines are graphed independently. This is visually useful to allow appropriate y-axes scales but is predominantly done to highlight that FF/RL values cannot be compared between cell lines. Different cell lines, particularly from different species, have their own (unmeasured) promoter efficiencies for each HR5-IE1 (in the FF plasmid) and OPIE2 (in the RL plasmid). The ratio in efficiency between HR5-IE1 and OPIE2 is not constant between cell lines. This effect cannot, therefore, be separated from the effect of experimental conditions on FF/RL. Instead, trends between experimental conditions can be compared between cell lines (e.g. fold change in activity).

Blank Page

Results and Discussion

A series of preliminary experiments were carried out to validate the experimental method and data quality control aspects, these are described in Appendix C.

Panel of fifteen plasmids representing five TIS and three 3'UTR sequences, in five insect cell lines

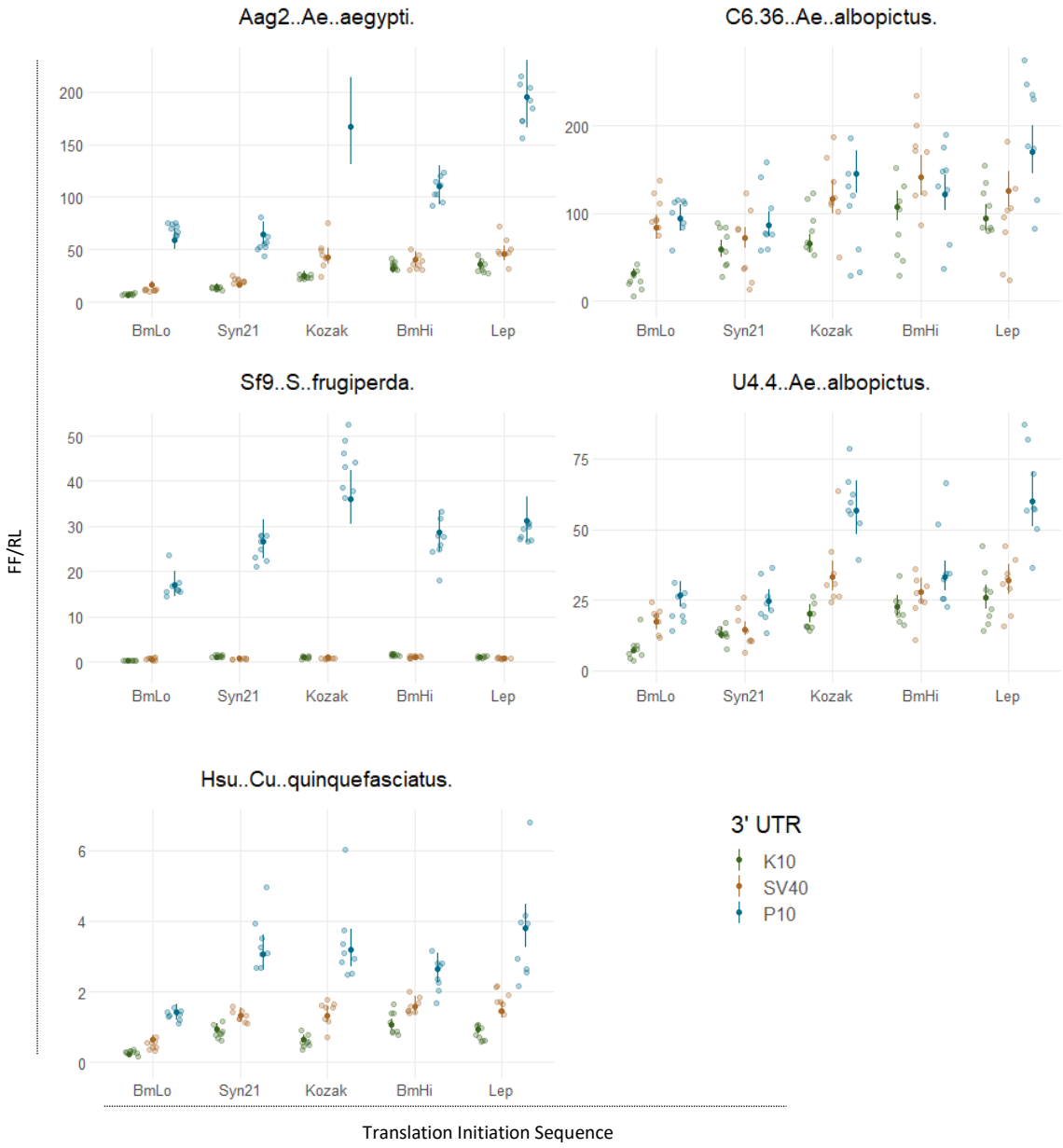


Figure 12: Graphs showing experimental results overlaid with predictions from the statistical model used for analysis. Results are shown for transfection of fifteen different FF expressing plasmids in five insect-origin cell lines. The FF/RL (y-axis) value for each sample is shown as a faded circle, separated by translation initiation sequence (TIS) on the x-axis and by 3'UTR sequence in interleaved colours. Data from each cell line is represented on independent graphs with independent y-axis scales as direct comparisons of FF/RL cannot be made between cell lines. Mean and 95% confidence interval (CI) for each combination of cell line, TIS and 3'UTR have been predicted by a customised generalised linear model and are shown overlaid as a solid circle and line. By comparing the spread of the actual data against the prediction of the model, we can assess the fit of the model. By plotting the results (predictions) of the model we can visually assess the differences in transgene expression attributable to TIS and 3'UTR (if the 95% CI does not overlap than the means are significantly different). Finally, these graphs offer a visual assurance of the actual results before moving into the necessarily abstract analysis based on predictions of the model. Data for Aag2:Kozak:P10 was not generated (operator error), but the model's predicted results are shown.

A dual-luciferase assay experiment with fifteen firefly expressing plasmids (five TIS and three 3'UTR) was carried out in five cell lines concurrently (Figure *12*). Each transfection was repeated in 8 wells of cells, using master mixes wherever possible. This data set was analysed through use of a customised generalised linear model (kindly carried out by Phil Leftwich).

Before analysis, results were quality controlled using a protocol developed across preliminary experiments (Appendix C). In this process, measurements of firefly luciferase (FF) and Renilla luciferase (RL) activity were individually screened for each sample. FL expression above 10^6 arbitrary light units (ALU) was demonstrated to be incompletely quenched in this version of the dual luciferase assay. This resulted in 'bleed through' to the RL measurement for that sample, which would skew a FF/RL standardisation. Samples with FL expression above 10^6 ALU were therefore excluded. A RL minimum threshold was calculated in order to exclude background expression that could be measured in the absence of a RL expressing plasmid (Appendix C). This threshold was derived from control transfections in each cell line and any samples whose RL activity did not exceed this threshold were excluded from further analysis so that they did not skew the FF/RL standardisation.

Samples that passed quality control were then each transformed by FF/RL to generate the values shown as transparent circles in Figure 12. These are the values that were then analysed by statistical model.

The custom generalised linear model (the model) uses three variables ("cell line", "TIS" and "3'UTR") and each two-way interaction (e.g. "Cell line: TIS") as categorical predictors. Three-way interactions were not included as the data set lacks sufficient power (number of repeats) for such analysis. The intercept (value against which each other was compared) was set using Aag2:Kozak:SV40 (Cell line:TIS:3'UTR). Further details are described in the methods section (page 53) and the full detail of the model is presented in Appendix C.

To accommodate the non-normal distribution of the data, a simulation-based approach was used to produce interpretable results in the form of estimated means (opaque circles) and 95% confidence intervals (opaque vertical lines) for each combination of variables (Cell line:TIS:3'UTR). These are plotted in Figure 12 as the "predicted results" (opaque), overlaying actual results (transparent) (which is the same data that was used to build the model). Looking at Figure 12 we can see that the predicted values properly overlay the actual results in all but two conditions - Aag2:Kozak:P10 where there is no actual data set (excluded due to operator error[3]) and SF9:Kozak:P10 where the model under-predicts actual results. Numerical analysis of the goodness of fit of the model is described in Appendix C.

[3] The plasmid was omitted from the transfection mixture by accident. Based on evidence from preliminary experiments and on the reliability of the statistical model, it was decided to not repeat this panel experiment to rectify the error.

Analysis of the effect of each variable and two-way relationship on FF/RL

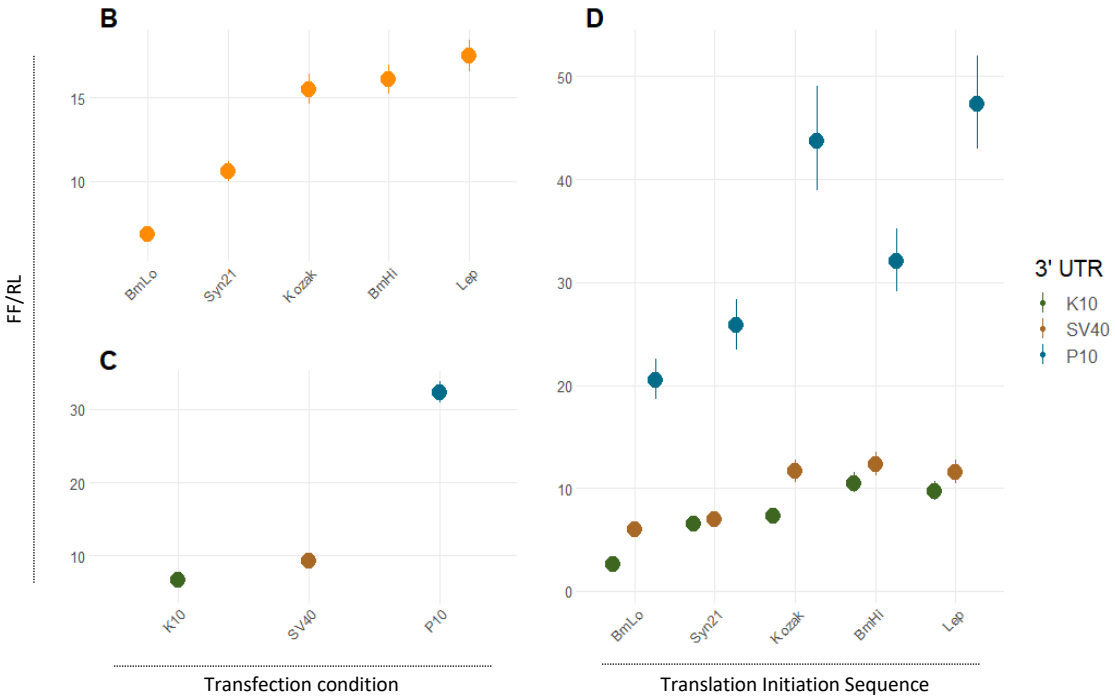


Figure 13: Graphs of predicted results for individual variables. The generalised linear model can show results for a single variable or pair of variables, by holding the remaining variable(s) constant. Panel B shows results based on TIS (cell line and 3'UTR held constant). Panel C shows results based on 3'UTR (cell line and TIS held constant). Panel D shows results from TIS and 3'UTR (cell line held constant). Each graph shows FF/RL on the y-axis with different scales. Each sequence within the examined variable is shown on the x-axis; in panel D the TIS sequences are shown on the x-axis and 3'UTR sequences are differentiated by colour. These graphs show the independent effect of each variable(s) on FF/RL, which is used as a measure of translation efficiency.

Accepting the model as fit for purpose, it can be used to generate data sets showing FF/RL results with specific variable(s) held constant and results therefore affected only by the remaining variable(s). For example, it can show the effect on FF/RL of different TIS sequences (with "cell line" and "3'UTR" held constant) or the effect on FF/RL of different combinations of TIS and 3'UTR sequences (with "cell line" held constant). Such results are shown graphically in Figure 13 and presented in full in Appendix C.

Figure 13 shows graphical results (FF/RL) of analysing two variables independently ("TIS" in Panel B, "3'UTR" in Panel C) and their pair-wise interaction (Panel D). Analysis of the effect of "cell line" on FF/RL is excluded. With the understanding that FF/RL values could not be compared between cell lines due to differential (and unquantified) expression from the promoters on each luciferase plasmid, the experimental design then uses different amounts of FF and RL plasmid in the transfection for each cell line. There is also the opportunity to use

different cell lysate volumes in the dual luciferase assay. These choices modulate luciferase results to sit within the dynamic range of the assay but obscure any trends that could otherwise have been gleaned from analysing the impact of "cell line" alone on FF/RL. For this reason, the summary data for the impact of "cell line" on FF/RL is not shown and it is not discussed further.

Figure 13 is a summary representation of data that is described in full in Appendix C, this data is generated through simulations using the model. The circles represent estimated mean and, where visible, the vertical lines indicate the 95% confidence interval (CI) of that estimated mean. Due to the nature of this (simulated) data, it is true that for any two estimated means where the 95% CI do not overlap, there is a statistically significant difference between those estimated means.

To look at the effect of a single variable on FF/RL (transgene expression efficiency), the other variables are each held constant to a theoretical average (e.g. average "cell line" and average "3'UTR" in order to examine different "TIS" sequences). To look at pair-wise interactions, the third variable is held constant and the effect of each combination of variable one and variable two on FF/RL is examined for any effect beyond that expected of each variable on its own (e.g. is the increase in expression brought about by Lep:P10 greater than the sum of the effects attributable to each Lep and P10 independently?).

Variable: Translation Initiation Sequence

To look at the variable translation initiation sequence ("TIS") in isolation, variables "3'UTR" and "cell line" were each held constant (to their average across the whole dataset). The results of this are shown in Figure 13 Panel B, with FF/RL on the y-axis and TIS sequence on the x-axis. The estimated mean of FF/RL is shown as a solid circle and the 95% CI is shown as a vertical line.

Looking at the mean values for each TIS, and where their 95% CI overlap, we can derive a rank order for the impact of TIS sequence on FF/RL:

$$\text{BmLo} < \text{Syn21} < \text{Kozak} \leq \text{BmHi} \leq \text{Lep}$$

By asking the model to contrast each combination of pairs of TIS sequences, we can produce ratios of activity that have properly accounted for the error. These are shown in Table *12* as a 95% CI for the mean ratio of the comparison; where the 95% CI of the mean does not span 1.00, the estimated mean ratio is given.

Table 12: Summary results of comparisons from Figure 38 panel B. Comparisons of FF/RL results for each TIS sequence are shown. "3'UTR" and "Cell line" are held constant.

Comparison	95% CI	Mean Ratio
Kozak / BmHi	0.86 – 1.08	ns
Kozak / BmLo	2.02 – 2.54	2.27
Kozak / Lep	0.79 – 0.99	0.89
Kozak / Syn21	1.31 – 1.64	1.46
BmHi / BmLo	2.11 – 2.62	2.35
BmHi / Lep	0.83 – 1.03	ns
BmHi / Syn21	1.36 – 1.69	1.52
BmLo / Lep	0.35 – 0.44	0.39
BmLo / Syn21	0.58 – 0.72	0.65
Lep / Syn21	1.48 – 1.84	1.65

Using these comparisons, we can say that each TIS sequence has a statistically significant difference in effect (on transgene expression) than each other TIS sequence, with two exceptions. The comparison of Kozak and BmHi shows that their mean ratio is not different from 1.00 – they do not have different effects on transgene expression. The same is seen for the comparison of BmHi and Lep, indicated in both cases by "ns" in place of a mean ratio. Although BmHi has the same effect as either Kozak or Lep, Kozak and Lep are statistically different from one-another.

These results confirm that the TIS sequence of a transgene affects the efficiency of protein expression (how much protein can be generated in a fixed time), in line with the literature (Kozak, 1987b, Cavener, 1987, Sano et al., 2002). Although other studies have pinpointed this effect to changes at translation initiation (Kozak, 1986, Chang et al., 1999) such specificity is outside of the scope of this project. Looking at different TIS sequences within this experiment, we can assess how much change can be made to the efficiency of protein expression (as measured by amount of reporter protein generated in a fixed time). Looking first at the greatest change, there is an increase of 2.56 fold from TIS "BmLo" to TIS "Lep". This is a smaller change in expression efficiency than is reported in the literature. Pfeiffer et al. (2012) found a 7.5 fold change between their standard construct (7nt BmHi sequence) and Syn21 (their best performing sequence), expressing GFP in neuronal tissue in *D. melanogaster* (*in vivo*). Tatematsu et al. (2014) found a 10 fold increase in activity when they looked at TIS sequences BmLo and BmHi in a *B. mori* cell line (a dual luciferase assay). Of interest, they found that the magnitude of change was mediated by coding sequence (only a 4 fold change when expressing eGFP) and by tissue when they did *in vivo* experiments (15x, 4x, 48x in different tissues, eGFP).

As this project looks only at a single reporter protein in cell lines, it is limited in the comparisons that can be drawn with the literature. Another way to look at these results is with TIS "Kozak" as the benchmark against which other TIS sequences are compared. The ubiquity of "Kozak" in insect transgenesis (excepting *Drosophila* constructs) makes this a useful, practical perspective. With the aim of increasing transgene expression, TIS "Lep" can be used in insect cell lines to increase reporter expression (from TIS "Kozak") by 1.12x. TIS "BmLo" can be used to decrease reporter expression to 0.44x from "Kozak". These results are independent from cell line identity across those used in this project, representing three mosquito species (*Ae. aegypti*, *Ae. albopictus* and *C. quinquefasciatus*) and one moth species (*S. frugiperda*). This speaks to the versatility of using validated TIS sequences to modulate protein expression efficiency, corroborating the literature in flies (*D. melanogaster*) and commercially relevant moths (*B. mori*), as well as work in vertebrates (zebrafish) (Pfeiffer et al., 2012, Horstick et al., 2015, Tatematsu et al., 2014).

Given the range of changes in expression efficiency mediated by TIS sequences reported by Tatematsu et al. (2014) (4 to 10 fold change *in vitro* with different reporter proteins; 4 fold change *in vitro* to 47 fold change *in vivo* with the same transgene), the results from this project cannot be used to predict the specific amount of change obtainable under other conditions (e.g. different coding sequence or *in vivo* expression). No further work was done with alternate reporter proteins or *in vivo*.

Going back to the rank order of expression efficiency of TIS sequences, we can examine relative performances of each sequence against expectations derived from the literature.

$$\text{BmLo} < \text{Syn21} < \text{Kozak} \leq \text{BmHi} \leq \text{Lep}$$

Starting from the left, the performance of TIS "BmLo" is in line with expectations held from the start of the project. This sequence was designed by (Tatematsu et al., 2014) as a consensus of least common nucleotides at each locus in the TIS in *B. mori*; it is the TIS sequence against which they referenced fold changes in expression caused by other TIS sequences. Knowing that there is adequate transgene expression in mosquitoes (*in vivo*) using vertebrate derived TIS "Kozak", it was expected that effects seen in other insect species (*B. mori* in this case) would also occur in the cell lines in this experiment. TIS "Syn21" is the unexplained exception to this.

Pfeiffer et al. (2012) designed TIS "Syn21" to increase expression efficiency in *D. melanogaster* by combining elements of baculovirus *Malacosoma neustria* nucleopolyhedrovirus (MnNPV) with a *D. melanogaster* consensus TIS sequence and report a 7.5 fold increase in expression (*in vivo*) as compared to their usual TIS (the same *D. melanogaster* consensus sequence, which is a truncated version of TIS "BmHi"). In this project, TIS "Syn21" consistently reduced expression efficiency as compared to either TIS "Kozak" or TIS "BmHi" (0.68 fold and 0.66 fold respectively). This outcome was also seen in each preliminary experiment (including the two discussed in this chapter) and in all five cell lines examined. Once more leaning on the vertebrate nature of the Kozak sequence, it was predicted that the insect- and baculovirus-based sequence would increase expression efficiency – especially as it had already been shown to do so in other Diptera (an order that includes *D. melanogaster* and the Culicine mosquitoes).

Although changes in magnitude of effect on expression efficiency are reported in the literature, there are no reported instances of the direction of change being affected by *in vitro* vs *in vivo* or by changes in coding sequence (Tatematsu et al., 2014). Furthermore, efficacious TIS sequences appear to have a conserved function between vertebrates and invertebrate insects (i.e. the adequate function of TIS "Kozak" in transgenic insects). In attempting to explain the unexpected effect of TIS "Syn21" on expression efficiency, all design and practical protocols were re-verified (including plasmid sequences) with no apparent flaws. There is no indication that the baculovirus from which TIS "Syn21" is partially derived has a host restriction (*Malacosoma neustria* nucleopolyhedrovirus (MnNPV), host "invertebrates" (NCBI:txid38012; accessed 07/2021; Schoch et al. (2020)). The intron sequence used in this project was *alcohol dehydrogenase* intron (*D. melanogaster* origin), which is different from the sequence used by Pfeiffer et al. (2012), *myosin heavy chain* intron (also of *D. melanogaster* origin). The changes in intron and in experimental species (*D. melanogaster* rather than mosquito or moth) are suggested as possible causes for the differences seen between results of this project and those reported by Pfeiffer et al. (2012), though no specific mechanism is known. Given the unpredictable behaviour of Syn21 between *D. melanogaster* and culicine mosquitoes (despite consistency between culicine cell lines and a lepidopteran cell line), it would be sensible to validate a TIS in a representative cell line before using it for transgenics of a new species (e.g. anopheline mosquitoes).

The relative efficacy of TIS "Kozak" is not specifically discussed as it was selected as the familiar standard against which to compare efficacy of other TIS sequences.

TIS sequences "BmHi" and "Lep" gave the greatest increase in expression efficiency in this project. These are both consensus sequences derived from Lepidopteran data, "BmHi" is the most common consensus from *B. mori* and "Lep" is the most common consensus from *S. frugiperda* (with ambiguities resolved using "BmHi") (Table 6) (Tatematsu et al., 2014, Sano et al., 2002). Being insect sequences, "BmHi" and "Lep" were expected to have greater expression efficiency than "Kozak" in insect cell lines, particularly in cell line Sf9 which is of *S. frugiperda* origin. It is notable that the efficiency of "Kozak" is not significantly different from "BmHi" and that a small (1.12 fold) change from "Kozak" is mediated by "Lep". Together, these suggest that much of the expression efficiency that can be obtained through altering the TIS is already achieved by use of the Kozak sequence. Though reassuring to those already using the Kozak sequence in transgenes, it is notable that a vertebrate sequence is so successful in insects.

Variable: 3'UTR

Moving on to the 3'UTR sequence, variables "cell line" and "TIS" were held constant. This data is shown graphically in Figure *13* panel C with 3'UTR sequence on the x-axis and FF/RL (as a proxy for translation efficiency) on the y-axis. The estimated mean and its 95% CI are shown. Where the 95% CI of two groups do not overlap, the means are significantly different from one another. This data shows a rank order of efficacy of 3'UTR sequences used in this experiment:

$$K10 < SV40 << P10$$

As with TIS, the model can be used to contrast each combination of pairs of 3'UTR sequences and produce ratios of activity that fully account for error. This data is summarised in Table 13. As all comparisons are significantly different from 1 (no difference in activity between the two sequences), data is reported as mean ratio and standard error.

Table 13: Summary results of comparisons of FF/RL of different 3'UTR sequences (Figure 38 panel C). "TIS" and "Cell line" are held constant

Comparison	**Mean Ratio**	**SE**
SV40 / K10	1.41	0.04
SV40 / P10	0.29	0.01
K10 / P10	0.21	0.01

As noted, Table 13 shows that the effect of each 3'UTR sequence on FF/RL is significantly different from that of each other 3'UTR sequence. There is a comparatively moderate change in activity between 3'UTR K10 and SV40, a 1.41x increase and a greater change in activity

from SV40 to P10 with a 3.45x increase (1/0.29). The rank order (K10 < SV40 << P10) is then reflected in the change from K10 to P10, a 4.76x increase (1/0.21). The amount of change in FF/RL achieved by changing 3'UTR sequence is of the same order of magnitude as changes achieved with different TIS sequences. The change of 3'UTR to P10 confers the biggest single change in FF/RL of any individual change to either TIS or 3'UTR sequence. These results are congruent with expectations based on the literature.

Although the mammalian (virus) origin of SV40 3'UTR is incongruent with its relative activity in insect cell lines, it is widely accepted as a suitable 3'UTR for coding sequence termination in insect transgenesis (Vlak et al., 1990, Tamura et al., 2000, Grossman et al., 2001, Gong et al., 2005, Fu et al., 2010). The decrease in FF/RL mediated by K10 3'UTR (from *D. melanogaster fs(1)K10*) corroborates anecdotal observations (personal communication, Luke Alphey and Tim Harvey-Samuel). While a 3'UTR derived from a closer evolutionary species might be expected to increase translational efficiency, this does not seem to outweigh the evolutionary advantage of viral sequences evolved to maximise parasitic use of host cell machinery.

The increase in FF/RL mediated by 3'UTR P10 is congruent with both its origin (baculovirus) and its reported activity in *D. melanogaster* (Pfeiffer et al., 2012). Pfeiffer et al. (2012) found a 17 fold increase in activity as compared to 3'UTR SV40, *in vivo*, in *D. melanogaster*.

Interaction: TIS and 3'UTR

By holding variable "cell line" as constant, the interaction between "TIS" and "3'UTR" can be examined. Data for this interaction is presented in Table 14 as the summary results of Figure *13* panel D. These are presented as "absolute values" of FF/RL rather than as ratios in comparison with a reference, with 95% CI indicated in brackets. As with the other model-predicted data sets, where there is no overlap of the 95% CI there is a statistically significant difference between two groups.

Table 14: Summary results for each interaction of TIS and 3'UTR. Shown as estimated means (FF/RL) with 95% CI. "Cell line" is held constant

		Estimated mean FF/RL (95% CI)		
		3'UTR		
		SV40	**K10**	**P10**
TIS	**Kozak**	11.62 (10.56-12.78)	7.33 (6.66-8.06)	43.68 (38.91-49.03)
	BmHi	12.30 (11.18-13.53)	10.50 (9.54-11.55)	32.04 (29.12-35.25)
	BmLo	6.03 (5.48-6.64)	2.59 (2.35-2.85)	20.49 (18.63-22.55)
	Lep	11.57 (10.52-12.73)	9.70 (8.81-10.67)	47.32 (43.01-52.06)
	Syn21	7.03 (6.39-7.73)	6.54 (5.94-7.19)	25.81 (23.46-28.39)

Looking at Figure 13 panel D and Table 14 we can examine the interaction of TIS and 3'UTR on FF/RL. This is of interest as Pfeiffer et al. (2012) identified a synergistic interaction between TIS Syn21 and 3'UTR P10. Looking at Figure *13* panel D in comparison with panel C (page 59), there is no obvious change in the activity of 3'UTR sequences with and without TIS as a variable. For TIS (Figure *13* panel D and panel B) there is some change to the relationship between TIS sequences as 3'UTR changes, but the rank order is largely unaffected, apart from 3'UTR P10 where activity of TIS BmHi drops below that of Kozak and Lep.

Using Figure *13* the TIS and 3'UTR sequence combination with the lowest estimated mean (BmLo:K10, 2.59) can be compared with the sequence combination with the highest estimated mean (Lep:P10, 47.32). This comparison yields a value for the single largest change in efficiency of protein expression mediated by changing the TIS sequence, the 3'UTR sequence or both: 18.27x.

Conclusion

Establishment of a reliable assay

Simultaneous assessment of two variables ("TIS" and "3'UTR"), their interaction and each of their interactions with a third variable "cell line" mandates a high degree of confidence in the protocols and assays used to generate results. Although the dual luciferase assay format was well established (generally and locally), a degree of inherent variability was identified in preliminary experiments that could not be resolved by user discipline or by careful selection of reagents and materials. To generate the power needed to accommodate for this confounding variation at the resolution of 2-fold changes, a novel format (locally) was generated using 96-well plates to vastly upscale sample size and take advantage of a local quirk using 1 in 10 reagent dilutions to decrease cost per sample. Once this assay had been validated (Appendix C), it proved to be a highly replicable and consistent format for this and other experiments (Chapter 4, 5).

Working with a panel of fifteen firefly luciferase (FF) plasmids spanning five translation initiation sequences (TIS) and three 3'UTR sequences, transfection of multiple cell lines produced remarkably consistent rank order effects. There is significant variation of TIS and 3'UTR activity dependent on cell line, but for TIS BmLo ≤ Syn21 < Kozak ≤ BmHi ≤ Lep and for 3'UTR K10 < SV40 << P10 was found to be true in four mosquito and one moth (*S. frugiperda*) cell lines in every experiment and preliminary experiment.

With sophisticated statistical analysis, kindly provided by Phil Leftwich, we were able to determine significance for these effects and for the interaction of each pair of variables ("TIS", "3'UTR" and "cell line"). Corroborating the literature, we confirmed a synergistic relationship between "TIS" and "3'UTR" (Pfeiffer et al., 2012). Paired with this dual luciferase assay format, this analytical method was conserved for future experiments (Chapter 4).

Changing the translation initiation sequence (TIS)

Broadly in line with the literature available in arthropods, we found that altering the TIS sequence produced a significant change in protein expression, which is attributed to changes in translational efficiency (Chang et al., 1999). All change attributed solely to changing TIS sequence was in the range of 1 to 2.6 fold. This was lower than expected from the literature (4 to 10 fold reported *in vitro* by Tatematsu et al. (2014)).

The hypothesis that species specificity (or match) would produce the greatest translation efficiency was corroborated in part by results showing that invertebrate TIS sequences (BmHi

and Lep) tended to outperform a vertebrate consensus TIS sequence (Kozak) in insect cell lines. Similarly, TIS BmLo was consciously designed by Tatematsu et al. (2014) to minimise translational efficiency in an insect species (lepidopteran *B. mori*) and it has performed as such in these experiments. The activity, or lack-thereof, of TIS Syn21 was contrary to expectations.

Syn21 was designed by Pfeiffer et al. (2012) as a hybrid TIS sequence based on the *D. melanogaster* consensus TIS (which is nearly identical to BmHi, Table *9*) and AT-rich elements from the *Malacosoma neustria* nucleopolyhedrosis virus (MnNPV) polyhedrin gene (Pfeiffer et al., 2012, Suzuki et al., 2006). It performed well *in vivo*, accounting for a 7.4 fold increase of reporter expression in *D. melanogaster* vs the original *D. melanogaster* consensus TIS sequence (Pfeiffer et al., 2012, Cavener and Ray, 1991). Based on these results and the relative closeness of mosquitoes and *D. melanogaster* (both order Diptera) as compared to mosquitoes and *S. frugiperda* or *B. mori* (both order Lepidoptera), TIS Syn21 was expected to outperform TIS Kozak, BmHi and Lep in mosquito cell lines, which it did not.

This outcome is not explained by the host species range of MnNPV. Comparing Pfeiffer et al. (2012)'s Syn21 construct (pJFRC13_Syn21) against the one used here (pIE1-Syn21-FF) there are two points of difference that may be relevant: 1) The nucleotide directly 3' of ATG is "ATGg" in pJFRC13_Syn21 ((Pfeiffer et al., 2012)) and "ATGt" in pIE1-Syn21-FF) (Table *6*). 2) The intron preceding Syn21 is "myosin heavy chain" in pJFRC13_Syn21 and "alcohol dehydrogenase" in pIE1-Syn21-FF (both of *D. melanogaster* origin). The nucleotide immediately 3' of ATG is known to be important in vertebrate translation initiation but is thought not to be so in invertebrates (Kozak, 1986, Kozak, 1987a). The effect of intron on translation initiation is described as dependent on promoter and coding sequence of the elements (Pfeiffer et al., 2010, Huynh and Zieler, 1999). It may be that the splicing activity *adh*-Syn21-FF in mosquito and moth cell lines is sufficiently different from splicing activity of Pfeiffer et al. (2012)'s construct in *D. melanogaster* to have hampered translation initiation efficiency. An interesting first step towards investigating this would be checking the activity of pIE1-Syn21-FF-SV40 in a *Drosophila* cell line, which was not explored in this project.

The cell lines used for these experiments are immortalised cell lines derived from embryonic, larval or pupal tissue of a specific species and are by no means a perfect simulation of individual cells within an adult insect, let alone of the whole organism. Referring again to the unexplained decrease in reporter output mediated by TIS Syn21, it is recommended that any

new TIS sequence be validated in a cell line model of a species of interest as a minimum standard and in advance of committing resources to the generation of stable transgenic lines.

Taking the work presented here to meet that standard, it is expected (but not known) that these *in vitro* results from cell culture models would show similar outcomes *in vivo* in whole insects. This assumption is based on the findings of Pfeiffer et al. (2012) and of Tatematsu et al. (2014), who each report a consistency of results from cell culture experiments to *in vivo* experiments.

Changing the 3'UTR sequence

In line with the available literature, we found that changing the 3'UTR sequence of the reporter gene cassette had a significant effect on reporter protein output. Using 3'UTR SV40 as a reference point, we found that 3'UTR *fs(1)*K10 (K10) reduced the rate of protein expression and 3'UTR P10 increased the rate of protein expression:

$$K10 < SV40 << P10$$

This rank order of efficacy of 3'UTR sequences was true in all five cell lines tested (four culicine mosquito, one moth). Comparing directly with the literature, Pfeiffer et al. (2012) describes "that the *p10* 3'-UTR can increase protein expression by more than a factor of 10 on its own". From Table 13 we can see that changes of 3.45 fold can be attributed to 3'UTR P10 (as opposed to SV40) when "TIS" and "cell line" are held constant. Looking at constructs with TIS sequence Kozak, we can see a 3.91 fold increase in reporter expression mediated by the change from Kozak:SV40:Aag2 to Kozak:P10:Aag2 (Appendix Table 14). Only in cell line Sf9 does 3'UTR mediate a change in reporter output to rival that reported by Pfeiffer – a 58 fold increase from Kozak:SV40:Sf9 to Kozak:P10:Sf9 (Figure *12*).

This data set provides a reference point for selection of a 3'UTR sequence when designing a transgene construct, as the 3'UTR sequence can mediate an increase or decrease in transgene expression. There are more practical limitations with choice of 3'UTR sequence than with choice of TIS, as the 3'UTR sequence can be involved in mediating temporal or spatial specificity of gene expression – one would preferably use the endogenous 3'UTR matching the promoter if specificity is a priority. The relatively large size (several hundred bp) of the 3'UTR sequence means that overall transgene construct size and sequence homology within and between constructs must be considered, which may be prioritised over expression efficiency when selecting a 3'UTR sequence. In summary, greater changes in

expression efficiency can be mediated by choice of 3'UTR sequence than of TIS, but this advantage is mediated by the decreased versatility of 3'UTR sequences as compared to TIS.

Interaction of TIS and 3'UTR

As with the other pairwise interaction, the interaction of "TIS" and "3'UTR" is found to contribute significantly to the goodness of fit of the analytical model (Appendix Table 13). Such an interaction was also noted by Pfeiffer et al. (2012) who found that combining the Syn21 and P10 elements further increased expression from their transgene construct.

Summary of changes achievable with these tools

The data presented in this chapter supports the use of alternate translation initiation sequences (TIS) as a tool to modulate the efficiency of transgenic protein expression in arthropod cell lines. Data around the 3'UTR sequence offers a quantitative measure of efficacy of established 3'UTR sequences for use in non-model insects. The data furthermore supports a combined approach to modulating protein expression efficiency where a transgene design can accommodate changes to both loci.

Work with the TIS is particularly intriguing as the locus in question can be as few as 9nt and is thought not to affect the spatial or temporal specificity of gene expression. It does not require changes to the coding sequence and the effect is seen in several different insect species of interest. Although these experiments have all been carried out *in vitro*, other publications support the relevance of such results to *in vivo* work, notably for stable genome integration of transgenes in *D. melanogaster* and *B. mori* (Pfeiffer et al., 2012, Tatematsu et al., 2014).

The ability to increase transgene expression has many practical applications, in applied genetic technologies as well as for research purposes. For mosquito genetic control strategies in particular, the ability to decrease protein expression efficiency may be useful where expression of toxic (intentionally or not) proteins is needed. The scale of changes achieved in these experiments is typically smaller than that reported by other groups – 2 to 3 fold increases in expression vs 4 to 10 fold increases in expression (Tatematsu et al., 2014). The extent of up or down regulation of expression efficiency is thought to be somewhat plastic, however, as Tatematsu et al. (2014) reported different results depending on the coding sequence, *in vitro* or *in vivo* context and on the tissue context in which transgenes were expressed in an adult *B. mori*. Changes in coding sequence were not investigated in this study and the work was limited to an *in vitro* context.

The consistent ability of some TIS sequences to enhance transgene expression in different species and in different cell lines (BmHi consistently increases expression in these experiments and is published as increasing expression *in vitro* and *in vivo* in *B. mori*, which is not a species context tested here) suggests that such tools can be easily adapted to use in non-model species of interest. This feature is particularly attractive in the design of transformation marker constructs and for the design of common parts in a genetic control strategy; being able to transfer a system from one mosquito species to another without altering every transgene reduces the resource cost of such a project. It is also of value for species such as *C. quinquefasciatus* where the genome sequence resources are less well developed than those of species that have been the subject of more intensive investigation historically (e.g. *An. gambiae*, *Ae. aegypti*).

That changes in expression efficiency can be achieved with such small nucleotide changes is advantageous for constructs where size is a priority or where size of homologous sequences is important. Such constraints are common in design of transgenes for stable genome integration and in design of transgenes where multiple constructs must ultimately be present in the same individual. Where such factors are not of concern, the 3'UTR sequence can also be altered to take advantage of sequences that increase (i.e. P10) or decrease (i.e. K10) expression efficiency. In combination with an altered TIS sequence, such changes can dramatically increase the expression efficiency (more so than is possible with one element or the other).

Limitations of these tools

Are they independent of experimental constants used here?

As has been discussed, the scale of the effect of TIS sequence on expression efficiency has been linked to the coding sequence and tissue context of the transgene (Tatematsu et al., 2014). Constructs in this experiment have been tested with a single coding sequence, plasmid backbone, promoter and cell/tissue context. It is therefore unknown whether these results will persist in the face of such changes. This is particularly important as every transgene is different and it is not possible to validate every factor before committing resources to developing the transgene or even to developing a transgenic line with a stable genomic insertion. The literature, particularly the breadth of literature, can offer a reasonable starting point to infer which changes might persist and which might be sensitive to particular contextual features. As TIS sequences have been tested *in vitro* and *in vivo* in matched species by Pfeiffer et al. (2012) and by Tatematsu et al. (2014), we have an expectation that the

trends seen in this data set will persist *in vivo*. Pfeiffer et al. (2012) used TIS sequences to enhance expression in different genomic contexts in *D. melanogaster.*

The literature reports that the coding sequence of the transgene effects the scale of change attributed to the TIS sequence (Tatematsu et al., 2014) but does not indicate that the direction of change is affected. The unexpected decreasing effect of TIS Syn21 in these experiments appears to be a sequence specific issue but suggests that caution is warranted in using new or unvalidated TIS sequences in your species of interest. Alternate coding sequences were not tested in this experiment, though such a test could be done by swapping the luciferase coding sequences between the experimental and control plasmids (firefly luciferase and Renilla luciferase).

To examine whether the effect of the TIS sequence is independent of the promoter sequence, further work was done by Phil Leftwich and Michelle Anderson. Three sets of TIS and 3'UTR were selected as "high" "medium" and "low" expressing combinations and were tested with a panel of other constitutively active promoters. This panel included both virus derived promoters and endogenous promoters (e.g. poly-ubiquitin). The ability of Lep:P10 and BmLo:K10 to increase and decrease (respectively) transgene expression was found to be independent of the promoter used (personal communication, Phil Leftwich and Michelle Anderson).

A fundamental limitation of this work is its use of consensus and virus-derived sequences to identify TIS, 3'UTRs and combinations thereof that are persistent across species and are designed to work ubiquitously throughout an individual. These features will not be suitable where a high degree of temporal or spatial specificity is required of transgene expression (e.g. germline expression). In such instances, the TIS and 3'UTR sequences of the endogenous gene from which the promoter is derived are likely to be more efficacious. This is supported by Tatematsu et al. (2014)'s work looking at transgene expression in different silk gland tissues of *B. mori*.

Although cell line representations of four insect species are included in this work, it is notable that three of those species are subfamily Culicinae and two of those are genus *Aedes*. The data generated shows that a rank order trend is consistent across all five cell lines for both TIS and 3'UTR. That the TIS Syn21 is effective in *D. melanogaster* but consistently decreases expression in the mosquito and moth cell lines tested suggests that caution is warranted in applying these findings to other insect species not represented here (e.g. Anopheline mosquitoes).

In balance of the literature and the findings of this experiment, the author is confident that the rank order of effects shown in cell line experiments will persist to other contexts. It is not expected that we can predict the specific scale of effects in other contexts, nor that we can confidently predict the effect of each sequence in species backgrounds that are not tested here.

Further work

The optimisation of the experimental protocol presents a system with which other points of curiosity can be explored. Such experiments would inherit the limitations of the cell culture nature of the work, but could be used to explore some of the questions raised in these results:

- Is the effect of TIS and 3'UTR sequences independent of inclusion of an intron and of the sequence of the intron?
- What is the behaviour of this panel of plasmids, particularly those using TIS Syn21, in a *D. melanogaster* cell line?
- Do the rank order trends persist if the *luc+* coding sequence is replaced with a different reporter sequence?

Although more resource intensive to implement, *in vivo* validation of these findings is the obvious next step. Such work would likely be done with a small selection of TIS and 3'UTR sequences.

Summary

In this work we have developed a robust method to ascertain that different translation initiation sequences (TIS) can alter rates of protein expression from a fixed amount of plasmid DNA. We have quantified the relative activity of three 3'UTR sequences in five insect cell lines and shown the relative activity of TIS, 3'UTR and their matrix of combinations independent of cell line. We have demonstrated that caution is warranted when using transgene motifs or sequences across species.

Chapter 4 – An *in vitro* CRISPRa assay for validating sgRNA activity

Introduction

The work described in this thesis is set in the framework of contributing to the development of a daisy-chain style gene drive system in *Ae. aegypti*. Moving on from the modulation of transgenic protein expression, as discussed in Chapter 3, Chapter 4 focuses on expanding the molecular toolkit available for the expression of single guide RNAs (sgRNAs) in transgenic mosquitoes. This work is achieved through local development of a CRISPR activation (CRISPRa) assay.

Transgenic expression of non-coding RNAs

Promoter sequences

Transgenic expression of non-coding RNAs (ncRNAs) (RNAs other than messenger RNA) has typically been achieved through use of RNA polymerase III promoter sequences as RNA polymerase II transcripts are typically transported to the cytoplasm (though some later re-enter the nucleus as ribonuclear protein complexes) and have post-transcriptional modifications that could inhibit ncRNA activity (Good et al., 1997). Only the type 3 subset of RNA polymerase III (RNA pol III) promoters is known to use a fully external (not internal to the RNA sequence) 5' promoter sequence (reviewed by Schramm and Hernandez (2002)). Several type 3 RNA pol III promoters are described for transgenic expression of ncRNAs in a human or mammalian context, including U6, 7SK and H1 (Good et al., 1997, Kabadi et al., 2014).

For expression of ncRNAs in an insect context (for RNA interference and for sgRNAs), U6 promoters are typically used and have been described as such in *D. melanogaster*, *Ae. aegypti*, *An. gambiae*, *B. mori* and *P. xylostella*, among others (Huang et al., 2017, Konet et al., 2007, Ma et al., 2014, Wakiyama et al., 2005). Arthropod 7SK genes are more recently described (Gruber et al., 2008a, Yazbeck et al., 2018, Gruber et al., 2008b). Although there are typically several U6-like genes and promoters identifiable in insect genomes, many are non-functional pseudogenes and only 1 to 3 functional U6 promoters are described for each of the aforementioned insect species (Hernandez et al., 2007).

Terminator sequence

The DNA terminator sequence for RNA pol III is a short (4 – 7nt) repeat of thymines (T) (Reviewed by Matera et al. (2007)). In the context of developing a transgenic organism, the short length of the RNA pol III terminator sequence is advantageous – it is cheaper to synthesise, easier to express and reduces the overall size of a transgenic cassette (size is inversely correlated with efficiency of genome integration (Geurts et al., 2003)). For a daisy-chain style gene drive system, this short sequence is furthermore advantageous in that it contributes less to stretches of sequence homology within and between transgenes that could result in undesired homologous recombination within a transgenic individual (Noble et al., 2019).

Expression of sgRNAs in species of interest

Though sgRNAs are foreign to eukaryotic systems, expression can be achieved using the same tools as for expression of other ncRNAs, this is typically achieved in insects by using U6 promoter sequences (Ma et al., 2014, Dong et al., 2015, Gratz et al., 2013, Hammond et al., 2016). To mitigate the risk of selecting for drive-resistant alleles in the target population of a CRISPR-based gene drive system, it is desirable to be able to express multiple sgRNAs in the same individual (Figure 14) (reviewed by Esvelt et al. (2014)). In a daisy-chain style gene drive system, this would be multiplied by the requirement to express sgRNAs targeting multiple genes (Figure 15) (Noble et al., 2019).

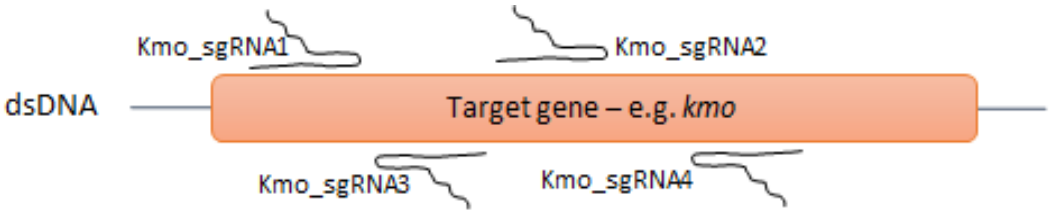


Figure 14: Cartoon representation of multiple sgRNA targets in a single target gene. In this representation there are four sgRNAs targeting four different loci within the target gene. This configuration acts as a mitigation against selection for resistant alleles in the population and is thought to be crucial for a successful CRISPR/Cas population control strategy.

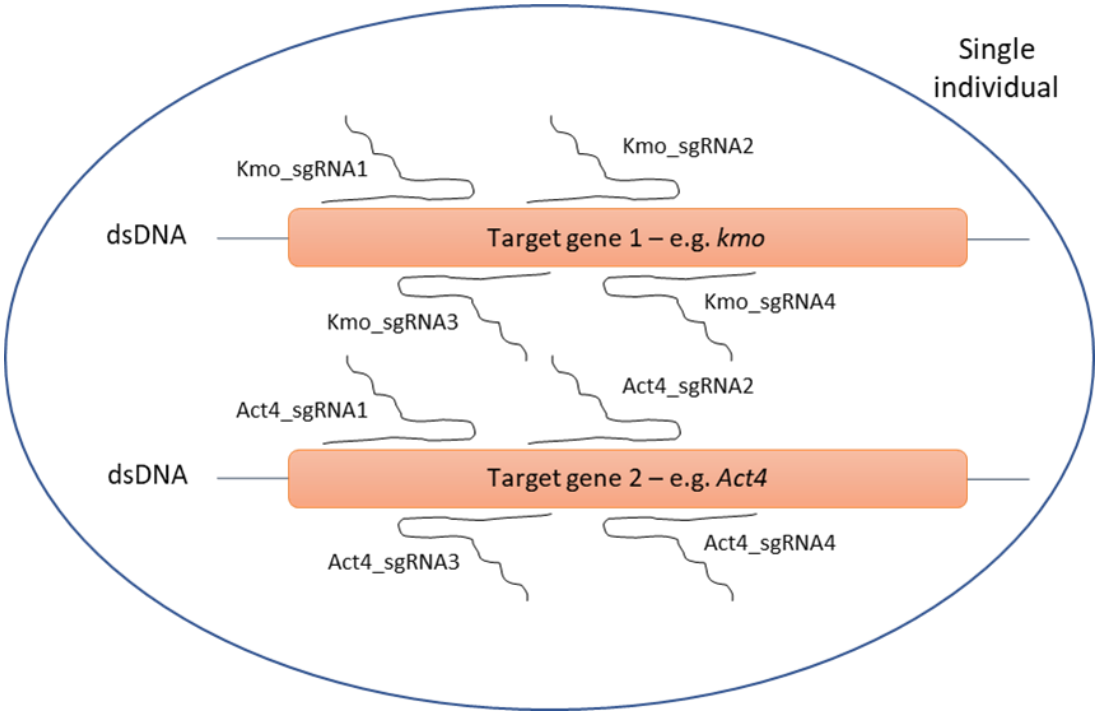


Figure 15: Cartoon representation of multiple sgRNA targets against two target genes in a single individual. In this representation there are four sgRNAs targeting different loci in target gene one and a further four sgRNAs targeting different loci in target gene two. Such a configuration allows simultaneous CRISPR/Cas targeting of multiple genes in a single individual, whilst maintaining mitigations against selection for resistant alleles.

It is not appropriate to use the same two RNA pol III promoters repeatedly in order to express the required number of different sgRNAs. Repetitive DNA sequences, especially identical, repeated DNA sequences, increase the risk of homologous recombination occurring within the transgene construct once it is inserted in the genome. For a 'split drive' or daisy chain gene drive, repeated DNA sequences increase the risk of different transgenes in the same individual recombining with one another. This scenario could result in the constrained gene drive becoming an unconstrained 'global drive' (as discussed in Chapter 1) (Noble et al., 2019).

With only 1 to 3 RNA pol III promoters validated in insect species of interest and a need to express at least eight different sgRNAs in order to construct the proposed daisy chain gene drive system, more RNA pol III promoters are needed, or a system to express multiple sgRNAs from each promoter.

Systems to express multiple sgRNAs from a single promoter are suggested in Noble et al. (2019) and revolve around identifying an enzymatic process to cleave transcribed sgRNAs from one another from a single transcript – akin to a bacterial operon. Micro RNA (miRNA)

and transfer RNA (tRNA) sequences are suggested as possible methods (Yan et al., 2016, Xie et al., 2015, Xie et al., 2017, Wang et al., 2017, Port and Bullock, 2016). Each of these methods were explored in preliminary experiments without convincing success in mosquito cell culture, though tRNA results reported in *D. melanogaster* could be reproduced in Drosophila cell line S2 (Appendix Figure 2). In light of these difficulties, it was decided to focus instead on the identification and validation of additional RNA pol III promoters.

Acquisition of additional RNA pol III promoters

Identification of novel RNA pol III promoters in insect species of interest

Konet et al. (2007) note cross-species activity of mosquito U6 promoters, mainly of an *An. gambiae* promoter in an *Ae. aegypti* cell line. Building on recent advances in the quality and number of genome sequences publicly available for mosquito species of interest (e.g. *Ae. aegypti*, *C. quinquefasciatus*, *Anopheles sp.*), it was hypothesised that U6 promoters of closely related species could be used to augment the availability of RNA pol III promoters for transgenesis in any one species of interest. This strategy is already employed for RNA pol II promoters, particularly where a model organism (such as *D. melanogaster*) is thoroughly described – e.g. *D. melanogaster* promoter of *Actin5C* and *hsp70* (heat-shock protein 70) are used in *Aedes. sp*. research (Pinkerton et al., 2000, Labbe et al., 2010).

The U6 promoter sequence is not well conserved outside of critical RNA pol III initiation motifs, TATA-like box and proximal sequence element (PSE) (described from page 96, alignment in Appendix D). The U6 gene product (ncRNA) has activity in the spliceosome and is extremely well conserved (reviewed by Hernandez et al. (2007) and Matera et al. (2007)). It is intended that U6 ncRNA sequence can be used as a search term to interrogate publicly available genome assemblies of species of interest and thus identify putative novel U6 promoter sequences.

Arthropod 7SK genes have been identified through bioinformatics methods, taking advantage of improvements in quality and availability of genomic resources for an increasing number of non-model species (Gruber et al., 2008a, Gruber et al., 2008b). Building on examples in mammalian systems (Kabadi et al., 2014), it is intended to identify and test novel 7SK promoter sequences of mosquito species of interest using the same methods as for U6 promoter sequences.

As the use of heterologous promoters is intended, this project is designed in such a way as to be easily adapted to work in any insect species of interest. This flexibility is exercised to

carry out experiments in several insect cell lines, representing multiple species of interest. *In vitro* (cell line) work is once again favoured (Chapter 3) as a practical representation of whole insects that confers increased throughput and reduced resource cost (as compared to whole insect work). The validity of this representation is discussed from page 104.

Validation of putative RNA pol III promoters

Verification of a promoter sequence can be carried out relatively simply by inserting the putative promoter sequence in the correct position in relation to a reporter gene, then placing that construct into an appropriate environment and recording the presence of the reporter gene product. For promoters where the gene product is a protein, there are popular reporter genes (e.g. luciferases or fluorescent proteins) that can be measured quantitatively. For an RNA pol III promoter where the gene product is ncRNA, quantitative reverse transcription PCR (qPCR) is the best way to directly quantify the presence of a specific RNA.

As with any RNA handling protocol, care (and therefore time) must be taken to minimise degradation of RNA between sample collection and quantification by qPCR. With the short length of sgRNA (~108nt) sample degradation would be a pressing concern. This could be mitigated by using a 'reporter' RNA that is expressed by the putative promoter sequences but is easier to isolate and quantify by qPCR than an sgRNA would be.

An alternative solution is to bypass the requirement to isolate and handle RNA altogether. Modifications of Cas9 have been described that can deactivate endonuclease activity without damaging its ability to bind to sgRNA or dsDNA (Jinek et al., 2012, Bikard et al., 2013, Qi et al., 2013) (reviewed by Brocken et al. (2018)). By conjugating this deactivated Cas9 (dCas9) to transcription factors, gene expression can be activated in a programmable fashion (using the sgRNA target sequence); this system has been termed CRISPR activation (CRISPRa) (Gilbert et al., 2013, Cheng et al., 2013). Predicated on an assumption that sgRNA availability is a limiting factor in reporter expression, a CRISPRa assay should be able to report relative abundance of a specific sgRNA and could therefore be used to report promoter activity. Protocols around a CRISPRa system could be designed to omit any RNA handling steps.

TRE-CRISPRa reporter assay

Tetracycline controlled trans-activator (tTA) inducible gene expression is an established system within the host lab and a Tet response element (TRE) reporter plasmid was used as the basis for designing an in house CRISPRa assay for use in insect cell lines. The TRE reporter plasmid is used in transgenic insects as an inducible/repressible expression system, it

requires the presence of tTA protein (and the absence of tetracycline) in order to express the reporter protein.

The TRE used consists of seven repeats of a tetracycline operator (TetO) sequence, followed by a minimal promoter sequence, *D. melanogaster* Hsp70, and a reporter protein coding sequence (CDS) (Figure 16). There is minimal expression of the reporter protein in the absence of tTA. In the presence of tTA (and the absence of tetracycline), the TetO sequences are bound by tTA proteins and the minimal promoter is stimulated by the transcription factor component of tTA. In the presence of tetracycline, tTA is preferentially bound to tetracycline and does not induce reporter expression (reviewed by Addgene at https://www.addgene.org/collections/tetracycline/, accessed Aug 2021).

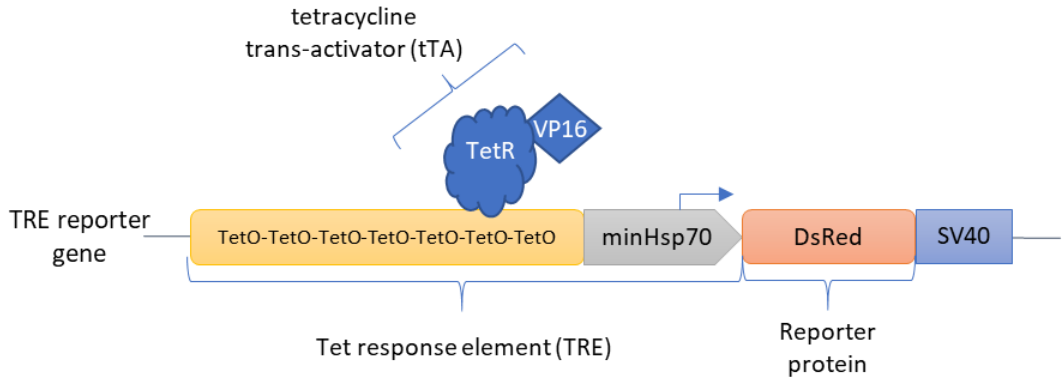


Figure 16: Cartoon representation of tTA inducible gene expression. Tetracycline trans-activator (tTA) consists of a tetracycline repressor protein (TetR) conjugated with a viral transcription factor (VP16). In the absence of tetracycline, tTA binds to the TetO sequence of the TRE and brings its transcription factor into proximity of the minimal promoter (minHsp70), prompting and enhancing expression of the DsRed reporter sequence. In the absence of tTA binding to the TRE, there is minimal expression of DsRed.

For the purposes of a CRISPRa assay, the TetO sequence can be used as a recognition site for enzymatically inert dCas9 to bind to the TRE and stimulate reporter expression by virtue of transcription factors conjugated to dCas9 (Figure 17). The seven repeats of the TetO sequence in the TRE may allow simultaneous binding of multiple dCas9 molecules, additionally stimulating the minimal promoter and increasing reporter expression. The use of a TRE based CRISPRa assay was locally conceived of by Tim Harvey-Samuel.

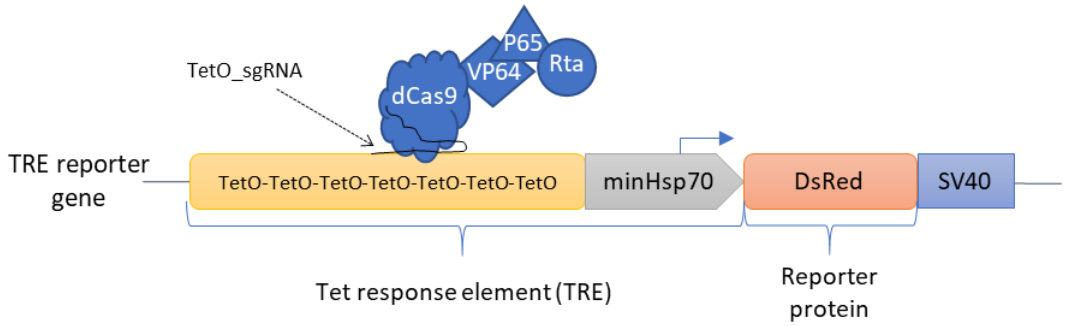


Figure 17: Cartoon representation of CRISPRa expression of the DsRed reporter gene. Using the same TRE reporter gene, TetO_sgRNA guides dCas9 to bind to the TetO sequence in the TRE. The dsDNA, sgRNA, dCas9 complex brings transcription factors conjugated to dCas9 into proximity of the minimal promoter, stimulating expression of the reporter protein (DsRed). Transcription factors VP64, P65 and Rta are known collectively with the enzymatically deactivated Cas9 as dCas9-VPR.

dCas9-VPR

From the identification of an enzymatically inert Cas9 (dCas9) mutant by Jinek et al. (2012), many versions of programmable DNA binding assays have been developed (reviewed by Brocken et al. (2018)). Chavez et al. (2015) empirically identify three transcription factor conjugates with dCas9 that create a gene activation assay with much increased activity as compared to single transcription factor conjugates. This dCas9-VPR assay uses VP64 (a tetrameric repeat adaptation of the herpes simplex virus VP16), P65 (nuclear factor κB 65kDa subunit) and Rta (Epstein-Barr virus R trans-activator), in that order, as C-terminus conjugates to the dCas9 protein ((Chavez et al., 2015), reviewed by Casas-Mollano et al. (2020)). The dCas9 protein itself is generated by including silencing mutations of the nuclease domains (D10A and H841A) (Jinek et al., 2012).

The dCas9-VPR assay has been demonstrated as efficacious when activating endogenous genes in *D. melanogaster* (Chavez et al., 2016) and is anticipated to perform similarly in other insect cell lines with a heterologous dsDNA source (plasmid). Complications around toxicity of the dCas9-VPR protein in moth and mosquito cell lines are a reasonably anticipated risk but can be easily determined in preliminary experiments. Such issues could be mitigated by decreasing the amount of dCas9-VPR protein introduced to the cells.

TetO specific sgRNA

The 'programmable' aspect of a CRISPR/Cas system is achieved through the short, non-coding single guide RNA (sgRNA), a synthetic hybrid of backbone (trans-activating CRISPR RNA (tracrRNA)) and target sequence specific (CRISPR RNA (crRNA)) RNAs (Jinek et al., 2012). The sgRNA facilitates binding of the Cas protein to dsDNA and is able to be designed to target

any DNA sequence, so long as there is a protospacer adjacent motif (PAM) 3' of the DNA target sequence (Jinek et al., 2012). To replace the function of tTA with CRISPR/dCas9 (Figure 17) on the TRE reporter gene, sgRNA(s) will be designed against the TetO repeat sequence (Figure 17). As the TetO sequence is synthetic and foreign to the host cell genome, the risk of off-target activity is greatly reduced, though this will be checked during sgRNA design.

For the initial design of the CRISPRa assay, quantifying the relative abundance of TetO-specific sgRNA as a proxy for promoter activity, the TetO-sgRNA sequence will remain constant once optimised. In later applications of a successful CRISPRa assay, the binding efficiency of different backbone structures of TetO-specific sgRNA could be explored, so long as the TetO-specific sequence remains constant. This idea is further explored in Chapter 5.

TRE reporter plasmid

Extant TRE reporter plasmids with the *D. melanogaster* Hsp70 minimal promoter have been locally demonstrated to function in transgenic *Ae. aegypti* and *P. xylostella* (personal communication, Tim Harvey-Samuel). These typically use a fluorescent reporter protein and a second, constitutively expressed transgenic cassette that expresses a separate fluorescent reporter protein as a visual marker for the presence of the TRE reporter plasmid and as a positive control if the TRE reporter is not seen when/where expected.

An *in vitro* (cell culture) CRISPRa assay for validating sgRNA activity

Bringing these factors together, it is intended that a dCas9-VPR CRISPRa assay can be adapted for use in cell line representations of insect species of interest, as a method for validation and quantification of sgRNA activity – which is used here as a proxy for RNA pol III promoter activity. If an assay can be established, then the aim of this work is to use that assay to determine whether further RNA pol III promoters for our species of interest can be sourced from closely related species.

Methods

Plasmids

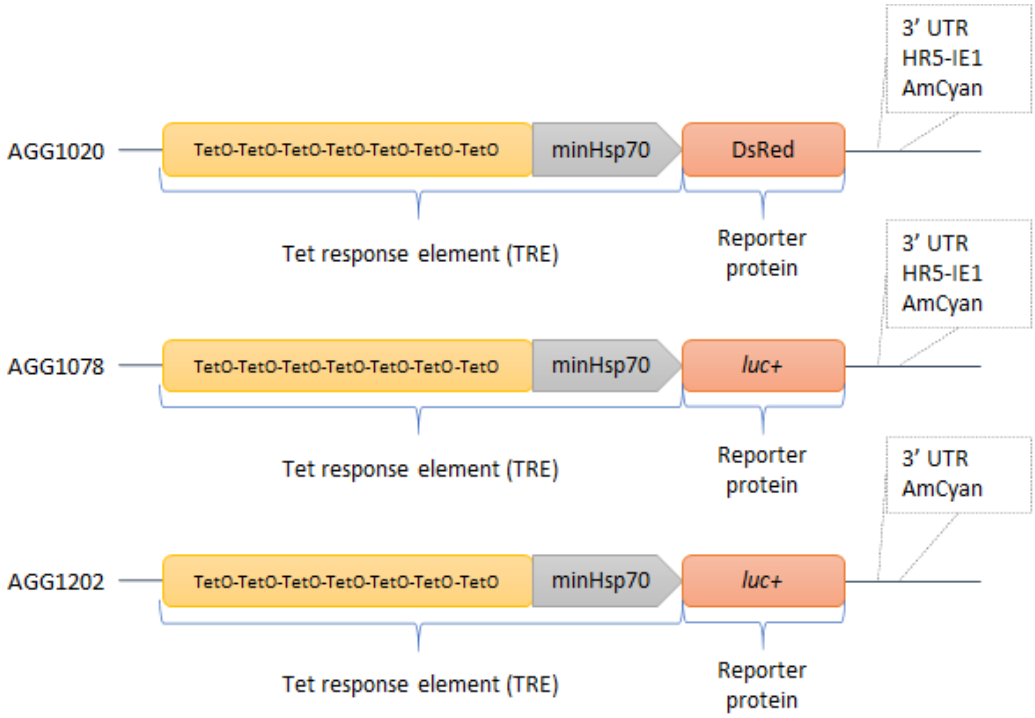


Figure 18: CRISPRa reporter plasmids. Three iterations of the reporter plasmid, AGG1020, AGG1078 and AGG1202 show development as the CRISPRa assay was adapted to luciferase reporting (AGG1020 -> AGG1078) and then optimised by removing a superfluous HR5-IE1 promoter (AGG1078 -> AGG1202).

Reporter plasmids

Reporter plasmid AGG1020 (Figure 18) was an existing insect transformation plasmid for constitutive expression of a fluorescent transformation marker, AmCyan, and tTA inducible expression of fluorophore DsRed. It was synthesised by Genewiz and kindly provided by Tim Harvey-Samuel. This reporter plasmid was used in preliminary experiments.

Luciferase reporter plasmid AGG1078

To adapt the CRISPRa assay to a dual luciferase reporter assay format, a new reporter plasmid (AGG1078) was developed by replacing the DsRed coding sequence (CDS) in AGG1020 with firefly luciferase CDS (*luc+*). The vector backbone was prepared by digesting plasmid AGG1020 with restriction enzymes AvrII and ApaI according to standard methods (Figure 19). These restriction sites are uniquely present in AGG1020, 5' and 3' of DsRed CDS. The linearised vector was dephosphorylated and prepared for ligation using standard techniques.

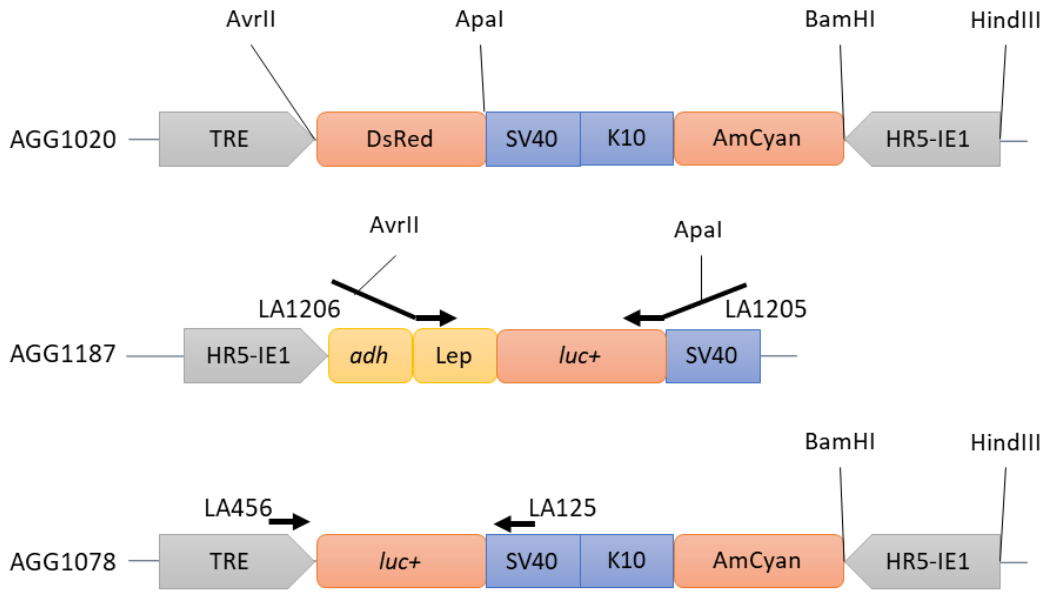


Figure 19: Cloning steps for creating reporter plasmid AGG1078. The fluorescent reporter plasmid AGG1020 was linearised with AvrII and ApaI to remove the DsRed CDS. A PCR product from template AGG1187 contained the firefly luciferase coding sequence with AvrII and ApaI restriction sites. Ligation of the prepared vector and template results in luciferase reporter plasmid AGG1078, which was sequence verified using primers situated outside of the replaced CDS. Unique restriction sites 5' and 3' of constitutive promoter HR5-IE1 are indicated.

Firefly luciferase CDS *luc+* was PCR amplified from AGG1187 (pIE1-Lep-FF-SV40, Chapter 3) using primers LA1205 and LA1206, which introduced AvrII and ApaI recognition sites 5' and 3' of the PCR product (Figure 19). PCR product LA1205-LA1206/AGG1187 was produced using a variant of touchdown-PCR, with a lowered annealing temperature for the first 10 cycles and a higher annealing temperature for the subsequent 25 cycles (Table 15). The PCR product was visually confirmed by agarose gel, then purified and prepared for ligation by digestion with AvrII and ApaI.

Table 15: PCR thermocycle for LA1205-LA1206/AGG1187

Step	**Cycles**	**Temp. (°C)**	**Time (s)**
Initial denaturation	1	98	60
Denaturation	10	98	10
Annealing		64	30
Extension		72	120
Denaturation	25	98	10
Annealing		72	120
Extension			
Final Extension	1	72	600

Linearised vector AGG1020/AvrII_ApaI was ligated with insert LA1205-LA1206/AGG1187/AvrII_ApaI and transformed into competent cells using standard methods. PCR colony screening was omitted and four independent clones were verified by Sanger sequencing using primers LA125 and LA456 (Figure 19). A successful clone was selected and the confirmed colony was grown to a larger volume to purify a stock of reporter plasmid AGG1078.

Luciferase reporter plasmid AGG1202

Luciferase reporter plasmid AGG1078 was modified to remove the constitutively active HR5-IE1 promoter from the legacy transformation/transfection marker cassette (Figure 18). HR5-IE1 was removed using unique restriction enzyme sites 5' and 3' of the sequence, BamHI and HindIII. Standard molecular protocols were used and the linearised vector was blunt-end cloned using Klenow (NEB, UK) and T4 ligase (NEB, UK). A post-ligation digestion was employed to reduce the presence of circular plasmids containing the HR5-IE1 sequence (XhoI). The reaction was heat inactivated and transformed into competent cells according to standard protocols.

Seven colonies were grown up, plasmid purified and confirmed by restriction enzyme digests (NcoI-HF; XhoI). Likely clones were confirmed by Sanger sequencing using primers LA291 and LA37, which sit in the translation initiation site of AmCyan and the 3' piggyBac flank, respectively. The sequence confirmed colony was grown to a larger volume to purify a stock of reporter plasmid AGG1202.

dCas9-VPR expressing plasmid AGG1068

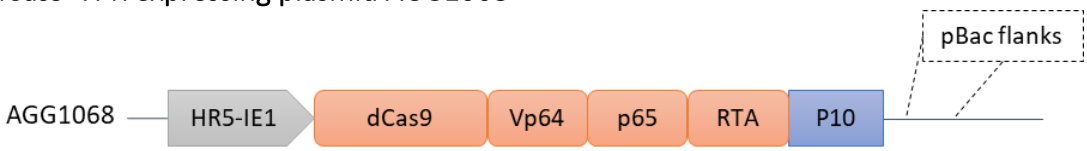


Figure 20: Representation of dCas9-VPR expressing plasmid, AGG1068. Constitutively active enhancer promoter HR5-IE1 drives expression of the dCas9-VPR

The dCas9-VPR expressing plasmid was designed based on SP-dCas9-VPR (gift from George Church (Addgene plasmid # 63798; http://n2t.net/addgene:63798)) (Chavez et al., 2015). The coding sequence was codon optimised for *Ae. aegypti* using Regenerator Vector NTI codon optimisation software (Thermofisher Scientific, USA) and the AGG1068 plasmid construct (Figure 20) was synthesized by Genewiz (Anderson et al., 2020). This work was carried out by Tim Harvey-Samuel.

Competent cells were transformed, grown up and plasmid purified according to standard protocols.

sgRNA expressing plasmids

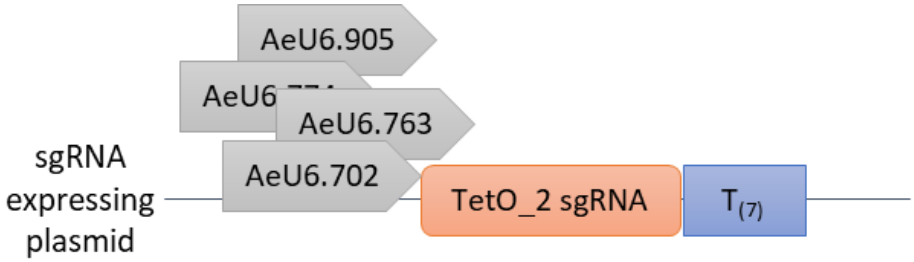


Figure 21: sgRNA expressing plasmids. The plasmid backbone with TetO_2 protospacer, standard sgRNA backbone sequence an $T_{(7)}$ terminator sequence was used as a vector. RNA polymerase III promoter sequences were cloned from genomic DNA or synthesised as gene fragments and could then be cloned into the vector. This system was used to test >30 RNA pol III promoters in the CRISPRa assay.

All sgRNA expressing plasmids were based on the same template, using pJet vector (ThermoFisher Scientific, USA) and the sgRNA sequence for TetO_sgRNA2 (see below for sgRNA design) (Figure 21). Promoter sequences were PCR amplified from genomic DNA where whole insects were available. Promoter sequences were otherwise taken from relevant publicly available databases (see below for *in silico* identification of RNA polymerase III promoters) and synthesised as gene fragments by Twist Bioscience (San Francisco, USA). All promoter sequences were kindly cloned by Sebald Verkuijl and Michelle Anderson. Some sgRNA expressing plasmids were synthesised as entire plasmids in the pTwist_amp vector by Twist Bioscience (San Francisco, USA) (Anderson et al., 2020). Standard protocols were used for cloning, transformation, and purification of plasmids. All plasmids were verified by Sanger Sequencing. The sgRNA expressing plasmids discussed in this chapter are those discussed in Anderson et al. (2020).

Renilla luciferase plasmid

Renilla luciferase expressing plasmid pRL-OpIE2 (AGG1080) was used in all dual luciferase assay experiments. Construction is described in Chapter 3 (Anderson et al., 2020).

RNA

sgRNA design

Three TetO specific sgRNAs were designed against the TetO repeats in the Tet response element (TRE) (Figure 18) using CHOPCHOP CRISPR design tool (Labun et al., 2016, Montague et al., 2014). Each TetO sgRNA was designed with a different sgRNA backbone variant from Noble et al. (2019) (Table 16). This work was kindly carried out by Victoria Norman.

Table 16: sgRNA sequences used in plasmids in this work

Name	**Proto-spacer name**	**Proto-spacer**	**Backbone**	**Backbone variant name**
TetO_sgRNA1	TetO_1	TCTCTATCACTGATAGGGAG	GTCCTAGAGCCATGAAAATGGCAAGTTAGGATAAGGCTAGTCCGTATTCAACGCTGAAAAGCGTGGCACCGAGTCGGTGC	sgRNA backbone 09
TetO_sgRNA2	TetO_2	ACTTTTCTCTATCACTGATA	GTTCCAGAGTCGTGCTGGGAACAGCACGACAAGTTGGAATAAGGCAAGTCCGTTATCATGCCGGAAGGCAGGCACCGATTCGGTGC	sgRNA backbone 25
TetO_sgRNA3	TetO_3	CACTTTTCTCTATCACTGAT	GTCGCAGAGCATCTGAAAAGATGCAAGTTGCGATAAGGCAAGTCCGTTATCAAGCTCGGGAGAGCTGGCACCGAGTCGGTGC	sgRNA backbone 29

sgRNA synthesis

in vitro transcribed sgRNAs (*iv* sgRNAs) used in this work were kindly synthesised from custom oligonucleotides by Victoria Norman, according to standard techniques. All *iv* sgRNA described in this work used a single sgRNA backbone sequence (LA988) and the same proto-spacers as described in Table 16. These primers are described in

Table *17*.

Table 17: Primers used to transcribe sgRNAs

Primer name	**Sequence (5' - 3')**
LA985: TetO_sgRNA1	GAAATTAATACGACTCACTATAGGTCTCTATCACTGATAGGGAGGTTTTAGAGCTAGAAA
LA986: TetO_sgRNA2	GAAATTAATACGACTCACTATAGGACTTTTCTCTATCACTGATAGTTTTAGAGCTAGAAA
LA987: TetO_sgRNA3	GAAATTAATACGACTCACTATAGGCACTTTTCTCTATCACTGATGTTTTAGAGCTAGAAA
LA988: sgRNA rev	AAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC

Transfection

Transfection experiments were carried out with standard protocols, as optimised in Chapter 3. A 96-well plate format is used with each well in a column representing a biological repeat for that transfection mix. A master mix was used wherever possible and multiple cell lines were transfected with the same master mix where possible. Components of the transfection (nucleic acids) varied for each experiment during optimisation and these amounts are presented in ng/well.

Dual luciferase assay

Transfected cells were harvested and read by dual luciferase assay, according to standard protocols. The volume of cell lysate processed by dual luciferase assay could be varied for each cell line in each experiment (as discussed in Chapter 3). Lysate volumes used in these experiments are noted in Table 18.

Table 18: Lysate volumes used for experiments in Chapter 4

Figure	**Experiment**	**Cell line**	**Lysate volume (ul)**
Figure 22	Optimisation 3	Aag2	5
Figure 23	Reporter plasmid testing	Aag2	1
		Hsu	5
		Sf9	5
Figure 24	Multi-species U6 promoters	Sua5.1	12

Figure	Experiment	Cell line	Lysate volume (ul)
Figure 25	Novel anopheline U6 promoters	Sua5.1	5
		4a_2	5

Identification of novel RNA polymerase III promoters

Where RNA polymerase III promoters (RNA pol III promoters) could not be drawn from the literature, they were identified through "BLASTn" search of the published genome sequence of the relevant species (Table 19). Each of these genome sequences were accessed through the VectorBase website (https://vectorbase.org/vectorbase/app/) (Giraldo-Calderon et al., 2015). The protocol for screening BLASTn matches is discussed from page 96.

Table 19: Genome assemblies used in identification of putative RNA pol III promoters

Species	Strain	Structural annotation version	Host website	Accessed on
Ae. aegypti	LVP_AGWG	Aaeg L3 & L5.1	Vectorbase	07/2018
Ae. albopictus	Foshan	AaloF1.2	Vectorbase	06/2018
C. quinquefasciatus	Johannesburg	CpipJ2.4	Vectorbase	07/2018
An. gambiae	PEST	AgamP4.9	Vectorbase	06/2018
An. funestus	FUMOZ	AfunF1.8	Vectorbase	07/2018
An. Stephensi	Indian	AsteI2.3	Vectorbase	07/2018
An. Stephensi	SDA-500	AsteiS1.6	Vectorbase	07/2018
An. albimanus	STECLA	AalbS2.5	Vectorbase	07/2018
An. arabiensis	Dongola	AaraD1.8	Vectorbase	07/2018

Analysis

The analytical methods discussed in Chapter 3 were used to manage and analyse data gathered from dual luciferase assays discussed in this chapter. In brief, luciferase readings in arbitrary light units (ALU) were visualised and analysed in GraphPad Prism (Version 9.0.0 for windows; GraphPad Software, USA). Microsoft Excel was used for some data manipulation (transformation to firefly luciferase / Renilla luciferase) and visualisation. The quality control checks developed in Chapter 3 were used for the work described in this chapter (Appendix D).

Experimental data validating and quantifying the activity of novel RNA pol III promoters was typically analysed using a non-parametric test, Kruskal-Wallis, followed with Dunn's multiple comparison to account for multiple comparisons being made with a single dataset. This analysis was carried out in GraphPad Prism (Version 9.0.0 for Windows; GraphPad Software, USA) and is discussed around each figure.

Blank Page

Results and Discussion

Several preliminary experiments were carried out to establish optimal parameters for carrying out the CRISPRa-TRE-dual luciferase assay. These focused intently on the amounts and ratios of each nucleic acid component of the transfection and are detailed in Appendix D. Much of this work was carried out with *in vitro* transcribed sgRNAs. Two preliminary experiments are described in this chapter: optimisation of the amount of sgRNA expressing plasmid used (to ensure results within the assay's dynamic range) and development of an improved TRE-luciferase reporter plasmid.

The pipeline used to identify and validate RNA pol III promoters *in silico* is discussed, followed by *in vitro* validation of those promoters using the CRISPRa assay.

Preliminary experiments

Optimisation of the amount of sgRNA expressing plasmid in the CRISPRa assay

Building on the results of preliminary experiments described in Appendix D, an optimisation experiment to characterise the effect of amount of sgRNA expressing plasmid on CRISPRa results was carried out in *Ae. aegypti* cell line Aag2. Plasmids expressing TetO_sgRNA in a single cassette were selected, AGG1120 and AGG1155. Each of these plasmids uses known RNA pol III promoter AeU6-702 to express TetO_sgRNA2; AGG1155 uses a 5' modification of the TetO_sgRNA that is reasonably expected to hobble sgRNA binding efficiency. This plasmid was developed for a separate project and was used here to represent a 'weak' RNA pol III promoter.

For the avoidance of doubt, this experiment is carried out at two transfection amounts each of reporter plasmid and of dCas9-VPR expressing plasmid (a continuation of preliminary work discussed in Appendix D). A four-step serial dilution is transfected for each sgRNA expressing plasmid. The background threshold is set to the upper 99.9% CI of the "No sgRNA control".

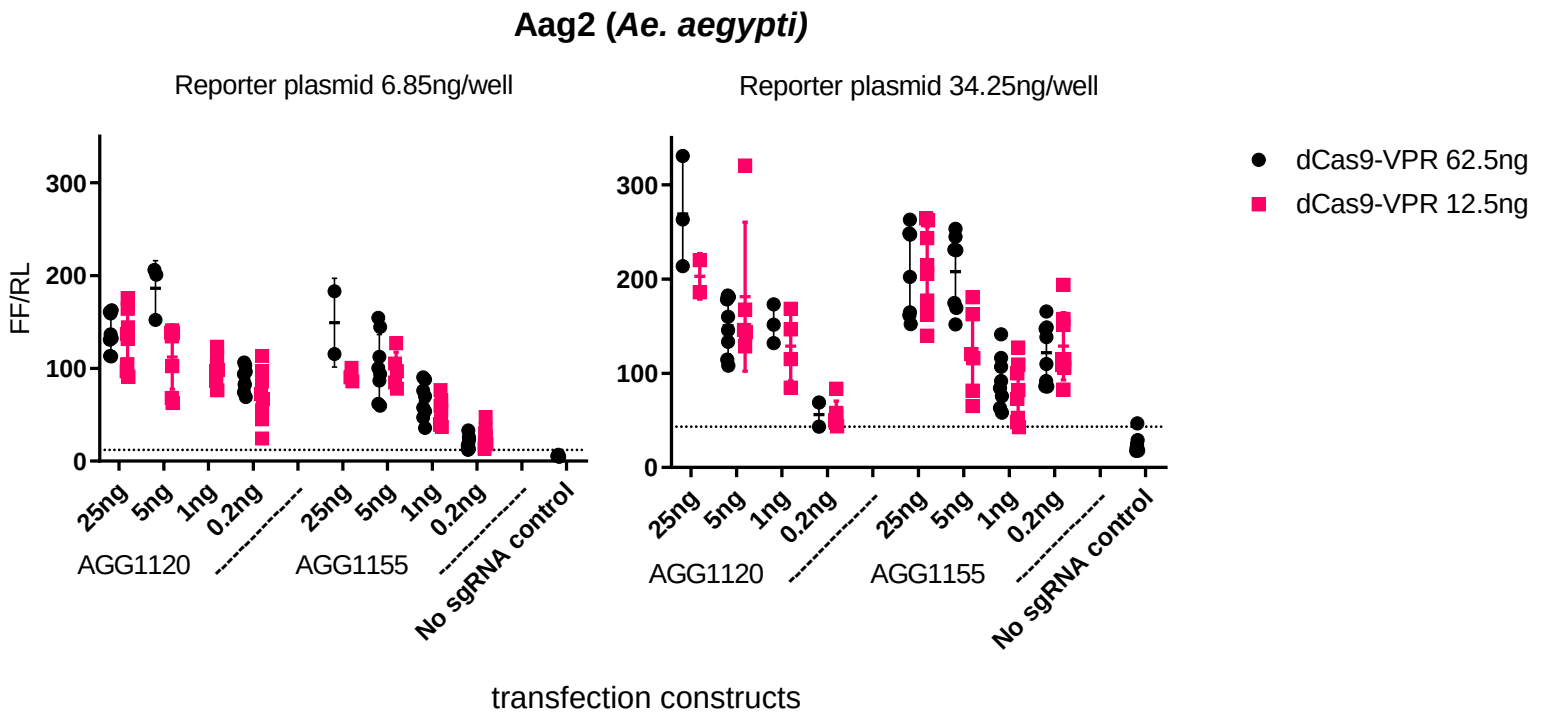


Figure 22: Graphs of results from optimising the amount of plasmid expressing sgRNA transfected in cell line Aag2. This work was done at two amounts of reporter plasmid (separate graphs with independent scales) and at two amounts of dCas9-VPR expressing plasmid (different colours) in order to account for any dose dependent interactions of the CRISPRa components. Data is reported as FF/RL (y-axis) and the background threshold is indicated with a horizontal dotted line, set by the 99.9% CI FF/RL measurements of the negative control (No sgRNA control). Results from each transfected well are shown as solid circles with mean and standard deviation indicated by horizontal and vertical lines, respectively. The x-axis specifies the sgRNA expressing plasmid and amount of the plasmid that were used in the transfection. Two sgRNA expressing plasmids are used (AGG1120 and AGG1155) at a 4-step serial dilution of amounts. The only difference between these plasmids is the U6 promoter used in the plasmid – *Ae*U6X or *Ae*U6X. N = 2-8. Data points were excluded where FF exceeded the quenching threshold.

Looking first at the change in quantity of dCas9-VPR plasmid transfected (interleaved colours) in Figure 22, there is no clear difference in results generated with 12.5ng/well dCas9-VPR plasmid vs 62.5ng/well (5x). It is concluded that saturation of dCas9-VPR in the CRISPRa assay occurs from 12.5ng/well, corroborating data from previous preliminary experiments (Appendix D). Changes attributable to a 5x increase in reporter plasmid (6.85ng/well vs 34.25ng/well) are also in line with those seen in previous experiments (Appendix D). Although there is an increase in luciferase activity with increased amount of reporter plasmid, it is not a linear increase. The effect does not appear to be dependent on other factors (e.g. amount of sgRNA expressing plasmid).

Each sgRNA (differing in efficiency of the TetO_sgRNA) was transfected in a four-step, 5x serial dilution. At 6.85ng/well reporter plasmid, each sgRNA amount (for both plasmids) has clear overlap of luciferase activity with its 5x and 0.2x amounts. For plasmid AGG1120 (fully functional TetO_sgRNA) there is overlap in luciferase activity for the 25ng/well and 0.2ng/well groups. For plasmid AGG1155 (hobbled TetO_sgRNA) there may be a trend of mean luciferase activity decreasing with amount of sgRNA, but only at 25ng/well and 0.2ng/well (125x decrease) are the luciferase results separated. Of note, the 0.2ng/well amount for plasmid AGG1155 has results overlapping with background. There is no distinction between luciferase results of the two sgRNA plasmids except at 0.2ng/well with the higher amount of dCas9-VPR plasmid.

The results are very similar at the higher amount of reporter plasmid (right graph, Figure 22), though there is better distinction between the highest and lowest amounts of sgRNA plasmid AGG1120 than at 6.85ng/well reporter plasmid. The background threshold increases as the amount of reporter plasmid increases and luciferase activity from the lower amounts of sgRNA expressing plasmid (1ng/well and 0.2ng/well) overlap with background. Distinction of sgRNA plasmid AGG1120 from AGG1155 at 0.2 – 1ng/well is no clearer than at the lower amount of reporter plasmid.

It is concluded from Figure 22 that the amount of sgRNA plasmid transfected can saturate the CRISPRa assay from very low quantities as the luciferase activity is largely unresponsive to the amount of sgRNA plasmid used. Focusing on smaller transfection quantities may allow the assay to discriminate between different quantities of sgRNA (i.e. differently efficient promoters driving sgRNA expression), especially with a greater number of repeats. This opportunity is limited by the amount of background luciferase expression from the

unstimulated reporter plasmid ("No sgRNA control"). It was decided to re-optimise the reporter plasmid in order to decrease unstimulated reporter expression.

Reporter plasmid optimisation

Background expression of the unstimulated reporter gene was noted throughout preliminary experiments and is anecdotally observed in transgenic *P. xylostella* (personal communication, Tim Harvey-Samuel and Victoria Norman). It is considered that elements of the HR5-IE1 enhancer promoter are acting in cis on the Tet response element (TRE) Hsp70 mini-promoter, causing this effect (Guarino et al., 1986, Pullen and Friesen, 1995).

As a straightforward optimisation for the CRISPRa assay, the HR5-IE1 element was removed from the constitutive marker cassette in reporter plasmid AGG1078, creating a second-generation TRE-luciferase reporter plasmid, AGG1202. These reporter plasmids were tested side-by-side in three cell lines (Figure *23*). Two "No sgRNA" conditions were used to assess background luciferase expression from the unstimulated reporter plasmids, with and without the dCas9-VPR plasmid. Three sgRNA conditions were used to look at activation of the reporter plasmid with a range of sgRNA availability: sgRNA expressing plasmids AGG1120 and AGG1155 (sgRNA hobbled by 5' modification), both using promoter AeU6-702, and *in vitro* transcribed sgRNA. This experiment aimed to determine whether background expression was reduced in assays using the second-generation reporter plasmid as compared to the first (via removal of the HR5-IE1 cassette) and to assess whether such a change improved the dynamic range of the CRISPRa assay.

Data is shown in Figure *23* as three graphs, grouped by cell line. All data is shown as FF/RL values and individual samples are shown alongside their mean and standard deviation. The different reporter plasmids are shown interleaved for each cell line, separated by colour. Some samples recorded firefly luciferase (FF) activity in excess of the assay's quenching limit, these were excluded from further analysis. An entire group of data was lost for *iv* sgRNA in cell line Sf9 with reporter plasmid AGG1202 for this reason (marked with italics and asterisks on the x-axis label). The positive control groups (those containing sgRNA) are labelled on the x-axis by the type of sgRNA that was used. The negative control groups are labelled by which CRISPRa component they are missing. Statistical analysis was carried out to determine if reporter plasmid AGG1202 reduced FF/RL as compared to AGG1078 – this is reported in Table 20.

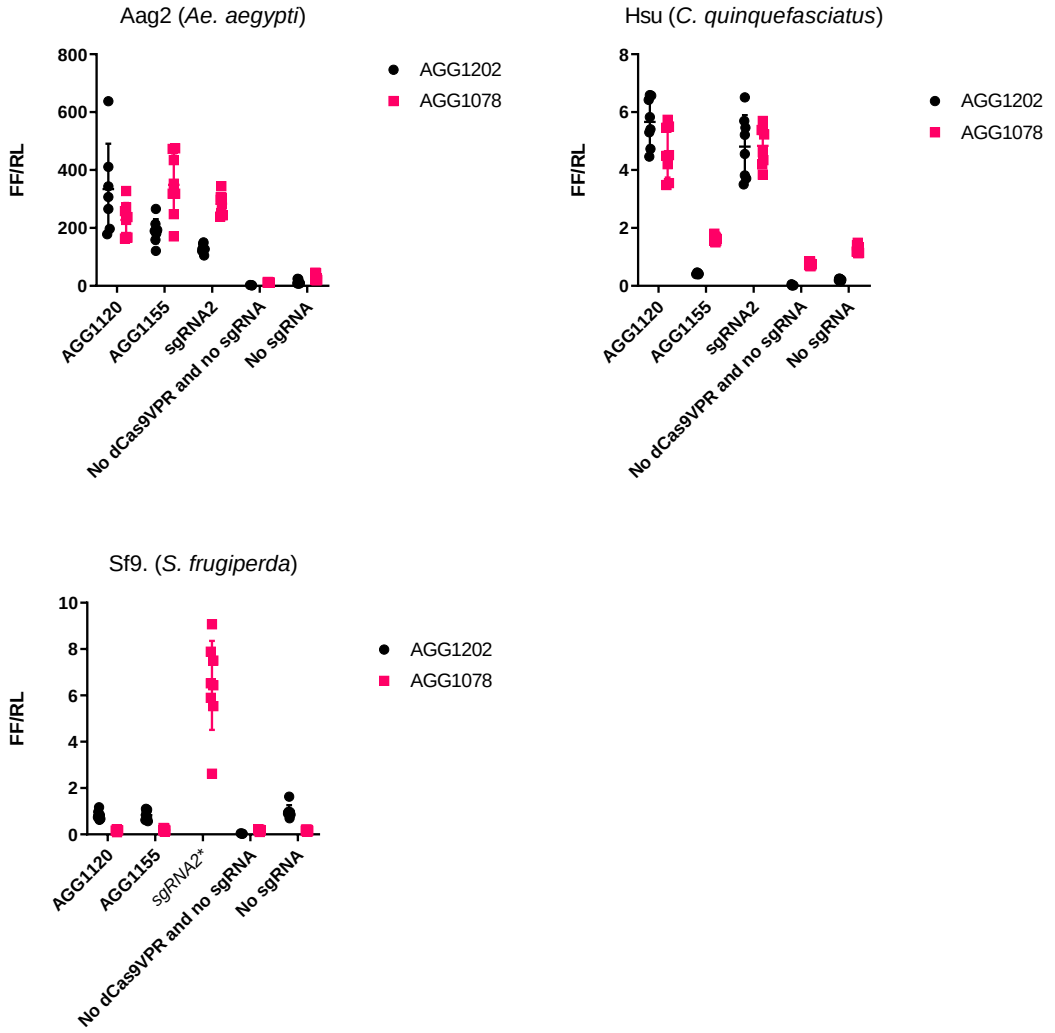


Figure 23: Graphs showing results of luciferase reporter plasmid testing in three cell lines, Aag2, Hsu and Sf9. Data for each cell line is shown on independent graphs with independent y-axis scales. The transfection conditions are shown on the x-axis (all transfections included the dual luciferase assay control, RL). Each data point is shown for N = 7 - 8 and mean and SD are shown where visible. Plasmids AGG1120 and AGG1155 are identical sgRNA expressing plasmids, apart from a 5' sgRNA modification in AGG1155 that hobbles binding efficiency. "*iv* sgRNA" refers to *in vitro* transcribed sgRNA. Where a condition is missing (*iv* sgRNA2, AGG1202, Sf9) the group name is noted in italics with an asterisk. The reporter plasmid identity (AGG1202 or AGG1078) is denoted by colour (legend to right).

Looking at the negative control groups ("No dCas9VPR and no sgRNA"; "No sgRNA"), we can see that there are differences between reporter plasmids (AGG1202 and AGG1078) depending on cell line context. In all three cell lines there appears to be good separation between the positive groups (sgRNA containing) and the negative controls. Table 20 shows the summary results of a multiple Mann-Whitney test carried out for each cell line (independently) testing whether the results acquired with reporter plasmid AGG1202 were lower than the results acquired with reporter plasmid AGG1078. The multiple comparisons

were adjusted using a false discovery rate (FDR) set to 1.00% using GraphPad Prism version 9.0.0 (USA). Results are reported as P values with "ns" where the P value is >0.05. As this is a directional (one tail) test, only incidents of AGG1202 having significantly lower FF/RL than AGG1078 can be reported (i.e. it is not distinguishable if AGG1202 has significantly higher FF/RL than AGG1078).

Table 20: P value results of multiple Mann-Whitney tests (independent by cell line) to determine whether results obtained with new reporter plasmid AGG1202 are lower than those obtained with previous reporter plasmid AGG1078. A hyphen represents a comparison that could not be done and "ns" indicates "not significant" (P value > 0.05)

	P value		
	Aag2	**Hsu**	**Sf9**
AGG1120	ns	ns	ns
AGG1155	0.007	<0.001	ns
sgRNA2	<0.001	ns	-
No dCas9VPR and no sgRNA	<0.001	<0.001	<0.001
No sgRNA	ns	<0.001	ns

The aim of optimising the CRISPRa luciferase plasmid was to reduce background (unstimulated) expression. Table 20 shows that AGG1202 is a successful optimisation in all three cell lines for negative control "No dCas9VPR and no sgRNA", but only in cell line Hsu for negative control "No sgRNA".

The positive controls (sgRNA expressing groups) were included in this experiment to screen for unintended side-effects of changing the reporter plasmid (e.g. changes in expression trends between groups). Although there are some indications of the change in reporter plasmid decreasing expression from some positive control groups (Table 20), it is thought that this is explained by the decrease in background expression in sgRNA containing groups where the CRISPRa assay is not necessarily saturated with TetO-sgRNA. Of note, the AeU6-702 promoter does not appear to function in cell line Sf9 and the modified sgRNA (AGG1155) does not appear to function in cell line Hsu.

It was concluded from this experiment that removing the cis HR5-IE1 sequence from the CRISPRa reporter plasmid can reduce the background (unstimulated) expression, particularly in cell line Hsu. This reporter plasmid, AGG1202, was used in all further experiments.

Finalised CRISPRa assay

The iteration of the CRISPRa assay carried forward uses the 48hr time point for sample collection and transfects each the dCas9-VPR and reporter plasmids (AGG1068 and AGG1202) at 25ng/well. Plasmid expressed sgRNA is used at 0.3ng/well and *iv* sgRNA at 40ng/well. The amount of Renilla luciferase plasmid (AGG1080) varies by cell line: 50ng/well for Aag2, U4.4 and *An. gambiae* cell lines Sua5.1 and 4a_2; 1ng/well for Hsu, Sf9 and S2.

Validation of RNA pol III promoters using the CRISPRa assay

An example of the experimental use of the CRISPRa dual luciferase assay is in the validation of novel RNA polymerase III (RNA pol III) promoters; to some extent these promoters can also be ranked by their relative activity. By altering the promoter sequence of the TetO-sgRNA2 expressing plasmid, the relative activity of putative RNA pol III promoters can be quantified.

Identification of putative RNA pol III promoter sequences

The identification of putative RNA pol III promoter sequences was done *in silico*, using publicly available genome sequences; different methods were used for U6 gene promoters and 7SK gene promoters.

U6 gene promoters

During optimisation of the CRISPRa assay, it was noted that the *Ae. aegypti* U6-702 promoter is functional in a cell line derived from *C. quinquefasciatus* (Hsu) (Figure *23*). The same dataset showed that this cross-species activity did not extend to lepidopteran cell line Sf9 (*S. frugiperda*). Konet (2007) have previously demonstrated some cross-species activity of *Ae. aegypti* and *An. gambiae* U6 promoter sequences, though mostly of *An. gambiae* promoter activity in *Ae. aegypti* cells. Cross-species activity of lepidopteran U6 promoters was noted in Figure 44, where *P. xylostella* U6 promoters function in a *S. frugiperda* cell line. This cross-species activity buoyed the intention to mine additional RNA pol III promoters for species of interest from other, closely related, species where genome assemblies are available.

U6 promoter sequences bear little sequence homology outside of the prescribed motifs – proximal sequence element (PSE) and TATA-like box – so the U6 gene sequence (107bp with good homology within and between *Ae. aegypti* and *An. gambiae*) was used as a search term to identify putative U6 promoters within the *Ae. aegypti* genome and in the genomes of other mosquito species (Konet et al., 2007, Hernandez et al., 2007). The U6 gene sequence from *Ae. aegypti* U6-702 (AAEL017702) was selected as this promoter was functional in preliminary experiments and is previously published (Konet et al., 2007).

AAEL017702 U6 gene sequence

GTCCTAGCTTCGGCTGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGCCCCTGC
GCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTT

BLAST searches were done using default settings (see page 88 for genome assemblies used) and putative U6 genes were first screened for the presence of an RNA pol III terminator sequence ($T_{\geq 4}$). Sequences without a functional terminator sequence were assumed to have non-functional promoter sequences and were excluded from further analysis. Remaining sequences were used to isolate 600bp 5' of the putative U6 sequence, which was taken as the putative U6 promoter sequence. Putative promoter sequences were then screened for the presence of the required motifs, PSE and TATA-like box, as described by Konet et al. (2007) and Hernandez et al. (2007).

~30bp 5' of ncRNA – TATA-like sequence – TATATA

~65bp 5' of ncRNA – PSE sequence – (A/G)T(C/G)CA(T/C)(C/T)GCTAGAA

This work was carried out in eight published mosquito genomes Table 19, page 88) and is summarised in supplemental information (Appendix Table 18). In brief, 29 likely U6 promoter sequences were identified with this method, several of which are previously published. These sequences are presented in full in supplemental information (page 193).

7SK gene promoters

The 7SK gene does not have the same duplication in mosquitoes that is seen with U6 genes and is more reliably annotated in publicly available genome sequences (Yazbeck et al., 2018). The *Ae. aegypti* annotated 7SK gene sequence was used to identify the 7SK promoter – taken as 600bp 5' of the gene start. The invertebrate 7SK promoter is described as having similar sequence motifs to the U6 promoter – a TATA-like box and proximal sequence element (PSE), but with an additional distal sequence element (DSE) that is thought to be necessary for activity (Gruber et al., 2008a). Assuming poor sequence homology of the promoter, the 7SK gene sequence (AAEL018514) was used to search for 7SK sequences in the *Ae. albopictus*, *C. quinquefasciatus* and *An. gambiae* genomes. Each genome search yielded one good match, each annotated as "arthropod 7SK RNA".

AAEL018514 7SK gene sequence

GGAGGTGTGTGTCTTCGTCTGTGATGGCACGATAACTGAACATTGATCGCTTTACGTGTTAGTTTGC
AGATCTGCTCAGTGGCAACCCGTCACACCTTGATACAATCGTCTGGCAGTCCGGATCTGGTATCACG
GGTGAACTCTCGCTGCACGGCGCCGGGCCGAACGCACGATTGATGTCATTTGTGATACAAGACTAC
TGCCGTTCTTACCCAACTCTTTCCAAATTGTTGAGTATAAAAATCGTAATTTAATACAGATAGCTTAG
CTTCGGATTAAAATTACATTGTTCAGAACGCTTCCATATCACTAGGGCACCGCCGAGCGGTCGGCCC
ATTCTTTTG

Ae. aegypti – AAEL018514; *Ae. albopictus* – AALF029648; *C. quinquefasciatus* – CPIJ039933; *An. gambiae* – AGAP028235.

To extend this search to other Anopheles species, the 7SK gene sequence of AGAP028235 was used as a search term against the genome databases for *An. arabiensis*, *An. funestus*, *An. stephensi* and *An. albimanus*. Although two hits were found for some species, there was only one 7SK gene per species that included the full gene, RNA pol III terminator, TATA-like box and PSE.

An. arabiensis – AARA015292; *An. funestus* – AFUN015339; *An. stephensi* – ASTEI12173; *An. albimanus* – AALB015206.

It is noted that AARA015292 and AGAP028235 (the 7SK promoters) are nearly identical, which is attributed to the evolutionary closeness of *An. gambiae* and *An. arabiensis*. All 7SK promoter sequences are presented in full in supplemental information (page 201).

Validation of novel RNA pol III promoters

The CRISPRa dual luciferase assay was used to validate activity of the putative RNA pol III promoters identified *in silico*. As the promoter sequences and sgRNA are relatively short (<650nt), promoters expressing TetO-sgRNA2 were synthesised as gene fragments that could then be cloned into standard expression vectors (e.g. pJet, ThermoFisher Scientific, USA). In conjunction with the 96-well plate format of the CRISPRa dual luciferase assay, almost every putative promoter was tested for expression activity *in vitro*.

A proportion of this work, focusing on published and novel RNA pol III promoters in Culicine cell lines (Aag2, Hsu and U4.4) has been published as (Anderson et al., 2020). Unpublished data is presented in Figure *24*, characterising the activity of a multi-species panel of RNA pol III promoters in *An. gambiae* cell line 4a_2. Further data is presented in Figure *25*, looking at the activity of novel *Anopheles* species U6 promoters in two *An. gambiae* cell lines.

Multi-species panel of RNA pol III promoters in *An. gambiae* cell line

Building opportunistically on work in Culicine mosquito cell lines (published as Anderson et al. (2020) and not discussed further here), and an availability of *An. gambiae* cell lines, a panel of RNA pol III promoters representing six insect species was tested in *An. gambiae* cell line Sua5.1 (Table 21).

Table 21: RNA pol III promoters tested by CRISPRa in *An. gambiae* cell line Sua5.1

Species origin	**Gene accession**	**Local promoter ID**	**Plasmid ID**	**Reference**
Ae. aegypti	AAEL017702	*AeU6-702	AGG1120	(Konet et al., 2007)
Ae. albopictus	AALF029744	AbU6-744	AGG1252	n/a
C. quinquefasciatus	CPIJ039596	CqU6-596	AGG1131	n/a
An. gambiae	AGAP013557	*AgU6-557	AGG1256	(Konet et al., 2007)
An. gambiae	AGAP013695	*AgU6-695	AGG1164	(Konet et al., 2007)
D. melanogaster	FBgn0004190	*DmU6-3	AGG1173	(Wakiyama et al., 2005)
P. xylostella	Not annotated	*PxU6-3	AGG1210	(Huang et al., 2017)
An. gambiae	AGAP028235	Ag7SK	AGG1261	n/a

These promoters were selected as 'strong' examples of U6 promoters from each *Ae. aegypti*, *Ae. albopictus* and *C. quinquefasciatus* based on previous experimental data (Anderson et al., 2020). All three *An. gambiae* RNA pol III promoters discussed (from page 96) are included. Fly promoter *D. melanogaster* U6-3 is included, as is moth promoter *P. xylostella* U6-3; each of these sequences were obtained from their original publications (Huang et al., 2017, Wakiyama et al., 2005) and they are included to give an indication of how far (evolutionarily speaking) cross-species activity can be observed with RNA pol III promoters.

The results of the CRISPRa dual luciferase assay are shown in Figure 24 and demonstrate the utility of this assay for validating or characterising RNA pol III promoters in cell line representations of species of interest. In this example, eight promoters from six species are compared side by side in a single cell line. Only one experiment was carried out, but biological replicates were produced using eight different wells of cells for each transfection group. The results are standardised to the luciferase control (RL) and are presented as FF/RL with a CRISPRa negative control "No sgRNA" to indicate background expression from the unstimulated reporter plasmid. The upper 99.9% CI of the "No sgRNA" group is marked with a horizontal line (0.098) but is obscured by the x-axis in Figure 24. The promoters tested

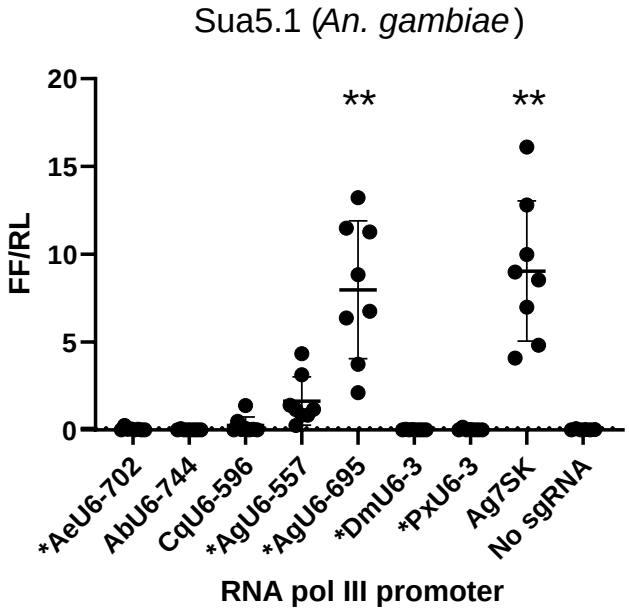


Figure 24: Graph showing CRISPRa results for a multi-species panel of RNA pol III promoters in *An. gambiae* cell line Sua5.1. Data is shown as FF/RL (y-axis) with individual samples plotted and mean and SD indicated. The promoter driving sgRNA expression in each group is listed on the x-axis; previously published promoters are denoted with an asterisk. N = 7 – 8. Where a group has been determined to be significantly different from background expression, the P value is marked with asterisks above the sample results. "**" indicates P value < 0.01 and > 0.001.

(Table 21) are indicated on the x-axis and include an asterisk where the promoter is previously described in the literature.

Statistical analysis of these results is limited by the small dataset (N = 7 – 8) and large number of comparisons (9 experimental groups). A Kruskal-Wallis test was used to determine that there is a significant difference in results between experimental groups ($P < 0.0001$) and follow up analysis was done with Dunn's multiple comparisons test to identify which groups (promoters) gave reporter activity (FF/RL) significantly greater than background ("No sgRNA"). These groups are marked with asterisks above the sample values. Promoters AgU6-695 and Ag7SK each have FF/RL activity that is significantly greater than background (P values < 0.01 and > 0.001).

Although promoter AgU6-557 is not significantly different from background expression in Figure *24*, it is noted that this promoter is reported to be active, but to a lesser degree than AgU6-695, in another *An. gambiae* cell line AG-55 (Konet et al., 2007). The Ag7SK promoter appears to have activity indistinguishable from that of Agu6-557, which is currently a popular *An. gambiae* RNA pol III promoter (Hammond et al., 2016), and greater activity than AgU6-

557, the other published *An. gambiae* RNA pol III promoter. This work therefore contributes a third, active, promoter for use in *An. gambiae*.

No activity distinguishable from background is noted for the *Ae. aegypti*, *Ae. albopictus*, *D. melanogaster* or *P. xylostella* promoters. Previous experiments have indicated that these *D. melanogaster* and *P. xylostella* promoters are not active in Culicine mosquito cell lines (Aag2, Hsu, U4.4), but that the *An. gambiae* promoters tested here are (Anderson et al., 2020). Konet et al. (2007) describes a similar relationship of *An. gambiae* promoters active in an *Ae. aegypti* cell line, but *Ae. aegypti* promoters not active (or weakly active) in an *An. gambiae* cell line. Positive control experiments were done in *D. melanogaster* cell line S2 and moth (*S. frugiperda*) cell line Sf9 to confirm that the plasmids used in this experiment can express sgRNA in a favourable host context (supplemental information, Appendix Figure *4* and Appendix Figure *5*).

Panel of Anopheline putative U6 promoter sequences in two *An. gambiae* cell lines

Having established that Culicine U6 promoter sequences do not appear to be active in *An. gambiae* cell line Sua5.1, plasmids using the putative Anopheline promoter sequences (Table 22) to express TetO_sgRNA2 were constructed for testing in *An. gambiae* cell lines. In the absence of cell lines representing multiple Anopheline mosquito species, two different *An. gambiae* cell lines were used.

Table 22: RNA pol III promoters tested by CRISPRa in *An. gambiae* cell lines

Species origin	**Gene accession**	**Local promoter ID**	**Plasmid ID**	**Reference**
An. gambiae	AGAP013557	*AgU6-557	AGG1256	(Konet et al., 2007)
An. gambiae	AGAP013695	*AgU6-695	AGG1164	(Konet et al., 2007)
An. albimanus	AALB015132	AalbU6-132	AGG1276	n/a
An. arabiensis	AARA015171	AaraU6-171	AGG1277	n/a
An. arabiensis	AARA015449	AaraU6-449	AGG1278	n/a
An. funestus	AFUN015538	AfunU6-538	AGG1279	n/a
An. funestus	AFUN015704	AfunU6-704	AGG1280	n/a
An. stephensi	ASTEI11842	AsteiU6-842	AGG1281	n/a
An. stephensi	ASTEI11858	AsteiU6-858	AGG1282	n/a
An. stephensi	ASTEI11917	AsteiU6-917	AGG1283	n/a

The results of the CRISPRa dual luciferase assay are shown in Figure *25*, with independent graphs for each *An. gambiae* cell line. Ten promoters from five Anopheline species are tested,

and results are shown as individual data points (FF/RL) with mean and SD where visible. The promoters are labelled on the x-axis, in the same order for each graph. Analysis was carried out once more using a Kruskal-Wallis test to determine significant difference between experimental groups and Dunn's multiple comparison test as a follow up to determine which groups are significantly different from the control, "No sgRNA". Statistical analysis is done independently for each cell line and where there is a significant difference from background, the P value is indicated with asterisks above the results. The same P value key is used: P < <.05, *; P < 0.01, **; P < 0.001, ***. Asterisks along the x-axis labels indicated previously published promoter sequences.

The results shown in Figure 25 demonstrate that there is cross-species activity of U6 promoter sequences in *An. gambiae* cell lines, so long as the promoters originate from other Anopheline species. As in the previous experiment (Figure 24), this dataset has limited statistical power (N is not great enough to support the degrees of freedom present). This is best exemplified by the inability to distinguish activity of *An. gambiae* U6 promoters from background, though they have previously been shown to be active (AgU6-557 and AgU6-695; Figure 24, (Anderson et al., 2020, Konet et al., 2007)).

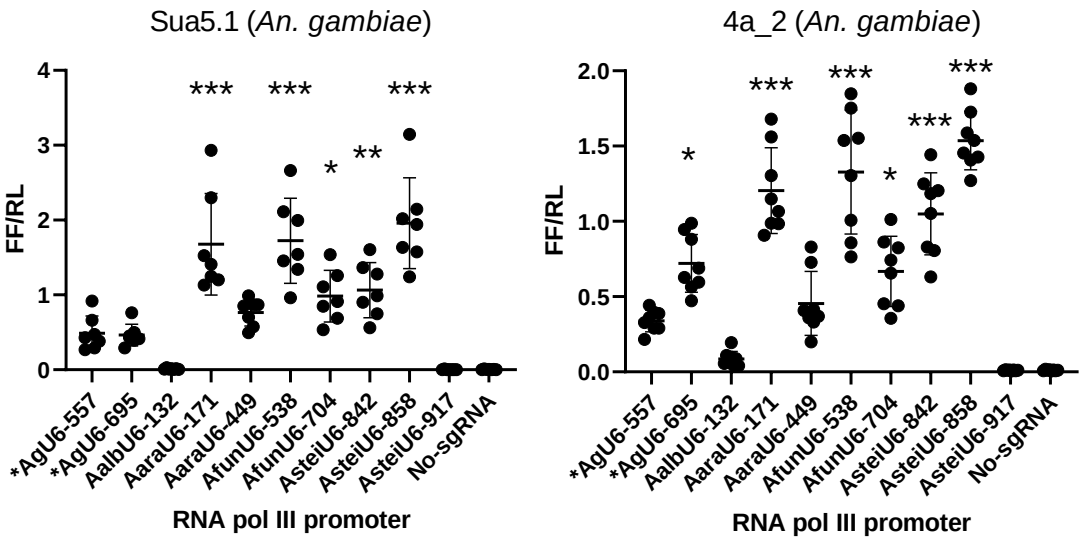


Figure 25: Graphs showing CRISPRa results for novel Anopheline U6 promoters in two *An. gambiae* cell lines. Data is shown as FF/RL with individual sample results plotted and mean and SD indicated; y-axis scales are independent for each cell line. The promoter driving sgRNA expression in each group is listed on the x-axis, where previously published promoters are indicated with an asterisk. Asterisks above sample results are used to indicate P value where a group has activity significantly different from background expression ("No sgRNA"). "*" $P < 0.05$; "**" $P < 0.01$; "***" $P < 0.001$. N = 7 – 8.

Looking at previously un-tested promoter sequences, AalbU6-132 and AsteiU6-917 yield results that cluster tightly background ("No sgRNA") expression in both cell lines. Promoter AraU6-449 gives a relatively higher mean ff/RL, but was, nonetheless, not found to be significantly different from background expression. The remaining five promoters (AaraU6-171, AfunU6-538, AfunU6-704, AsteiU6-842 and AsteiU6-858) are significantly distinguishable from background in both cell lines (Figure 25). Each of these appears to have greater than or equal to activity with the *An. gambiae* U6 promoters. Although comparisons between groups are not made statistically with this data set (insufficient power), it is observed that results are similar between cell lines. The change in magnitude of results (FF/RL) between cell lines cannot be attributed to a single cause as the relative activity of the OpIE2 promoter expressing control luciferase (RL) is not known for these cell lines.

These results demonstrate that there are a number of mosquito species of interest for which additional RNA pol III promoters can be mined from closely related species. The work shown in Figure *25* was an extension to the preceding experiments in *Culicinae* (Anderson et al., 2020) and was not further built upon in this body of work.

The CRISPRa dual luciferase assay as a method to validate and characterise RNA pol III promoters in mosquito cell lines

Although the datasets presented here (Figure 24 and Figure 25) could be considered 'preliminary' due to their lack of statistical power, data collected in Culicine cell lines and presented in Anderson et al. (2020) confirms that this CRISPRa assay can be used to validate and characterise activity of RNA pol III promoters in mosquito cell lines of interest. The work presented in Figure 25 suggests 5 – 6 novel U6 promoters that are at least as active in *An. gambiae* cell lines as the two promoters currently described (Hammond et al., 2016, Konet et al., 2007). The implications of these results are discussed further from page 104.

The applicability of dCas9-VPR experimental data generated *in vitro* (in cell lines) to Cas9 applications *in vivo* is also discussed from page 104.

Conclusion

Establishment of a fit-for-purpose CRISPRa assay

Through multiple iterations of TRE reporter plasmid and comprehensive optimisation experiments to determine behaviour of the CRISPRa dual luciferase assay according to each variable and potential component interactions (Appendix D), a standardised CRISPRa dual luciferase assay was established. In this configuration (page 96), sgRNA abundance is a limiting factor in reporter expression and the assay can be used to determine (in cells) the validity and the relative activity of a suite of putative RNA pol III promoters.

Of particular interest, the assay is characterised with both *in vitro* transcribed (*iv*) sgRNAs (Appendix D) and plasmid expressed sgRNAs. This confers versatility in application of the assay to different aspects of the questions around expressing multiple sgRNAs in a single individual whilst minimising sequence homology. This cell culture assay provides a robust resource for testing of sgRNA components of an insect transgene before committing resources to generating transgenic insects.

Acquisition of additional RNA pol III promoters

The work presented in this chapter and in Anderson et al. (2020) characterises the binary (presence/absence) and relative (comparing between promoter sequences) activity of many putative RNA pol III promoters of the U6 and 7SK genes in four mosquito species of interest. It adds to our understanding of the applicability and limitations of cross-species activity of these promoters. This represents a major contribution to the tool kit available for expression of non-coding RNAs in economically and medically important insect species.

In silico identification of putative RNA pol III promoters (U6 and 7SK)

A series of putative RNA pol III promoters were identified through an *in silico* workflow based around highly conserved promoter sequence motifs and the availability of high quality genome sequences. Putative promoter sequences were selected as 600bp regions 5' of conserved U6 and 7SK ncRNAs. A screening process was used to short-list only putative promoter sequences containing the critical sequence motifs. Putative promoter sequences were tested in cells, validating the *in silico* method.

There are several putative U6 promoter sequences that appear functional *in silico* but do not confer activity significantly above background expression in the CRISPRa assays. Although putative promoter sequences were scrutinised for homologies that could be predictive of this lack of activity (Appendix D), none were identified past those described in the initial method. Where promoters were tested that did not pass *in silico* screening (data not shown), no significant promoter activity was detected.

Validation of putative RNA pol III promoters in homologous cell lines

In the absence of sequence motifs that can reliably predict the validity and relative activity of an RNA pol III promoter sequence, the convenience of this *in vitro* (cell culture) screening assay is crucial for the confirmation of putative promoter sequences. The 96-well plate format, in combination with an automated plate reader (GloMax multi+, Promega, UK) allows for tens of constructs to be tested concurrently and is applicable to a wide range of adherent insect cell lines.

This workflow has been validated through homologous testing of published U6 promoters in *Ae. aegypti*, *An. gambiae* and *D. melanogaster* cell lines. Novel putative U6 promoters have been tested in homologous cell lines for *Ae. aegypti*, *Ae. albopictus* and *C. quinquefasciatus* cell lines (Anderson et al., 2020) and putative 7SK promoters have been validated for expression of sgRNA in each cell lines of the above mosquito species.

Activity of RNA pol III promoters in heterologous cell lines

The cross-species activity of these RNA pol III promoter sequences in culicine mosquito cell lines is discussed in detail in Anderson et al. (2020). Focusing then on the activity of Anopheline putative RNA pol III promoter sequences in heterologous (*An. gambiae*) cell lines, Figure 25 demonstrates that there are several promoters with activity comparable to that of the established homologous U6 promoters. This data indicates poor activity of culicine promoter sequences in *An. gambiae* cell line Sua5.1. There is no activity of promoters of more evolutionarily distant species *D. melanogaster* or *P. xylostella*.

For the *Anopheles sp.* promoters that are not active in *An. gambiae* cell lines (AalbU6-132 and AsteiU6-917), no positive control was done with which to determine if these promoters would be active in a homologous cell line. It is noted too that the data presented in Figure 24 and Figure 25 is from a single experiment each and does not have the statistical power (degrees of freedom) necessary to fully analyse relative activity of the different putative promoters.

Limitations

In vitro CRISPRa dual luciferase assay

As with the work discussed in Chapter 3, it is not known conclusively whether results obtained in an immortalised cell line will be directly relatable to gene expression in a whole insect. It is expected that U6 and 7SK promoter sequences would have constitutive expression, based on the role of U6 and 7SK ncRNAs in basic cellular functions (splicing and transcription regulation respectively); alleviating concerns around tissue-specific expression patterns that would not be seen in cell culture (derived from embryonic, larval or pupal origins). A further consideration is the extra-genomic nature of the plasmid DNA used in these experiments. Positional effects on RNA pol II promoters are commonly occurring and it is anticipated that this would occur to RNA pol III promoters as well.

Assurances can be drawn from comparisons of the results gathered *in vitro* with similar experiments conducted *in vivo* by other authors. Appendix Table 19 reviews the reported activity of various U6 promoter sequences in various species of interest, as available in the literature. Such comparisons were not available for 7SK promoters. Broadly speaking, the data collected in this project is corroborated by the literature. Relative efficacies are given as an estimate and there are some changes in rank order, which may be attributable to the type of assay used to quantify promoter activity and to the context (*in vitro* vs *in vivo*) in which the assay was done.

Novel RNA pol III promoters

The key limitation of the described workflow for identification and screening of putative novel RNA pol III promoters is the availability and quality of genome sequence databases for species of interest. As no link is identified between promoter sequence and its measured activity, these sequences are vulnerable to changes as genome assemblies are improved and sequence reads are increased. Such discrepancies could be mitigated by isolating putative promoter sequences directly from genomic DNA of the species in question. For the time considerations involved, this was not felt to be necessary for the work presented here. It

could be more applicable if maximum output from a single promoter was desired, or for species where there were significant concerns about the quality of the genomic sequence available.

In the context of expressing multiple sgRNAs in a single cell, this work is limited in that it does not examine whether there is an interaction between RNA pol III promoters present concurrently in a cell. Understanding whether there is an interaction between promoters, or a saturation point where additional promoters diminish the activity of pre-existing promoters is important in the design of a multiplexed sgRNA transgene. This topic could be explored through the use of non-specific sgRNAs (dummy sgRNAs) that do not code for the TRE reporter gene or for the host genome (e.g. ZsGreen-specific sgRNA). Single-promoter plasmids could be transfected concurrently or multiple promoter-sgRNAs could be synthesised on a single plasmid. By measuring abundance of TetO-sgRNA from a single promoter, any effect from additional promoters could be identified. This work may need to be conducted for several iterations and combinations if effects are dependent on promoter identity (e.g. homologous U6 vs heterologous U6).

Advantages

The key advantage of this workflow for identification and validation of novel RNA pol III promoter sequences is that it is relatively rapid and can characterise multiple promoters and sgRNAs in parallel. The fidelity of the cell line model and the CRISPRa model to the intended Cas9-transgenic insect appears to be sufficient to merit use of such a system to identify and validate a large number of putative RNA pol III promoter sequences in a short time.

The ability to work with cell line representations of multiple species of interest concurrently allows for more comprehensive validation of putative RNA pol III promoter sequences in homologous and heterologous cell lines. In addition to expanding the toolbox available for expression of ncRNAs in several mosquito species, this work has contributed to our understanding of the limitations on cross-species activity of U6 promoters, which may be of interest to those specialising in the evolution of such genes.

Chapter 5: Design improvements to express multiple sgRNAs on a single transgene

Introduction

Building on the work of Chapter 4, improvements to the design of a multiple sgRNA expression cassette (one transgene expressing more than one sgRNA) are considered in Chapter 5. Understanding that transgene size is inversely correlated with genome integration efficiency in mosquito transgenesis (Geurts et al., 2003), it is explored whether the ~600bp RNA polymerase III (pol III) promoters validated in Chapter 4 can be used at shorter lengths. This work arose from the consideration that there is no obvious sequence similarity between validated U6 RNA pol III promoter sequences 5' of the proximal sequence element (PSE) (Anderson et al., 2020) (page 105), which itself is approximately 65bp 5' of the snRNA sequence.

Furthermore, Chapter 5 examines whether variations to the sgRNA conserved 'backbone' sequence can be used in mosquito species of interest without unduly sacrificing the affinity of sgRNA/DNA/Cas protein interactions. This work is informed broadly by the body of literature examining natural (species) variations and optimisations of the synthetic sgRNA sequence (Dang et al., 2015, Briner et al., 2014, Chylinski et al., 2013) and specifically by the work of Noble et al. (2019) who published the design of such a panel, and results of testing it in a human cell line. If these sgRNA backbone variations are similarly efficacious in mosquito cell lines, they represent a step towards improving transgene stability by minimising sequence homology within and between transgenes where multiple sgRNAs are expressed in a single individual.

Decreasing the length of the RNA pol III promoter

As noted in Chapter 4 and in the corresponding publication, Anderson et al. (2020), there are no obvious homologies or motifs amongst the ~600bp U6 RNA pol III promoter sequences beyond those noted in the literature and initially used to identify the putative promoters *in silico*. The PSE and TATA-like box are each noted ~65bp and ~20bp 5' of the snRNA in these sequences, as described in the literature for U6 promoters in general. The restriction of conserved motifs to this region within 100bp of the translation initiation site suggests that

the guideline of 600bp may not be necessary for promoter activity. This hypothesis can be readily tested using the CRISPRa dual luciferase assay method described in Chapter 4.

Sequence modifications in the sgRNA 'backbone'

A single guide RNA (sgRNA) is a synthetic amalgamation of the two RNAs that make up a naturally occurring guide RNA - crisprRNA (crRNA) and tracrRNA (trRNA). crRNA is complementary to the DNA target sequence and fuses to the scaffold trRNA sequence that is the conserved 'backbone' of the guide RNA, mediating Cas protein binding (Jinek et al., 2012). Work has been done to examine changes in binding efficiency of sgRNA mediated by changes to the backbone sequence (Noble et al., 2019, Dang et al., 2015, Briner et al., 2014, Chylinski et al., 2013), demonstrating that there are several nucleotide positions tolerant of variation. Noble et al. (2019) built on this research to test of a panel of 32 sgRNA 'variants' in a human cell line with some success.

If these results can be validated in a cell line model of mosquito species of interest, then these variants could be used in a system expressing multiple sgRNAs (against different target loci) to decrease sequence repeats in the transgene. The efficacy of different sgRNA backbone sequences can be assessed using the CRISPRa dual luciferase assay and by maintaining a constant target sequence of the sgRNAs (TetO_2, as used in Chapter 4).

The repetitive nature of sgRNA sequences is a cause for concern in transgene stability as sequence repetition can encourage undesired homologous recombination within or between transgenes in a single individual (Simoni et al., 2014). This is particularly urgent for constrained gene drives where recombination could join two transgene cassettes and create an unconstrained, global gene drive (Noble et al., 2019).

Panel of sgRNA 'backbone' variants

Remaining within the broader focus of building a constrained CRISPR/Cas gene drive in mosquito species of interest (specifically a Daisy-chain drive), the panel of sgRNA variants described by (Noble et al., 2019) was directly copied for testing in mosquito cell lines. Figure 26 shows the full sgRNA panel tested by Noble et al. (2019). Several constructs from their pilot study were not carried through to their full experiment, the results of which are shown

in Appendix E. The full panel of sgRNA "variants" described in Figure 26 were synthesised with TetO_2 protospacer, for CRISPRa assay testing in mosquito cell lines.

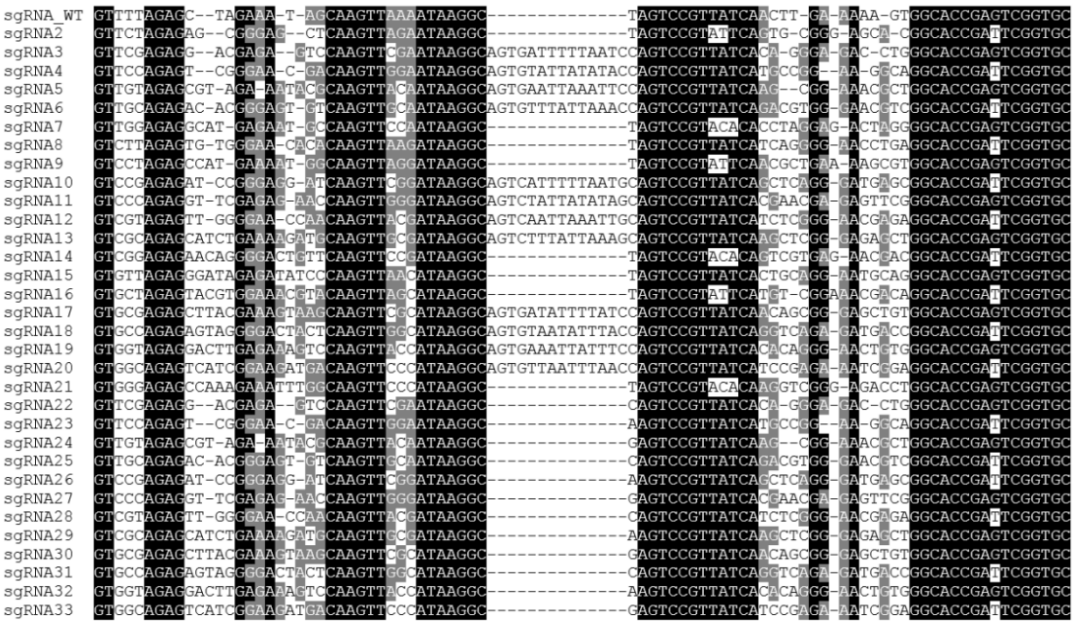


Figure 26: Alignment of a panel of sgRNA sequences that include variations in the otherwise conserved 'backbone' region. This figure is replicated from Noble et al. (2019), supplementary information, who test these sgRNA variants in a human cell line with a transcriptional activation assay, similar to CRISPRa.

Methods and Materials

Plasmids

Plasmid components of the CRISPRa assay (dCas9-VPR expressing plasmid AGG1068, reporter plasmid AGG1202 and Renilla luciferase plasmid AGG1080) were carried forward from work described in Chapter 4.

sgRNA expressing plasmids using truncated lengths of novel RNA pol III promoter sequences were generated through custom order of synthetic DNA fragments (Twist Biosciences, USA) that were then cloned into pJet vector backbones (Fisher Scientific, UK). This work was kindly carried out by Michelle Anderson, Sebald Verkuijl and Josh Ang.

In vitro transcribed sgRNAs

Sequences for each sgRNA variant were custom synthesized as DNA oligonucleotides (oligos) (Appendix Table 20) Oligos were annealed according to standard methods and then *in vitro* transcribed using MEGAscript T7 kit (Ambion, USA). This work was kindly carried out by Michelle Anderson. A negative control sgRNA, specific for an irrelevant sequence (Kmo447) was used in some experiments. This was made using LA925 (Appendix Table 20) and used the sgRNA_WT backbone.

CRISPRa assay

Cell transfections were carried out according to the principles determined in Chapter 4 (page 96).

Endonuclease assay

To mitigate for experimental artifacts arising from the use of dCas9 in the CRISPRa assay (as opposed to endonuclease-capable Cas9), a cell based endonuclease assay was developed by Sebald Verkuijl. This assay uses an adapted luciferase reporter plasmid to generate signal in the presence of specific endonuclease activity (Figure 27). By transfecting cells with the reporter plasmid (AGG1217), a Cas9 expressing plasmid (HR5-IE1-Cas9, AGG1089), the *iv* sgRNA and *Renilla* luciferase plasmid (AGG1080), a relative measure of endonuclease activity can be taken. This assay is not expected to be as sensitive as the CRISPRa assay, but provides valuable corroboration of CRISPRa results for this experiment.

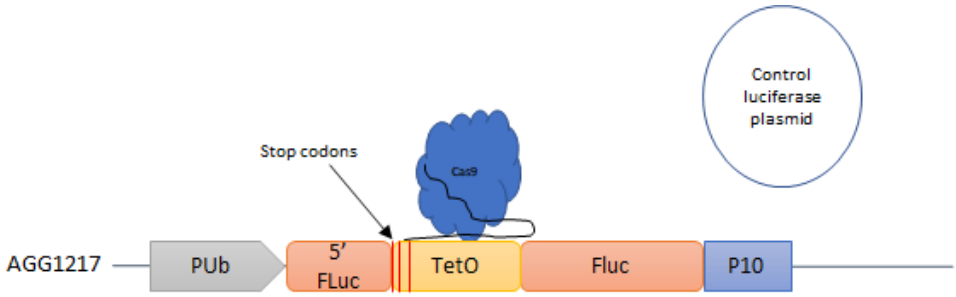


Figure 27: Representation of a dual luciferase assay endonuclease assay that can be used in cell culture to measure abundance or efficacy of TetO specific sgRNA(s). Sebald Verkuijl kindly designed and constructed a firefly luciferase (FLuc) reporter plasmid (AGG1217) that is intended to report specific endonuclease activity. This was achieved first by disrupting FLuc expression, introducing a TetO sequence with premature stop codons in each frame, near the 5' of the coding sequence. The disrupting sequence is followed by the entire FLuc coding sequence, leading to a 5'FLuc – TetO – Fluc design. When the TetO sequence is cut by endonuclease activity, there will be a percentage of repair events that use the homology of the 5' FLuc sequence at each side of the cut to make a repair that codes for functional FLuc protein; this output is expected to reflect the frequency of endonuclease activity. The reliance of reporting on the repair mechanism of the endonuclease activity is expected to reduce assay sensitivity as compared with CRISPRa.

Dual luciferase assay

Dual luciferase assays were carried out according to previously determined principles (Chapter 3). For the experiment shown in Figure 28, 2µl of lysate was used for the dual luciferase assay. For Figure 29, 5µl. Figure *30* data for cell line Aag2, which used 5µl lysate, and Hsu, which used 2µl . Figure 31 was generated from data in cell line Aag2 and uses 1µl lysate for each assay type.

Analysis

Dual luciferase assay results were screened for quality as previously described (Chapter 3, 4) and results were visualised using GraphPad Prism (Version 9.0.0 for windows; GraphPad Software, USA). Basic statistical analysis was carried out using the same software and is described alongside each data set.

Results and Discussion

Reducing the length of the RNA pol III promoter

To test shorter RNA pol III promoter sequences, an initial panel of seven novel U6 RNA pol III promoter sequences were synthesised at four lengths (bp) each: ~600, 400, 200, 100. Each of these was cloned into a pJet vector backbone, expressing the same TetO_2 sgRNA. These plasmids (a panel of 28) were transfected into Aag2 (*Ae. aegypti*) cells and into Hsu (*C. quinquefasciatus*) cells for measurement of expression activity through the CRISPRa dual luciferase assay. The experiment was conducted in both cell lines in parallel (using a master mix of transfection components wherever possible) and care was taken to maintain the *iv* sgRNAs on ice until the last possible moment (to minimise RNase-mediated degradation).

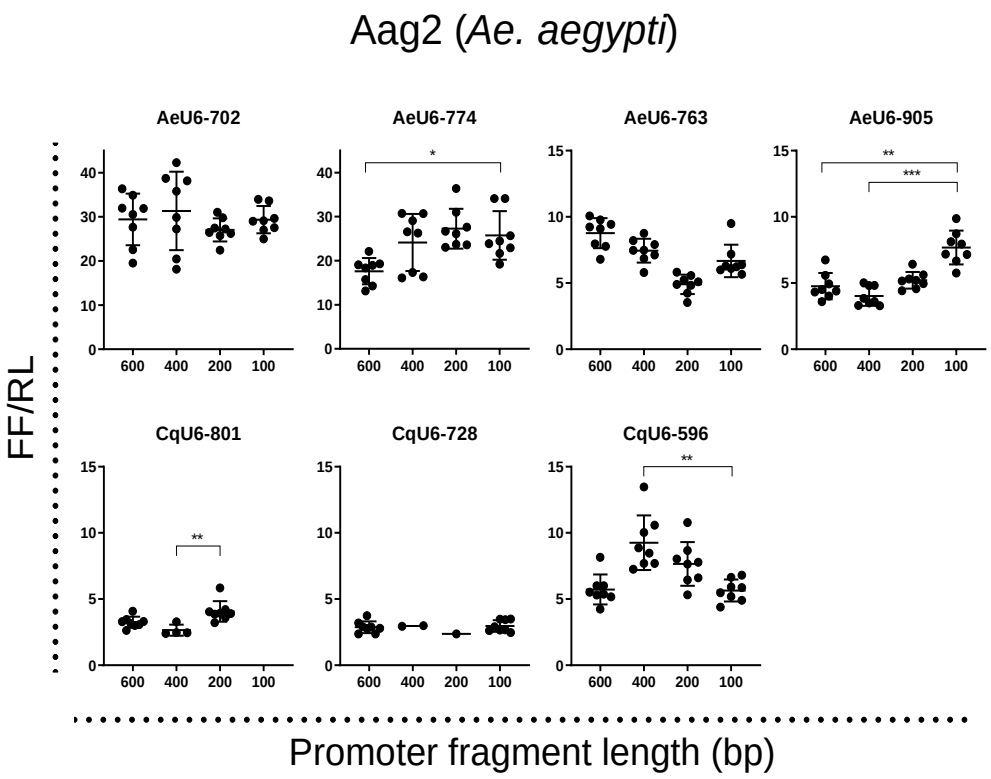


Figure 28: Panel of graphs showing FF/RL activity of seven RNA pol III promoters at four sequence lengths each in cell line Aag2. FF/RL activity for each promoter is measured on the y-axis in ALU, different graphs have independent scales. Each graph shows activity for a different promoter identity, all tested in cell line Aag2. The promoter sequence fragment length (measured as sequence 5' of translation initiation) is noted on the x-axis of each graph (bp). Data points are shown for each repeat for each experiment, with mean and SD indicated. Where a result is not significantly different from background, it is excluded from the figure. Where more than 3 data points are present, basic statistical analysis was done to determine which promoter fragment lengths are significantly different from the shortest fragment. Where $P < 0.05$, this is indicated with one asterisk ("**" for $P < 0.01$ and "***" for $P < 0.001$).

Figure 28 shows the results for decreasing promoter fragment length for seven U6 promoters in *Ae. aegypti* cell line Aag2. Each U6 promoter is represented on an independent graph, with *Ae. aegypti* and *C. quinquefasciatus* promoters on different lines. Each promoter was tested at four sequence lengths, 600bp, 400bp, 200bp and 100bp, measured 5' of the G nucleotide initiation site of the U6 snRNA and shown on the x-axis. Two y-axis scales are used, based on expression levels of the promoter in question. FF/RL is presented on the y-axis, transformed to the mean value of No-sgRNA control for this experiment to eliminate background expression from the unstimulated reporter plasmid. Values that are not greater than the upper 99.9% CI of the No-sgRNA control (Aag2: 2.35 ALU, Hsu: 1.59 ALU) have been excluded from analysis and are not pictured (note that CqU6-801 100bp is excluded entirely on this premise). Individual values for each group are reported to indicate where N < 8 due to such exclusions. The mean and standard deviation are noted where present.

Basic statistical analysis was done within each promoter group to determine whether any sequence length showed activity significantly different from that of the shortest fragment. Where there is a significant difference, it is noted with an asterisk ("*" P <0.05; "**" P< 0.01; "***" P< 0.001). Where a group has fewer than three data points, it has been excluded from statistical analysis. Data in Figure *29* is presented in the same manner and shows results from the duplicate experiment in *C. quinquefasciatus* cell line Hsu. As N ≤ 8 for each group, there is not sufficient statistical power to carry out more complex analysis of promoter fragment length between different promoters or between cell lines.

In cell line Aag2 (Figure 28), *Ae. aegypti* promoters AeU6-702 and AeU6-774 show a higher level of activity than the other five promoters, which is reflected in an adjusted y axis scale. These activity levels are in line with results seen in Chapter 4. These results are described below:

- AeU6-702 No significant difference between promoter fragment lengths.
- AeU6-774 Significant difference only between 600bp and 100bp fragments; the mean activity for each fragment length is ~19 ALU (600bp) and ~25 ALU (100bp).
- AeU6-763 No significant difference between promoter fragment lengths.
- AeU6-905 There is a significant difference between 600bp and 100bp and between 400bp and 100bp; the mean activity for each fragment length is ~5 ALU (600bp and 400bp) and ~7.5 ALU (100bp).

- CqU6-801 Expression from the 100bp fragment was not significantly different from background expression of the assay. Some individual values for each other fragment length were excluded for the same reason, but each of 600bp, 400bp and 200bp had mean activity greater than background (~3 to 4 ALU). There is a significant difference between 400bp and 200bp promoter fragments, which respectively have mean activity of ~2.5 ALU and ~ 4 ALU.
- CqU6-728 No significant difference between promoter fragment lengths. Several data points were excluded for not exceeding the assay's background threshold.
- CqU6-596 Significant difference only between 400bp and 100bp fragments; the mean activity for each fragment is ~9 ALU (400bp) and ~6 ALU (100bp).

Ae. aegypti U6 promoters tested in *Ae. aegypti* cell line Aag2 are consistently unencumbered by decreasing fragment length to 100bp, when activity is measured by the CRISPRa dual luciferase activity. This trend is less clear with *C. quinquefasciatus* U6 promoters in the same *Ae. aegypti* cell line, where the 100bp promoter fragments appear to decrease activity as compared to the longer fragments (twice out of three promoters). It is noted that these changes in activity are small and that the relative activity of these CqU6 promoters in cell line Aag2 is low.

In cell line Hsu (Figure 29), AeU6-702 and AeU6-774 are more active than the other promoters, as they were in cell line Aag2 (Figure 28). Promoter CqU6-596 is more active in cell line Hsu than it was in cell line Aag2, with mean activity typically exceeding that of AeU6-702 or AeU6-774 (in Hsu cells); the y-axis scales are adjusted to reflect this. Figure 29 shows that only these three, higher expressing promoters show any significant difference between fragment lengths in the CRISPRa assay - for promoters AeU6-763, AeU6-905, CqU6-801 and CqU6-728 there is no significant change in mean activity based on fragment length. Of note, there are several instances amongst these promoters where data has been excluded for not exceeding the background threshold and statistical analysis was not done for any group with fewer than three data points.

In cell line Hsu (Figure 29),there is significant difference when comparing the mean activity of 600bp and 400bp promoter fragments with the 100bp fragment for both AeU6-702 and AeU6-774. For AeU6-702, mean activity at 100bp is ~7 ALU, and for 600bp and 400bp it is ~12 ALU and ~15 ALU respectively. For AeU6-774, mean activity at 100bp is ~14 ALU, and for

600bp and 400bp it is ~8 ALU and ~11 ALU respectively. The Kruskal-Wallis statistical test used to analyse the datasets shown in Figure 28 and Figure 29 was two-tailed and does not report a direction for the difference between two groups, so there can be no statistical confidence in whether one group has greater mean activity than another, only that the two are statistically significantly different from one another.

For promoter CqU6-596 in cell line Hsu (Figure 29), there is no significant difference between 600bp fragment length and 100bp fragment, but there is a significant difference between each 400bp and 200bp fragments each compared to the 100bp fragment. 400bp and 200bp fragments have mean activity ~34 ALU whereas 100bp and 600bp fragments each have mean activity ~15 ALU.

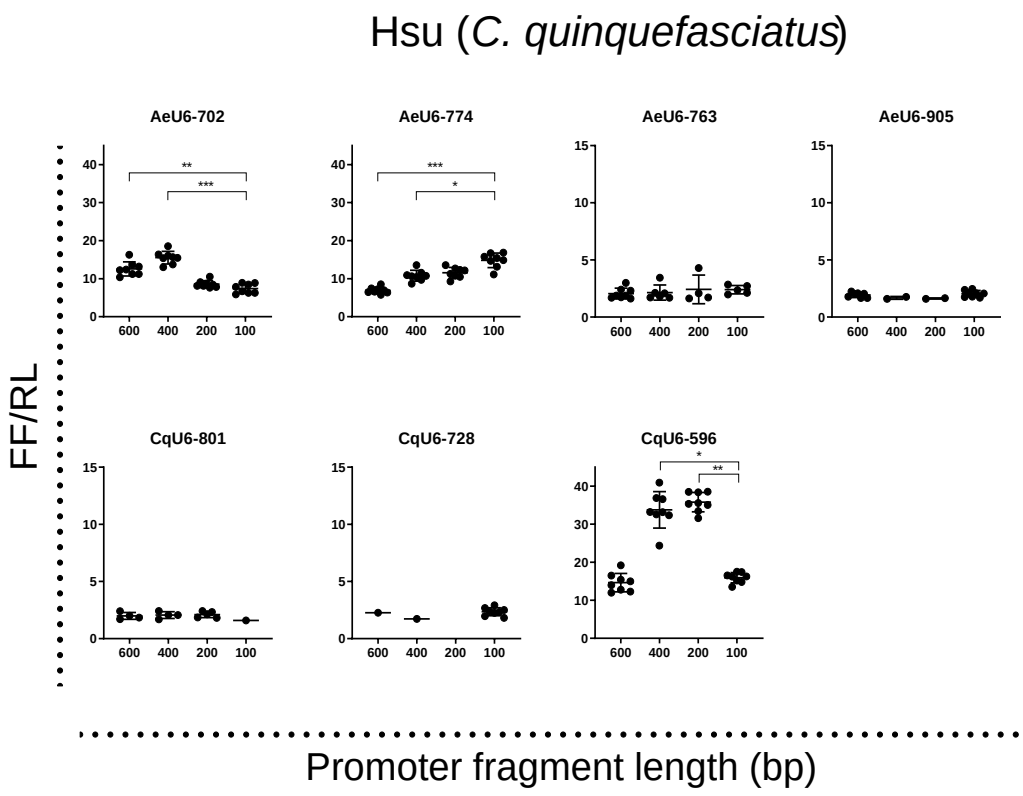


Figure 29: Panel of graphs showing FF/RL activity of seven RNA pol III promoters at four sequence lengths each in cell line Hsu. This data set was generated concurrently with the data shown in cell line Aag2 (Figure 28) and is presented in the same format. Data is separated by RNA pol III promoter identity (graph) and by promoter fragment length (bp) on the x-axis. FF/RL is reported on the y-axis, with independent scales per graph. Where a data point was not significantly different from the background threshold, it is excluded. Remaining data points are shown individually with mean and SD indicated. Where at least three repeats are present, basic statistical analysis was carried out to determine if each fragment length is significantly different from the shortest fragment length. Where there is a significant difference, the comparison and result is noted with an overhead line and asterisk ("*" $P < 0.05$; "**" $P < 0.01$; "***" $P < 0.001$).

Looking at the data in Figure 28 and Figure 29 holistically, it must be considered that these results were all generated from a single experiment with a single measurement of plasmid concentration per plasmid used to transfect each well of cells, for both cell lines. Without looking at repeated experimental data, it cannot be excluded that the pattern in mean activity for promoter fragments of CqU6-596, for example, is an artifact of errors in pipetting or of nanodrop measurement of nucleic acid concentration. Nonetheless, conclusions can be broadly drawn across the repetition of seven different U6 promoters and across two cell lines.

Comparing between data generated in cell line Aag2 versus cell line Hsu, there is not a distinguishable effect of cell line on the relationship between fragment lengths, only on the overall activity of each promoter (i.e. the identity of the promoter with the greatest activity varies by cell line, but the relationship between fragment lengths of a single promoter does not vary meaningfully by cell line). This is consistent with results generated in Chapter 4 and with our proposed understanding of the mechanism of any change in activity between promoter fragment lengths - that there are not obligate promoter motifs more than 100bp 5’ of translation initiation for U6 promoters.

Looking at the effect of promoter fragment length on activity of each of seven promoters across two cell lines, it is concluded that decreasing promoter fragment length from 600bp to 100bp does not typically decrease promoter activity. There are nuances in this trend that could be further explored with repetition of the experiment (as discussed around promoter CqU6-596). If these results can be validated with repetition, then decreasing U6 promoter sequence length to 100bp 5’ of the translation initiation nucleotide (“G” for U6 snRNA) is a viable strategy for reducing overall transgene size where design calls for multiple RNA pol III promoters each expressing different sgRNAs.

sgRNA backbone variants – CRISPRa assay

An initial panel of sgRNA backbone variants were tested in cell lines Aag2 and Hsu using the CRISPRa dual luciferase assay; these results are presented in Figure *30*.

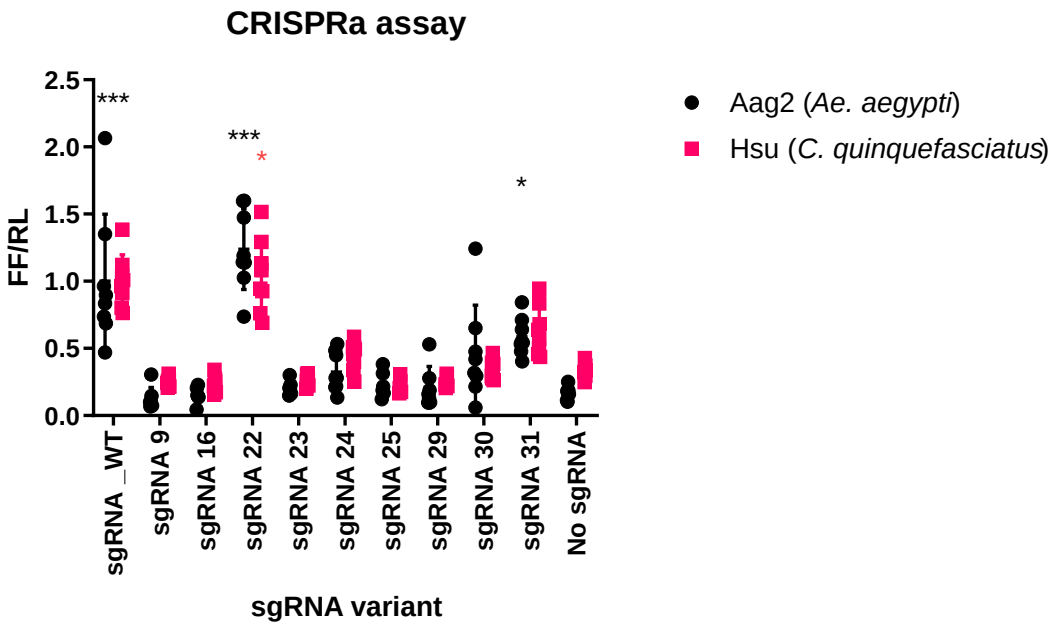


Figure 30: Graph showing results from CRISPRa assay testing a panel of sgRNA variants (altered backbone sequences) in cell lines Aag2 (black) and Hsu (pink). Results are shown as individual data points with mean and SD indicated. Data is reported on the y-axis as FF/RL relative to sgRNA_WT as 1 (for each data set). Each sgRNA variant identity is shown on the x-axis with the negative control group No sgRNA presented on the far right. Where a group has a mean significantly different from background, it is indicated by asterisk in the colour for that dataset ("*" $P < 0.05$; "***" $P < 0.001$).

Results in Figure *30* are presented relative to the activity of the standard sgRNA backbone sequence, labelled "sgRNA_WT" for each cell line, in keeping with the presentation of Noble et al. (2019). It must be noted that the negative control group, "No sgRNA" is included at the far right of the x-axis and that no values have been excluded on the basis of falling within the 99.9% CI of the No sgRNA mean (for each cell line). In deference to the low or absent activity of several experimental groups, data is instead presented only with values above the Firefly luciferase threshold ($1x10^6$ ALU) excluded. Basic statistical analysis was carried out to determine which experimental groups are significantly different from background expression (the negative control, No sgRNA). A Kruskal-Wallis test was used with Dunn's multiple comparison to compare each group against No sgRNA, the P-values for each test are represented in Figure *30* as asterisks above groups that are significantly different ($P < 0.05$ "*", $P< 0.001$ "***").

First comparing the two cell lines, Aag2 and Hsu, represented by interleaved colours in Figure 30, it is noted that there is no meaningful difference in activity of any experimental group as compared to their corresponding sgRNA_WT. The mean FF/RL of No_sgRNA control for cell line Aag2 appears to be slightly lower than that of cell line Hsu; this difference is suggested to explain the differences between cell lines in which experimental groups show statistically significant mean values from the negative control. There was no expectation of cell line identity influencing the relative activity of sgRNA backbone variants.

In cell line Aag2, there are three experimental groups that show statistically significant differences in mean from that of the negative control: sgRNA_WT, sgRNA 22 and sgRNA 31. For cell line Hsu, only sgRNA 22 has a mean statistically different from No_sgRNA. This is an unexpected outcome as every variants included here shows significant activity in Noble et al. (2019). That sgRNA_WT and some sgRNA variants have activity different from background demonstrates that there was no global failure of the assay.

Out of an abundance of caution, the sgRNA sequences were each re-confirmed *in silico* and the *in vitro* transcribed sgRNAs were kindly re-synthesised by Michelle Anderson. This was done to mitigate the possibility of RNAse contamination and degradation having compromised the initial experiment.

sgRNA backbone variants – CRISPRa and CRISPR endonuclease assays

A complete panel of *in vitro* transcribed sgRNA variants were synthesised alongside reproduction of the sgRNAs used in the initial experiment. These were transfected into a single cell line, Aag2, and were concurrently used in an additional CRISPR endonuclease assay. The dCas9-VPR protein used in the CRISPRa assay is a modification of Cas9 protein, to remove endonuclease activity. There is a possibility that this change in activity of the Cas-sgRNA-DNA complex causes changes in the binding kinetics of the complex (i.e. does the complex dissociate at the same rate if there is no enzymatic activity). To mitigate this risk, an additional assay was developed by Sebald Verkuijl and uses an endonuclease reporter plasmid that contains a TetO sequence corresponding to TetO2_sgRNA (Figure 27).

At a baseline, the reporter plasmid (AGG1217) expresses non-functional Firefly luciferase, a mutant created by inclusion of the TetO sequence and premature stop codons at the 5’ of the coding region. In the presence of Cas9 and TetO2_sgRNA, the included TetO sequence is cut and a double stranded break is created. In a proportion of these events, the cut plasmid will be repaired by non-homologous end-joining in a fashion that causes gain of function of the Firefly luciferase gene; this in turn is detected in a subsequent dual luciferase assay.

This endonuclease assay is less sensitive than the CRISPRa assay as not every TetO endonuclease event will result in a measurable outcome (gain of Firefly luciferase function). Furthermore, there is a higher background expression of Firefly luciferase from the intact reporter plasmid than is seen in the CRISPRa assay. For these reasons, the endonuclease assay is not used as a primary assay in sgRNA experiments. It is suitable, however, for use in this circumstance where we would like to exclude the deactivated Cas9 in the CRISPRa assay as a confounding variable or causal agent of the unexpected results. Further validation of this assay is included in Appendix Figure 6.

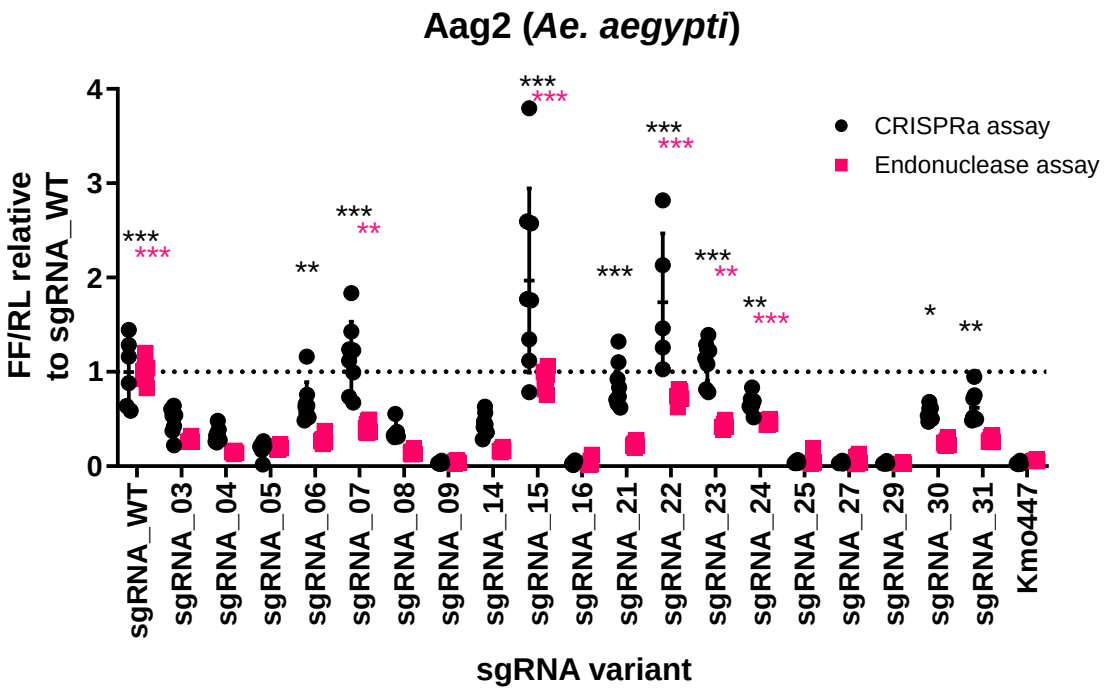


Figure 31: Graph showing relative activity of a panel of sgRNA variants in two cell culture assays, tested in cell line Aag2. Each assay is differentiated by colour and denoted in the legend. Data is reported as FF/RL transformed to the mean of sgRNA_WT as 1 for each assay (y-axis), emphasised with a dotted line. The sgRNA identity is reported on the x-axis, with sgRNA_WT on the far left and the negative control (an irrelevant sgRNA, Kmo447) at the far right. Data points are shown individually for each repeat with mean and SD indicated. Where a group is significantly different from background, it is indicated with asterisks (colour coded per assay) ("*" $P < 0.05$; "**" $P < 0.01$; "***" $P < 0.001$).

Figure 31 uses the same format as Figure 30 and shows the results for cell line Aag2 of 19 sgRNA backbone variants tested with two assays, relative to sgRNA_WT as a positive control and an irrelevant sgRNA, Kmo447, as a negative control. The results from each the CRISPRa assay and the Endonuclease assay are shown as interleaved colours and each dataset is transformed to the mean of its sgRNA_WT results as 1. A horizontal line is included at y = 1 for enhanced clarity. The sgRNA variants are shown on the x-axis with the positive control

(sgRNA_WT) on the far left and the negative control (Kmo447) on the far right. The y-axis shows results as FF/RL relative to sgRNA_WT. Firefly luciferase data was quality controlled to exclude any results above the FF threshold and each biological replicate is shown as an individual symbol. Mean and standard deviation are indicated where visible.

As in Figure 30, simple statistical analysis was carried out for each dataset to confirm if there was a significant difference in means attributable to the experimental groups (Kruskal-Wallis test) with Dunn's multiple comparison follow up used to test if the mean for each group is significantly different from the mean of the negative control group (Kmo447). Where there was a significant difference, this is indicated with asterisks above the results ("***" $P < 0.05$; "**" $P < 0.01$; "*" $P < 0.001$).

First comparing the two assays, it appears that the mean results for each experimental group are lower in the endonuclease assay than for the same group in the CRISPRa assay. This corroborates expectations that the endonuclease reporter plasmid would be less sensitive than the CRISPRa reporter plasmid. There is, however, significant expression from the positive control (sgRNA_WT) in the endonuclease assay and the results for each sgRNA variant do mimic those seen with the CRISPRa assay. This corroborates expectations that the dCas9 does not have meaningfully different sgRNA-dsDNA binding activity as compared to Cas9 and offers validation of the use of a CRISPRa assay as a model for endonuclease activity in this scenario (further supported by Appendix Figure 6).

There are five sgRNA variants that appear to have mean activity at the same level or higher than sgRNA_WT: sgRNAs 7, 15, 21, 22 and 23. Three further variants have means lower than sgRNA_WT but are significantly different from background in at least the CRISPRa assay: sgRNAs 6, 24, 30 and 31. These results are encouraging as they validate nine sgRNA backbone sequences that can be used to minimise sequence homology in a transgene or transgenes expressing multiple sgRNAs in the same individual or cell. Based on the consistency of results between Aag2 and Hsu cell lines, it is expected that these results could be applied to mosquito or insect species of interest more widely. There are hesitations, however, raised by the discrepancies between the results in Figure 30 and Figure 31 as compared to the results shown in Noble et al. (2019), reproduced in Appendix E and summarised in Table *23*.

Comparing results in Aag2 with those reported in a Human cell line

Table *23* is presented as a summary of mean (transformed against WT) results for each sgRNA variant in each of the three experiments described here and additionally those of (Noble et al., 2019) (grouped by Author, cell line and assay type).

Table 23: Mean activity of each sgRNA variant, relative to sgRNA WT from each experiment in this chapter and additionally the results of Noble et al. (2019)

Author	**Noble**	**Purcell**			
Cell line context	**Human**	**Hsu**	**Aag2**	**Aag2**	**Aag2**
Assay	**CRISPRa**	**CRISPRa**	**CRISPRa**	**CRISPRa**	**Endonuclease**
sgRNA WT	1	1	1	1	1
sgRNA 2	0.75	-	-	-	-
sgRNA 3	0.5	-	-	0.5	0.3
sgRNA 4	1	-	-	0.3	0.2
sgRNA 5	0.9	-	-	0.2	0.2
sgRNA 6	0.6	-	-	0.7	0.3
sgRNA 7	1	-	-	1.2	0.4
sgRNA 8	1	-	-	0.4	0.2
sgRNA 9	1.75	0.2	0.1	0	0.1
sgRNA 14	0.8	-	-	0.5	0.2
sgRNA 15	1.1	-	-	2	1
sgRNA 16	1	0.2	0.1	0	0
sgRNA 21	1.1	-	-	0.9	0.2
sgRNA 22	1	1	1.2	1.7	0.7
sgRNA 23	1.4	0.3	0.2	1	0.4
sgRNA 24	1	0.4	0.3	0.7	0.5
sgRNA 25	1.4	0.2	0.2	0	0.1
sgRNA 26	0.5	-	-	-	-
sgRNA 27	0.5	-	-	0	0.1
sgRNA 28	0.5	-	-	-	-
sgRNA 29	1.5	0.2	0.2	0	0
sgRNA 30	1.5	0.3	0.5	0.6	0.3
sgRNA 31	1.8	0.6	0.6	0.6	0.3
sgRNA 32	0.4	-	-	-	-
-ve control	0	0.3	0.2	0	0.1

These values are an estimate presented to indicate the apparent differences in results generated by each author. These differences are attributed to the host context of the experiment, Human vs mosquito cell lines, and potentially to any differences in the transcriptional activation assay used by Noble et al. (2019) vs the CRISPRa assay used by

Purcell. Values in Table 23 have not been analysed statistically and were generated by eye for Noble et al. (2019) and through the actual mean values of transformed results for Purcell. Whether sgRNA variants were significantly different from background in the Purcell assays is not included in this table, though the mean value for negative controls is noted. Shading of cells in Table *23* is used to indicate results with a mean activity ≥ 1.

Accepting that results generated in Human cell lines do not translate perfectly for experiments in mosquitoes, this data set suggests four sgRNA variants that can be used in Culicinae mosquitoes with at least equal activity to the standard sgRNA sequence. There are a further seven variants that could be confirmed to have activity greater than background (with further experimental repeats), though less than sgRNA_WT. The limitations of these experiments lead to the recommendation that frequency analysis be carried out for endonuclease activity generated by each sgRNA variant where multiple sgRNAs are present in a single individual. It is reasonably expected that each would demonstrate the desired activity, but noted that drops in efficacy as compared to sgRNA_WT may need to be accounted for in simulation models of the effect of such transgenes on a complex biological control strategy. It is furthermore recommended that sgRNA variant sequences be validated in a species specific representation of each insect species of interest, particularly in light of the unexpected discrepancies between results reported in Human cell culture as compared to those shown in mosquito cell culture.

Conclusion

Additional use of the CRISPRa dual luciferase assay format established in Chapter 4 has offered results that suggest viable improvements to transgene design, particularly for transgenes expressing multiple sgRNAs in a single individual (or cell). Reducing the sequence fragment length of the RNA pol III U6 promoters was particularly successful and suggests that a meaningful decrease in transgene size can be achieved, which is expected to increase the efficiency of genomic integration of the transgene. Seven U6 promoter sequences are explored at four sequence lengths each (600bp, 400bp, 200bp and 100bp) in each *Ae. aegypti* cell line Aag2 and *C. quinquefasciatus* cell line Hsu.

Comparing the activity of 100bp promoter fragments against their 600bp counterparts finds that the change in size is typically not detrimental to the amount of sgRNA then expressed from that plasmid (transgene). These results are less clear in promoters with lower overall expression activity, but have been validated in sufficient numbers to confirm that a 100bp U6 promoter sequence is sufficient for use in transgenes for mosquitoes. This work is built

upon and described further in Anderson et al. (2020). As in previous chapters, this work is limited by the *in vitro*, cell culture nature of the work and it would be advantageous to validate the activity of these 100bp promoters *in vivo*. Particularly, the question of interactions between promoters or limitations of cellular machinery to express sgRNAs from multiple U6 promoters in a single individual is not explored in this work. Such a question would be an important avenue of further work, particularly as results throughout this project (Chapters 3, 4 and 5) have often come across outlier or other unexplained events that are not yet predictable. Validated technologies are particularly valuable as creating a transgenic insect line is very labour intensive.

The second set of experiments described in this chapter examine whether variant sgRNA backbone sequences described by Noble et al. (2019) and presented in a Human cell context can be adopted for use in mosquitoes. Although the full panel of 20+ variants are not all suitably active in mosquito cells, several are and there are approximately 10 sgRNA backbone sequences that could be used to create multi-sgRNA transgenes with minimised sequence homology. This could be sufficient for suggested iterations of complex CRISPR population control strategies (e.g. tripartite Daisy-chain gene drive) and fulfils an important technological requirement for realising such a system.

It is unexpected that there is such a marked difference in activity between sequences described as active in a Human cell line and their activity in mosquito cell lines. The cellular machinery involved in CRISPR/Cas activity is highly conserved and not expected to differ between these species. Nonetheless, cell culture is an imperfect model and this is not the only instance of unexpected variation between cell lines noted within this project. This limitation, and that of a transcriptional activator (CRISPRa) assay as a model for Cas9 gene editing *in vivo*, guide the recommendation that any further such technologies be validated in as close a model as is practicable before investing resources into its inclusion in a transgenic insect line.

Chapter 6: Conclusion

The experiments described in this thesis sit within the aim of developing the toolkits available to researchers in pursuit of implementing complex CRISPR population control designs in mosquitoes and other pest arthropods – particularly a 'Daisy chain gene drive' in *Ae. aegypti*. This work was carried out exclusively in cell culture models of species of interest and used a 96-well plate, dual luciferase assay format to increase experimental capacity.

Modulation of transgene expression through translational modification

By adapting published data on the effect of the translation initiation sequence (TIS), on its own and in conjunction with the 3' untranslated region (UTR) sequence, it was determined that such sequences can also impact the rate of protein expression in mosquito cell culture between 2 fold and 10 fold. Five TIS sequences were tested in each combination with three 3'UTR sequences, in four mosquito cell lines and one moth cell line. This data was analysed through a custom statistical model. In particular, it was found that choice of TIS sequence alone could upregulate reporter gene expression by 1.12 fold and could downregulate expression by 2.27 fold (from "Kozak" TIS sequence as a standard).

That these changes could be achieved is in line with previous reports (e.g. Pfeiffer et al. (2012)). The magnitude of change, however, was smaller than those reported by similar work in *D. melanogaster* (Pfeiffer et al. (2012) report 7.5 fold increase) and *B. mori* (Tatematsu et al. (2014) report 4 to 10 fold increase *in vitro*). This does not preclude use of such sequences to modulate transgene expression in future work, but perhaps tempers expectations.

The immediate application of these findings lies in their use for transgene design, specifically for upregulating expression of a transgene (e.g. a fluorescent marker or toxic effector) or for downregulating expression of a transgene (e.g. a toxic effect with 'leaky' expression from its inducible promoter).

Although different constitutively active promoters can have different rates of mRNA expression in different insects, the RNA polymerase (pol) II promoter used for transgenics also controls the temporal and spatial activity of transgene expression and these factors are often primary in the choice of promoter for a transgene. It cannot be examined in cell culture experiments, but there is no indication in the literature (Pfeiffer et al., 2012, Tatematsu et al., 2014) that the TIS can effect temporal or spatial activity of a promoter. This frames the TIS as an opportunity to influence transgene expression efficiency, independent of the

promoter sequence. The 3'UTR is less independent of the temporal and spatial profile of a gene's activity, but presents a further opportunity to adjust transgene expression efficiency in a transgene. It would be interesting to be able to conduct such work in an empirical fashion, but the TIS sequences and 3'UTR sequences validated and characterised here could be used immediately in design of future transgenes where expression rate is anticipated to cause difficulties.

Follow on work in this area examined the consistency of results across a panel of (constitutively expressed) RNA pol II promoters. This work is as yet unpublished, but indicates that the effect of TIS and 3'UTR is independent of promoter identity (personal communication, Phil Leftwich and Michelle Anderson). Cell culture models are a poor representation of any specific insect tissue (they originate from crushed embryos, larvae or pupae) and so work with TIS and tissue-specific promoters was not explored. This would be an interesting avenue of exploration for *in vivo* experiments.

Tatematsu et al. (2014) raise an interesting line of enquiry around the use of tissue-specific TIS in transgenes with an intended tissue-specific expression profile. Their work is in the context of *B. mori*, a species used commercially to express exogenous proteins, and has therefore focused on maximising transgene expression rates. Although this commercial aspect does not exist in mosquitoes, there is an interest in expressing transgenes in very tight tissue specific fashion for activity in germ cells (e.g. Hammond et al. (2016) and Hammond et al. (2021)). If the resources were available to carry out germ cell expression experiments, it would be interesting to compare the activity of a native TIS against that of, for example, "Lep", and examine whether there is an associated difference in expression rates.

An *in vitro* CRISPRa assay for validating sgRNA activity

The work in Chapter 4 moved away from transgenic expression of proteins and into transgenic expression of sgRNAs for CRISPR/Cas applications. Recognising that direct, quantitative measurement of Cas endonuclease activity was not efficient, a transcription activation assay was adopted as a model for measuring sgRNA activity. This CRISPR activation (CRISPRa) assay was developed in cell lines of interest based on the work of Chavez et al. (2015), who demonstrated the assay in mammalian cell culture. It was furthermore tied to a dual luciferase reporting format, building on the resources established in Chapter 3.

Once characterised and optimised, the CRISPRa assay was used to measure abundance of reporter-sequence (TetO) specific sgRNAs. An experimental pipeline was developed for *in silico* identification of U6 and 7SK RNA pol III promoters and CRISPRa testing of the expression activity of these sequences. Using this method, a panel of published and novel promoters

were validated in mosquito cell lines. Advantage was taken of the non-species specific nature of the CRISPRa assay and this panel was tested in several cell lines, demonstrating a range of cross-species activity and increasing the resource pool for *Ae. aegypti*, *C. quinquefasciatus* and *An. gambiae* transgene expression. A section of this work was published as Anderson et al. (2020).

Although some further work was done with these promoters and with this assay in Chapter 5, a noted direction for additional investigation is the effect on sgRNA expression of multiple promoter-sgRNA transgenes being present in the same cell. This aspect is crucial for the use of complex CRISPR based gene drives and could be tested in the extant assay system (in cell culture).

The 'model' nature of this work (cell culture and CRISPRa rather than whole insect and endonuclease activity) was addressed by recent publication (by other groups) of the activity of the same promoters in whole insects, measured by endonuclease activity (Li et al., 2020). Comparisons across the published work are made in Chapter 4. These findings support the validity of this CRISPRa cell culture approach as a model for insect transgenesis.

Design improvements to express multiple sgRNAs on a single transgene

Chapter 5 developed the findings of Chapter 4 by using the established CRISPRa dual luciferase assay to test design changes intended to improve the design of a transgene expressing multiple sgRNAs. This was looked at in two areas, by reducing the length of the promoter sequence and by using sequence variations in the sgRNA conserved ('backbone') region.

Experiments looking at reducing the size of the U6 promoter fragment from ~600bp to 100bp showed that there was not consistent detriment to expression activity by making this change. If this finding were consistent in whole insects and with use of multiple RNA pol III promoters in a single individual, it would allow a decrease of 500bp per promoter. Complex CRISPR gene drives are recommended to use at least four sgRNAs (Noble et al., 2017) and so this could amount to a ≥2kbp reduction in size of the transgene cassette, with corresponding increase in transgene stability.

The presence of areas of sequence repetition in a transgene are a potential source of instability and unintended homologous recombination. Use of multiple sgRNAs on a single transgene would amount to highly repetitive regions and so Noble et al. (2019) suggest a panel of variants with changes to the conserved sgRNA 'backbone' sequence that reduce this

homology without compromising binding activity of the sgRNA. Noble et al. (2019)'s panel presents results from testing in human cell culture, and this experiment was reproduced using CRISPRa in mosquito cell culture.

There were variations in activity of the same sgRNA variant between human cell culture and mosquito cell culture, which was not expected and is not satisfactorily explained. In spite of this, four variants were identified that appear to not alter binding activity and a further five were shown to retain more than half the binding activity of the 'wild-type' sgRNA sequence (though less than 100% of activity). The veracity of these results was further tested through use of an endonuclease luciferase reporter assay that allows approximate measurement of the endonuclease activity conferred by reporter-sequence (TetO) specific sgRNAs. This assay is less sensitive and has higher background activity than the CRISPRa assay, but was shown with sgRNA variants and sgRNA abundance to produce results (when comparing between experimental groups) that are consistent with results generated by CRISPRa assay.

These two areas represent practical improvements in the tools available for design of transgenes expressing multiple sgRNAs and in turn design of Daisy chain gene drive in mosquitoes. Although linking of different components is a concern for any 'split drive' (where constraint is predicated on the endonuclease being split across more than one transgene), it is a particular risk for a tripartite Daisy chain drive where there are two constructs ("A" and "*C*") that each express multiple sgRNAs and could therefore have considerable sequence homology (Figure *2*). If sequence homology between these constructs facilitated homologous recombination between them, the overall system could be reduced to a pair of constructs (e.g. "A" and "B") that are together able to reproduce themselves. The findings presented in Chapter 5 offer insight to how such sequence homology can be minimised.

It is noted that this work has not been confirmed in whole insects. The inconsistencies between results in human cell culture and mosquito cell culture raises concerns that the cell culture model may be less representative in this case than it has been shown to be for other areas (e.g. testing of RNA pol III promoters).

Summary

The body of this work presents well characterised, high (manual) throughput assays for testing of variations in DNA based elements of transgene design. The CRISPRa assay in particular was used in three areas of work and can be the basis of future work that has yet to be considered. Consistencies between the results discussed here and the literature offer a validation of the use of a cell culture model for testing designs that will be used in whole insects. Similarly, the transcription activation (CRISPRa) assay developed here appears to

function as a good model of endonuclease activity for assessment of sgRNA abundance and binding activity.

These assays are used to validate design elements and improvements that represent significant gains for the community's capacity to thoughtfully design each element of a transgene, particularly in complex multi-part systems.

Outside of the scope of this project, work is still needed to develop the resources to build an effective, safe Daisy chain gene drive. Effector mechanisms for the desired phenotype, RNA polymerase II promoters with highly specific expression patterns and empirical demonstration of the concepts that are currently supported by mathematical modelling are all needed before a gene drive product could be considered to have been achieved. Such a product would need community and government support to be utilised 'in the field'. These requirements are not unique to a Daisy chain gene drive system, but highlight some of the scope of work that is needed.

The epidemiology of mosquito borne diseases is complex, involving interactions of the environment, the vector, the pathogen and the humans (or animals) that are affected. Complex problems often call for multi-factorial solutions and genetic control strategies for reducing size of a vector population are only a part of that picture. As work continues in this field, it must also continue in pharmaceutical, social and medical fields so that we can reach the fastest and most effective amelioration of the burden of these diseases.

Although the work in this project represents a contribution to one specific solution (Daisy chain gene drives in *Ae. aegypti*), it is done in a species non-specific fashion wherever possible and care is taken to ensure that the developed assays are well characterised and reproducible. These considerations are an acknowledgement that future uses of any scientific research cannot necessarily be predicted at the time that the work is done.

Appendix A : Supplemental Methods

Cell line species validation

Introduction

Cell line authentication is an important quality control check for any research done using cell culture, particularly where a cell line is not obtained directly from a verified source such as the American Type Culture Collection (ATCC). Cross-contamination and mis-identification of cell lines is thought to be a pervasive issue in modern scientific research, particularly within the field of human cancer biology (Chatterjee, 2007, ATCC, 2010).

Although the issue is less prevalent in insect cell lines, it remains of concern. As species of origin was the main feature of relevance for the experiments in this thesis, cell lines were species authenticated by 'PCR barcoding'. This was done for each new cryovial or flask of cells received.

'PCR barcoding' for species identification was based on the work of Folmer et al. (1994). Universal primers are used to amplify a conserved region of the mitochondrial gene cytochrome c oxidase subunit I (COI) of a sample, producing an amplicon that is species specific. The sequence of this amplicon can be compared against a database (Ratnasingham and Hebert, 2007) to identify the species of origin of the sample.

Each new cell culture and several extant genomic DNA (gDNA) samples were tested in this way. Fresh gDNA samples taken from visually confirmed adults of lab colonies as a positive control.

Methods

Appendix Table 1: Primers for PCR barcoding

Primer name	Sequence
LA806: LCO1490	GGTCAACAAATCATAAAGATATTGG
LA807: HC02198	TAAACTTCAGGGTGACCAAAAAATCA

Cell or insect samples were collected and their gDNA extracted according to manufacturer's recommendations, using the Machery-Nagel NucleoSpin Tissue Kit (Machery-Nagel, Germany). PCR primers (Appendix Table 1) were custom synthesised (Sigma-Aldrich, UK). The PCR primers, reaction (Appendix Table 2) and thermocycle conditions (

Appendix Table *3*) were adapted from Folmer et al. (1994) for use with Q5 Hot Start High-Fidelity DNA Polymerase (Q5 enzyme) (NEB, UK).

Appendix Table 2: PCR reaction for species validation

Reagent	Volume (µl)
Q5 buffer	5
dNTP (10mM)	0.5
Fw primer (10µM)	1.25
Rv primer (10µM)	1.25
Template gDNA	50ng
Q5 enzyme	0.25
dH_2O	To 25µl

Appendix Table 3: PCR thermocycle for species validation

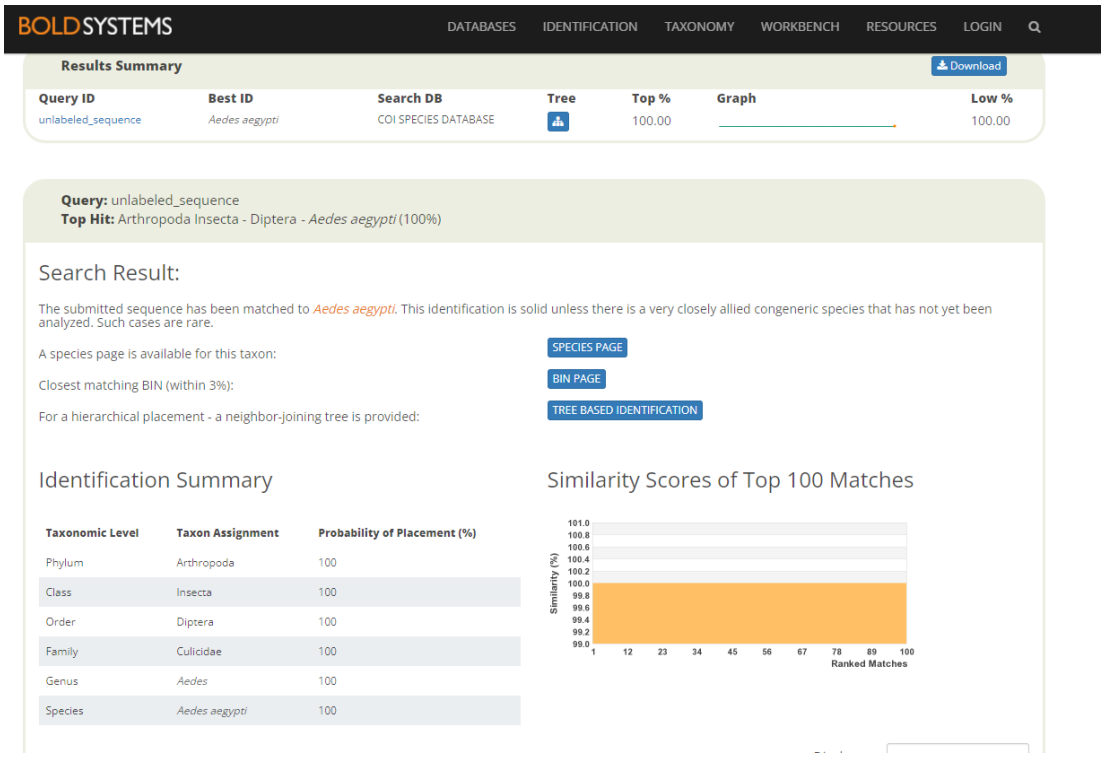
Cycles	Temperature (˚C)	Time
1	95	60s
35	95	30s
	40	30s
	72	60s
1	72	7min
1	16	Indefinite

These conditions often produced non-specific product, which could be cleanly separated from the desired band (720bp) by agarose gel electrophoresis. Purified bands were Sanger sequenced to provide data for comparison with the Barcode of Life Data Systems database ("Identification Engine") (historical databases Jul-2016 and Jul-2017) (BOLDSYSTEMS, 2007, Ratnasingham and Hebert, 2007).

Results

Clean DNA sequence results were obtained for each sample from cell culture or whole insect gDNA. These were compared against the database, which identified the likely sequence identity of the source gDNA. An example of this process is shown in Appendix Figure 1.

Each cell line used in this work was PCR validated with this method and all were correctly identified as the expected species. This was further confirmed by aligning results from cell culture samples against positive control samples from visually confirmed (species) adult insects.



Appendix Figure 1: Example query of BOLD database, using the sanger sequence from a sample from cell line Aag2. This result shows 100% confidence that the sequence originated from an *Ae. aegypti* sample, validating the cell line species origin. Species with very closely related species (i.e. within a species complex) cannot be distinguished from one another (e.g. *C. pipiens* and *C. quinquefasciatus*).

Discussion

PCR species validation with this method offers a simple and convenient method of authenticating cell lines. Although it cannot be used to distinguish between cell lines from the same species, it is a suitable verification tool for the purposes of this work.

Appendix B: Materials

Plasmids

Chapter 3: Modulation of transgene expression through translational modification

Appendix Table 4: Plasmids used in Chapter 3

Plasmid ID	Plasmid Name	Description	Source
AGG1079	pRL-CMV	CMV-*Renilla*-SV40	Promega, UK
AGG1080	pRL-OpIE2	OpIE2-*Renilla*-SV40	JP
AGG1185	pIE1-FF-SV40	HR5-IE1-MCS-FF-SV40	JP
AGG1183	pGL3-Basic	Firefly-SV40	Promega, UK
AGG1032	pB-Hr5Ie1-DsRed-TRE-DsRED	TRE driving DsRed, Red marker	THS; Genewiz
AGG1186	pIE1-Koz-FF-SV40	Kozak ATG context, SV40 3' UTR	JP
AGG1187	pIE1-Lep-FF-SV40	Lepidopteran consensus ATG context sequence, SV40 3' UTR	JP
AGG1188	pIE1-BmHi-FF-SV40	*B. mori* consensus ATG context sequence, SV40 3' UTR	JP
AGG1189	pIE1-BmLo-FF-SV40	*B. mori* least common ATG context sequence, SV40 3' UTR	JP
AGG1190	pIE1-Syn21-FF-SV40	Syn21 ATG context sequence, SV40 3' UTR	JP
AGG1196	pIE1-Koz-FF-P10	Kozak ATG context, P10 3' UTR	JP
AGG1197	pIE1-Lep-FF-P10	Lepidopteran consensus ATG context sequence, P10 3' UTR	JP
AGG1198	pIE1-BmHi-FF-P10	*B. mori* consensus ATG context sequence, P10 3' UTR	JP
AGG1199	pIE1-BmLo-FF-P10	*B. mori* least common ATG context sequence, P10 3' UTR	JP
AGG1200	pIE1-Syn21-FF-P10	Syn21 ATG context sequence, P10 3' UTR	JP
AGG1191	pIE1-Koz-FF-K10	Kozak ATG context, fs(1)K10 3' UTR	JP
AGG1192	pIE1-Lep-FF-K10	Lepidopteran consensus ATG context sequence, fs(1)K10 3' UTR	JP

Plasmid ID	Plasmid Name	Description	Source
AGG1193	pIE1-BmHi-FF-K10	*B. mori* consensus ATG context sequence, fs(1)K10 3' UTR	JP
AGG1194	pIE1-BmLo-FF-K10	*B. mori* least common ATG context sequence, fs(1)K10 3' UTR	JP
AGG1195	pIE1-Syn21-FF-K10	Syn21 ATG context sequence, fs(1)K10 3' UTR	JP
AGG1019	pBac-ZsG-A4-tTAV3	Moth pBac transformation construct	SB
AGG1201	pHR5-IE1-ZsGreen	HR5-IE1-ZsGreen	JP
AGG1024	pBac-ZsGreen-tTAV3	Moth pBac transformation construct	THS
AGG1240	pIE1-FF-P10	HR5-IE1-MCS-FF-P10	JP
AGG1241	pIE1-FF-K10	HR5-IE1-MCS-FF-K10	JP
L5610	pTNT	Plasmid for T7 *in vitro* reactions	Promega

Chapter 4: An *in vitro* CRISPRa assay for validating sgRNA activity

Appendix Table 5: Plasmids used in Chapter 4

Plasmid ID	Plasmid Name	Description	Source
AGG1020	pUC57-pB-AmCyan-TRE-DsRed	Insect transformation plasmid, constitutively expressed AmCyan and tTa inducible DsRed	THS; Genewiz
AGG1068	pB-HR5/IE1-dCas9.VPR-P10	constitutive expression of dCas9-VPR	THS; Genewiz
AGG1078	pUC57-pB-AmCy-TRE-FF-SV40	Identical to AGG1020, with DsRed CDS replaced with FF CDS	JP
AGG1080	pRL-OpIE2	OpIE2-*Renilla*-SV40	JP
AGG1092	PxU6 (3) TetO_sgRNA	Three PxU6 prmoters each expressing TetOsgRNAs (1, 2 and 3)	THS
AGG1094	pUC57KAN_U63-KMO-rice tRNA-TetO	AeU6-702 driving Kmo_sgRNA_OstRNAGly_TetO_sgRNA	SV
AGG1120	*AeU6-702	AeU6-702 expressing TetO_sgRNA2	Twist Biosciences
AGG1131	CqU6-596	AqU6-596 expressing TetO_sgRNA2	Twist Biosciences
AGG1155	AeU6-702_KmosgRNA_TetOsgRNA2	AeU6-702 expressing KmosgRNA_TetOsgRNA2 dimer	Twist Biosciences

Plasmid ID	Plasmid Name	Description	Source
AGG1164	*AgU6-695	AgU6-695 expressing TetO_sgRNA2	Twist Biosciences
AGG1171	Dm tRNA control	DmU6-1 TetO_sgRNA	Twist Biosciences
AGG1173	*DmU6-3	DmU6-3 expressing TetO_sgRNA2	Twist Biosciences
AGG1187	pIE1-Lep-FF-SV40	Lepidopteran consensus ATG context sequence, SV40 3' UTR	JP
AGG1202	pUC57-pB-TRE-FF-SV40	Identical to AGG1078, with constitutive promoter removed	JP
AGG1210	*PxU6-3	PxU6-3 expressing TetO_sgRNA2	Twist Biosciences
AGG1252	AbU6-744	AbU6-744 expressing TetO_sgRNA2	Twist Biosciences
AGG1256	*AgU6-557	AgU6-557 expressing TetO_sgRNA2	Twist Biosciences
AGG1261	Ag7SK	Ag7SK expressing TetO_sgRNA2	Twist Biosciences
AGG1276	AalbU6-132	AalbU6-132 expressing TetO_sgRNA2	Twist Biosciences
AGG1277	AaraU6-171	AaraU6-171 expressing TetO_sgRNA2	Twist Biosciences
AGG1278	AaraU6-449	AaraU6-449 expressing TetO_sgRNA2	Twist Biosciences
AGG1279	AfunU6-538	AfunU6-538 expressing TetO_sgRNA2	Twist Biosciences
AGG1280	AfunU6-704	AfunU6-704 expressing TetO_sgRNA2	Twist Biosciences
AGG1281	AsteiU6-842	AsteiU6-842 expressing TetO_sgRNA2	Twist Biosciences
AGG1282	AsteiU6-858	AsteiU6-858 expressing TetO_sgRNA2	Twist Biosciences
AGG1283	AsteiU6-917	AsteiU6-917 expressing TetO_sgRNA2	Twist Biosciences
AGG1300	Dm tRNA control	DmU6-1 Kmo_sgRNA_tRNA_TetO_sgRNA OsGlycine	Twist Biosciences
AGG1301	Dm tRNA control	DmU6-1 TetO_sgRNA_tRNA_Kmo_sgRNA OsGlycine (truncated tRNA sequence, error)	Twist Biosciences

Plasmid ID	Plasmid Name	Description	Source
AGG1302	Dm tRNA control	tRNA (OsGly) Kmo_sgRNA	Twist Biosciences
AGG1303	Dm tRNA control	tRNA (OsGly) TetO_sgRNA	Twist Biosciences
AGG1304	Dm tRNA control	DmU6-1 Kmo_sgRNA_tRNA_TetO_sgRNA AeGlycine	Twist Biosciences
AGG1305	Dm tRNA control	DmU6-1 TetO_sgRNA_tRNA_Kmo_sgRNA AeGlycine	Twist Biosciences
AGG1306	Dm tRNA control	DmU6-1 Kmo_sgRNA_tRNA_TetO_sgRNA Scrambled tRNA seq	Twist Biosciences
AGG1307	Dm tRNA control	DmU6-1 TetO_sgRNA_tRNA_Kmo_sgRNA Scrambled tRNA seq	Twist Biosciences
AGG1308	Dm tRNA control	tRNA (AeGly) Kmo_sgRNA	Twist Biosciences
AGG1309	Dm tRNA control	tRNA (AeGly) TetO_sgRNA	Twist Biosciences
AGG1311	Dm tRNA control	DmU6-1 Kmo_sgRNA	Twist Biosciences

Chapter 5: Transgene improvements for expressing multiple sgRNAs

Appendix Table 6: Plasmids used in Chapter 5

Plasmid ID	Plasmid Name	Description	Source
AGG1089	pB-Pub-hCas9_T2A_GFP-Opie2-DsRED	*Ae. aegypti* poly-ubiquitin driving human codon optimised Cas9 with a GFP fusion and DsRed transformation marker	Kind gift of Omar Akbari
AGG1113	AeU6-763 (400bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience
AGG1114	AeU6-763 (200bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience
AGG1115	CqU6-801 (200bp)	*C. quinquefasciatus* U6 promoter truncation	Twist Bioscience
AGG1117	AeU6-774 (400bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience

Plasmid ID	Plasmid Name	Description	Source
AGG1118	AeU6-774 (200bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience
AGG1121	AeU6-702 (400bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience
AGG1122	AeU6-702 (200bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience
AGG1123	CqU6-801 (400bp)	*C. quinquefasciatus* U6 promoter truncation	Twist Bioscience
AGG1124	CqU6-596 (200bp)	*C. quinquefasciatus* U6 promoter truncation	Twist Bioscience
AGG1125	CqU6-596 (400bp)	*C. quinquefasciatus* U6 promoter truncation	Twist Bioscience
AGG1132	AeU6-763 (100bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience
AGG1133	AeU6-774 (100bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience
AGG1134	AeU6-702 (100bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience
AGG1159	CqU6-801 (100bp)	*C. quinquefasciatus* U6 promoter truncation	Twist Bioscience
AGG1160	CqU6-728 (100bp)	*C. quinquefasciatus* U6 promoter truncation	Twist Bioscience
AGG1161	CqU6-728 (400bp)	*C. quinquefasciatus* U6 promoter truncation	Twist Bioscience
AGG1162	CqU6-728 (200bp)	*C. quinquefasciatus* U6 promoter truncation	Twist Bioscience
AGG1163	CqU6-596 (100bp)	*C. quinquefasciatus* U6 promoter truncation	Twist Bioscience
AGG1168	AeU6-905 (400bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience
AGG1169	AeU6-905 (200bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience
AGG1170	AeU6-905 (100bp)	*Ae. aegypti* U6 promoter truncation	Twist Bioscience
AGG1217	pGL3-Basic-Pub_Nuclease-activated-Luc_G1	Endonuclease (FF) reporter plasmid	SV

Primers

Chapter 3: Modulation of transgene expression through translational modification

Appendix Table 7: Primers used in Chapter 3

Primer name	Length (nt)	Primer sequence
LA227	22	GCCTTATGCAGTTGCTCTCCAG
LA228	19	GCAAGTTGACACTGGCGGC
LA233	23	ACTGGGGTAACCTTTGAGTTCTC
LA236	13	AGAAGCCGCCACC
LA237	14	AGAAAACCAACAAC
LA238	14	AGAAAAAAATCAAA
LA239	14	AGAACCGCCGGCGT
LA240	25	AGAAAACTTAAAAAAAAAAATCAAA
LA246	58	AGCCCCATGGAGACGTCATGCATCGTCTCGTTCTAAAGGTGTTA TAAATCAAATTAGT
LA293	13	CCATGGTGGCGGC
LA294	14	CCATGTTGTTGGTT
LA295	14	CCATTTTGATTTTT
LA296	14	CCATACGCCGGCGG
LA297	25	CCATTTTGATTTTTTTTTTTAAGTT
LA392	41	CCGTTCTAGAGCGGCCGCATGAATCGTTTTTAAAATAACAA
LA393	29	GTTGTCGACGTTAACTCGAATCGCTATCC
LA444	27	CCACCGGATCTAGATAACTGGAGCTTG
LA445	37	CAGTCGACCCCAAACGCGCCAGTGGTAGTACACAGTA

Chapter 4: An *in vitro* CRISPRa assay for validating sgRNA activity

Appendix Table 8: Primers used in Chapter 4

Primer name	Length (nt)	Primer sequence
LA1205	39	TATCTGGGCCCATTACACGGCGATCTTTCCGCCCTTCTT
LA1206	31	TCATACCTAGGAACCAACAACATGGAAGACG
LA125	22	CCTCTACAAATGTGGTATGGCT
LA291	23	GGTGTTATAAATCAAATTAGTTT
LA37	24	CCCTAGAAAGATAATCATATTGTG
LA456	22	GGTACCCTGGCGTCTAATTGGG
LA985	60	GAAATTAATACGACTCACTATAGGTCTCTATCACTGATAGGGAGGTTT TAGAGCTAGAAA

Primer name	Length (nt)	Primer sequence
LA986	60	GAAATTAATACGACTCACTATAGGACTTTTCTCTATCACTGATAGTTTTAGAGCTAGAAA
LA987	60	GAAATTAATACGACTCACTATAGGCACTTTTCTCTATCACTGATGTTTTAGAGCTAGAAA
LA988	80	AAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC

Chapter 5: Transgene improvements for expressing multiple sgRNAs

Appendix Table 9: Primers used in Chapter 5

Primer name	Length (nt)	Primer sequence
LA1308	84	AAAAGCACCGAATCGGTGCCGACGTTCCCACGTCTGATAACGGACTGGCCTTATTGCAACTTGACACTCCCGTGTCTCTGCAAC
LA1416	86	AAAAGCACCGACTCGGTGCCAGCTCTCCCGAGCTTGATAACGGACTTGCCTTATCGCAACTTGCATCTTTTCAGATGCTCTGCGAC
LA1417	84	AAAAGCACCGACTCGGTGCCACGCTTTTCAGCGTTGAATACGGACTAGCCTTATCCTAACTTGCCATTTTCATGGCTCTAGGAC
LA1418	63	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTGCAGAGACACGGG
LA1419	63	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTCGCAGAGCATCTGA
LA1420	63	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTCCTAGAGCCATGAA
LA1543	86	AAAAGCACCGACTCGGTGCCCTGCATTCCTGCAGTGATAACGGACTAGCCTTATGTTAACTTGGGATATCTCTATCCCTCTAACAC
LA1544	86	AAAAGCACCGAATCGGTGCCTGTCGTTTCCGACATGAATACGGACTAGCCTTATGCTAACTTGTACGTTTCCACGTACTCTAGCAC
LA1545	80	AAAAGCACCGACTCGGTGCCCAGGTCTCCCTGTGATAACGGACTGGCCTTATTCGAACTTGGACTCTCGTCCTCTCGAAC
LA1546	82	AAAAGCACCGACTCGGTGCCAGCGTTTCCGCTTGATAACGGACTCGCCTTATTGTAACTTGCGTATTTCTACGCTCTACAAC
LA1547	86	AAAAGCACCGACTCGGTGCCACAGCTCCCGCTGTTGATAACGGACTCGCCTTATGCGAACTTGCTTACTTTCGTAAGCTCTCGCAC
LA1548	86	AAAAGCACCGAATCGGTGCCGGTCATCTCTGACCTGATAACGGACTGGCCTTATGCCAACTTGAGTAGTCCCCTACTCTCTGGCAC
LA1549	64	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTGCTAGAGTACGTGGA
LA1550	64	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTGTTAGAGGGATAGAG
LA1551	64	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTCGAGAGGACGAGAG

Primer name	Length (nt)	Primer sequence
LA1552	64	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTGTAGAGCGTAGAAA
LA1553	64	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTGCGAGAGCTTACGAA
LA1554	64	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTGCCAGAGAGTAGGGG
LA2162	60	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTCCAGAGTCGG
LA2163	94	AAAAGCACCGAATCGGTGCCTGCCTTCCGGCATGATAACGGACTGGTATATAATACACTGCCTTATTCCAACTTGTCGTTCCCGACTCTGGAAC
LA2164	60	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTGGAGAGGCAT
LA2165	84	AAAAGCACCGACTCGGTGCCCCTAGTCTCCTAGGTGTGTACGGACTAGCCTTATTGGAACTTGGCATTCTCATGCCTCTCCAAC
LA2166	60	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTCTTAGAGTGTG
LA2167	84	AAAAGCACCGAATCGGTGCCTCAGGTTCCCCTGATGATAACGGACTAGCCTTATCTTAACTTGTGTGTTCCCACACTCTAAGAC
LA2168	60	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTCGGAGAGAACA
LA2169	86	AAAAGCACCGAATCGGTGCCGTCGTTCGCACGACTGTGTACGGACTAGCCTTATCGGAACTTGAACAGTCCCCTGTTCTCTCCGAC
LA2170	60	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTGGGAGAGCCAA
LA2171	86	AAAAGCACCGACTCGGTGCCAGGTCTCCCGACCTTGTGTACGGACTAGCCTTATGGGAACTTGCCAAATTTCTTTGGCTCTCCCAC
LA2172	60	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTCCAGAGTCGG
LA2173	80	AAAAGCACCGAATCGGTGCCTGCCTTCCGGCATGATAACGGACTTGCCTTATTCCAACTTGTCGTTCCCGACTCTGGAAC
LA2174	60	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTCCAGAGGTTCG
LA2175	83	AAAAGCACCGACTCGGTGCCCGAACTCTCGTTCGTGATAACGGACTCGCCTTATCCCAACTTGGTTCTCTCGAACCTCTGGAC
LA2176	60	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTGCAGAGACAC
LA2177	98	AAAAGCACCGAATCGGTGCCGACGTTCCCACGTCTGATAACGGACTGGTTTAATAAACACTGCCTTATTGCAACTTGACACTCCCGTGTCTCTGCAAC
LA2178	60	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTGTAGAGCGTA
LA2179	96	AAAAGCACCTACTCGGTGCCAGCGTTTCCGCTTGATAACGGACTGGAATTTAATTCACTGCCTTATTGTAACTTGCGTATTTCTACGCTCTACAAC

Primer name	Length (nt)	Primer sequence
LA2180	60	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTCGAGAGGACG
LA2181	92	AAAAGCACCGACTCGGTGCCAGGTCTCCCTGTGATAACGGACTGGATTAAAATCACTGCCTTATTCGAACTTGGACTCTCGTCCTCTCGAAC
LA925	62	GAAATTAATACGACTCACTATAGGGCCATATAATGTGGGCGGCAGTTCCAGAGTCGTGCTGG
LA986	60	GAAATTAATACGACTCACTATAGGACTTTTCTCTATCACTGATAGTTTTAGAGCTAGAAA
LA988	80	AAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC

Materials

Appendix Table 10: Materials and reagents used throughout

Name	Manufacturer	Catalogue #	Description
Cell culture flasks with angled neck and plug seal cap	Corning, Fisher Scientific, UK	(e.g.) 10767442	Cell culture flask
Cell scrapers	Corning, Fisher Scientific, UK	10707441	Cell scrapers
CoolCell	Corning, Fisher Scientific, UK	15542771	Cryopreserving aid
DMSO	Sigma-Aldrich (Merck, Germany)	D2650	Cryopreservant
Dual luciferase assay kit	Promega, UK	E1960	Dual luciferase assay reagents
Fetal bovine serum (serum)	Labtech, Lewes, UK	SKU: FB-1001/100-500	Media supplement
Insect Xpress medium	Lonza, Switzerland	BELN12-730Q	Media
Leibovitz's L-15 medium	Gibco, Fisher Scientific, UK	11415056	Media
Lipofectamine 2000	Invitrogen, Fisher Scientific, UK	12566014	Transfection reagent
Nunc MicroWell 96-well plates	ThermoFisher Scientific, UK	10212811	Cell culture plate
Opti-MEM	Gibco, Fisher Scientific, UK	10149832	Media
PBS^{-}	Central Services, The Pirbright Institute	N/A	Ion-free phosphate buffered saline (PBS)
Penicillin/streptomycin solution (Pen/Strep)	Gibco, Fisher Scientific, UK	11528876	Media supplement
Schneider's Drosophila medium	Gibco, Fisher Scientific, UK	11590576	Media
Sterile reservoirs	Starlab, UK	E2310-1010	Reagent reservoir
Thermo Scientific™ White 96-Well Immuno Plates	ThermoFisher Scientific, UK	10537205	Optical plate
TransIT-Pro Kit (inc. Boost)	Mirrus Bio, GeneFlow	E7-0152	Transfection reagent
Trypan blue	Gibco, Fisher Scientific, UK	15250061	Cell stain
Tryptose phosphate broth (TPB)	Gibco, Fisher Scientific, UK	18050039	Media supplement

Name	Manufacturer	Catalogue #	Description
Virkon tablets	RelyOn+, Scientific Laboratory Supplies, UK	330013	Disinfectant
50x TAE	Fisher Scientific, UK	10399519	Buffer
Agarose	Sigma-Aldrich (Merck, Germany)	A9539-500G	Agarose for gel electrophoresis of nucleic acids
Ampicillin sodium salt	Fisher Scientific, UK	10419313	Antibiotic
ApaI	NEB, UK	R0114S	Restriction enzyme
AvrII	NEB, UK	R0174S	Restriction enzyme
BamHI	NEB, UK	R0136S	Restriction enzyme
BsmBI	NEB, UK	R0580L	Restriction enzyme
Buffer 3.1	NEB, UK	B7203S	Buffer
CloneJET PCR cloning kit	Thermo Scientific (Fisher Scientific, UK)	10765841	PCR cloning kit
CutSmart	NEB, UK	B7204S	Buffer
DEPC treated water	Ambion (FisherScientific, UK)	AM9906	RNA-safe water
DNA oligos	Sigma-Aldrich (Merck, Germany)	Custom	Synthetic oligonucleotides
dNTPs	NEB, UK	N0447L	Deoxynucleotide (dNTP) solution mix
DreamTaq buffer	Thermo-fisher scientific, UK	B65	Buffer
DreamTaq enzyme	Thermo-fisher scientific, UK	EP0701	DreamTaq DNA polymerase
Gel loading dye	NEB, UK	B7025S	Gel loading dye, purple (6x)
GeneRuler 50bp DNA ladder	ThermoFisher Scientific, UK	10794291	DNA ladder
Glycerol	Analar, VWR, UK	2438826	Reagent
HindIII	NEB, UK	R0104S	Restriction enzyme
HindIII-HF	NEB, UK	R3104S	Restriction enzyme, high fidelity
Hyperladder, 1kb DNA ladder	Bioline Reagents	BIO-33025	DNA ladder
Kanamycin sulphate	Fisher Scientific, UK	11815024	Antibiotic
Klenow	NEB, UK	M0210	Reagent
MEGAscript T7 kit (short)	Ambion (FisherScientific, UK)	AM1354	RNA transcription
NcoI-HF	NEB, UK	R3193S	Restriction enzyme, high fidelity

Name	Manufacturer	Catalogue #	Description
NsiI-HF	NEB, UK	R3127S	Restriction enzyme, high fidelity
NucleoSpin Gel and PCR Clean-up	Macherey Nagel, Germany	11992242	dsDNA purification kit
NucleoSpin Plasmid	Macherey Nagel, Germany	12353358	Plasmid prep kit
PacI	NEB, UK	R0547S	Restriction enzyme
Q5 buffer	NEB, UK	B9027S	Buffer
Q5 enzyme	NEB, UK	M0491L	Q5 High-Fidelity DNA Polymerase
rSAP	NEB, UK	M0371S	Recombinant shrimp alkaline phosphotase
SacII	NEB, UK	R0157S	Restriction enzyme
SalI-HF	NEB, UK	R3138S	Restriction enzyme, high fidelity
Sanger sequencing	Source Bioscience, UK	Custom	Sanger sequencing
SOC media	MP Biomedicals (Fisher Scientific, UK)	3031-012	Media
T4 DNA ligase	NEB, UK	M0202L	T4 DNA ligase
T4 ligation buffer	NEB, UK	B0202L	Buffer
Water, molecular grade	Millipore, Merk, Germany	H20MB0506	Water
XbaI	NEB, UK	R0145S	Restriction enzyme
XhoI	NEB, UK	R0146S	Restriction enzyme
XL 10-Gold ultracompetent cells	Stratagene (Agilent, UK)	#200315	Competent cells

Appendix C: Supplemental Information - Chapter 3

Optimisation experiments carried out in advance of the work shown in Chapter 3 are detailed here.

Optimisation experiment 1 looks at transfection of fifteen different firefly luciferase plasmids (Table 9) in two mosquito cell lines (C6.36 and U4.4). Experimental controls are shown in Figure 32 and results in Figure 33. This experiment was conducted to determine whether the experimental design could produce reliable results.

Single luciferase controls

The control conditions in Figure 32 were created as indicators of experiment reliability, which cannot be discerned from the experimental results as they are positive for both firefly luciferase (FF) and Renilla luciferase (RL) activity. Single luciferase controls allow us to screen

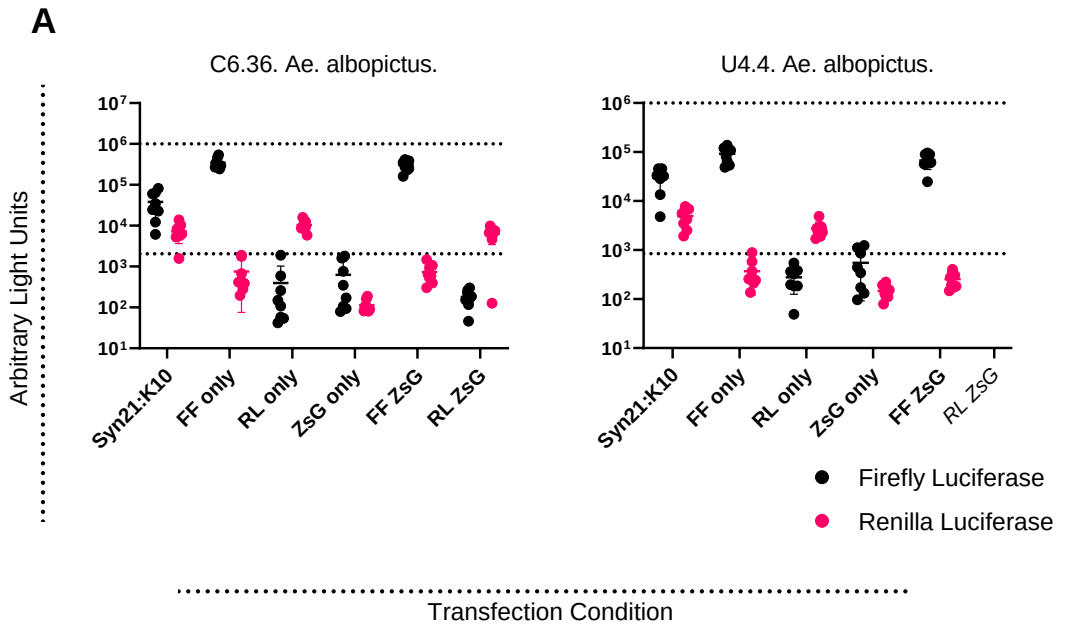


Figure 32: Graphs of controls for optimisation experiment 1. Control transfection conditions are created in each cell line alongside the main experiment (Figure 33). Controls are analysed before transforming the main results from independent luciferase activities (arbitrary light units (ALU)) to a ratio (FF/RL) for analysis (Figure 33). Control transfection conditions are designed to show FF expression and RL expression independently, but otherwise mimicking experimental conditions (e.g. transfection of FF plasmid without RL plasmid). A non-luciferase plasmid, ZsG (Control plasmid: luciferase-null (HR5-IE1-ZsGreen)) is included as a negative control. Experimental condition Syn21:K10 is included as a positive control. Data is shown with "transfection condition" on the x-axis and ALU on the y-axis. FF and RL activities are represented with different colours, FF quenching threshold is marked at 10^6 ALU and RL background threshold is marked for each cell line (99.9% CI of RL activity in "FF only"). Transfection condition "RL ZsG" was not completed in U4.4 due to a calculation error in reagent preparation. These graphs give an overview of data quality that cannot be seen once results are transformed to FF/RL.

for under or over-expression of each luciferase and provide the data to calculate the RL background threshold for each experiment (indicated as the lower horizontal line).

In initial experiments (data not shown) it was difficult to ensure that samples expressing Renilla luciferase did so at an amount distinguishable from background (light readings generated by the optical plate and cell lysate, independent of any luciferase activity). Arising from these difficulties, it was decided to calculate a 'background threshold' for RL

measurements based on the arbitrary light units (ALU) generated in the RL measurement of a sample not transfected with the RL plasmid.

The RL background threshold for each experiment is set at the upper 99.9% confidence interval (CI) for mean RL measurements of the transfection condition "FF only". This is done for each cell line. The 99.9% CI was selected as setting RL background threshold to the 95% and 99% CI resulted in a false positive rate of 1 in 4 and 1 in 8, respectively. Using the 99.9% CI as the RL background threshold results in a 1 in 16 error rate (across seven experiments).

There was a second major concern for error generated by the experiment protocol. It was identified (in preliminary experiments) that firefly luciferase (FF) activity could overwhelm the quenching capacity of the dual luciferase assay reagents, resulting in on-going FF activity (and light emissions) during the RL activity measurements[4]. This is an artifact of using dual luciferase assay reagents at a 1 in 10 dilution (to enable a higher sample processing capacity). A 'quenching threshold' was set at 10^6 ALU. A sample with FF activity above this threshold is considered to be at meaningful risk of having a RL measurement contaminated with FF activity. The quenching threshold is marked with a horizontal line on each graph in Figure 32.

Considering difficulties encountered in establishing a reliable experimental protocol, a conservative (cautious) approach was taken with regards to experimental controls. Optimisation experiment 1 represents the first occurrence of the experimental protocol that generated data sets unmarred by widespread error. Figure 32 shows the FF and RL results (in ALU) for a panel of control conditions in each cell line (U4.4 and C6.36, both *Ae. albopictus*).

From the left of each graph, a double-positive transfection condition (transfected with FF and RL) is included as a final check that the results seen in single-positive controls are not affected by presence of the other luciferase. For both cell lines, FF activity of the double-positive (Syn21:K10) is below the quenching threshold and RL activity is above the background threshold, apart from in one repeat (of 8) in cell line C6.36. This error rate in RL activity is within acceptable parameters. In the full results, any sample with RL activity below the background threshold is excluded from further analysis, as is any sample with FF activity above the quenching threshold.

[4] Each firefly luciferase and Renilla luciferase emit the same wavelength of light. Their activity can, therefore, only be distinguished by knowing the reagents present at the time a light measurement is made (PROMEGA 2015. Dual-Luciferase Reporter Assay System. *Instructions for use of Products E1910 and E1960.*).

In both cell lines, the single-positive controls each have all FF and all RL results below and above their respective thresholds.

To control for any change in luciferase activity mediated by transfection with multiple plasmids, a luciferase-null plasmid (pHR5-IE1-ZsGreen) was used to generate single-positive luciferase controls “FF ZsG” and “RL ZsG”. The transfection condition “ZsG only” is null for both luciferases. These results were analysed to determine if the presence of a second (luciferase-null) plasmid generated a significant difference in luciferase activity as compared to the single-positive controls “FF only” and “RL only” (Table 24). Of note, the “RL ZsG” results for cell line U4.4 were not completed, due to user error during transfection.

Table 24: Mann-Whitney test results summary

	Comparisons	P value	P value summary
C6.36	"FF only" FF vs "FF ZsG" FF	0.645	ns
	"FF only" RL vs "FF ZsG" RL	0.574	ns
	"RL only" FF vs "RL ZsG" FF	0.645	ns
	"RL only" RL vs "RL ZsG" RL	0.021	*
U4.4	"FF only" FF vs "FF ZsG" FF	0.235	ns
	"FF only" RL vs "FF ZsG" RL	0.505	ns

In deference to non-normal distribution of several groups, all statistics were completed using non-parametric tests. A Mann-Whitney test (two-tailed) was used to compare each pair of groups indicted in Table 24. N = 8 for each group and significance is indicated by “ns” (not significant) or “*” ($P < 0.05$).

For FF measurements there is no significant difference between the presence and absence of pHR5-IE1-ZsGreen in the transfection condition. For RL activity, there is a significant difference between the presence and absence of pHR5-IE1-ZsGreen for the RL-positive conditions (only cell line C6.36) but not for the RL-negative conditions. This pattern of results suggests that the presence of the pHR5-IE1-ZsGreen plasmid enhances RL expression rather than increasing expression of any plasmid (as a change in transfection efficiency caused by larger amounts of DNA would account for) or creating a false RL signal (which would be seen in RL-null conditions).

Although this single experiment is not sufficient to disprove the null hypothesis, one possible explanation could be that a trans-acting enhancement of RL expression (promoter OpIE2) is mediated by the HR5-IE1 promoter on the ZsG plasmid, for example by increased proximity

of additional transcription factors. Such an effect would not be replicated by the FF plasmids as they use the same HR5-IE1 promoter. Further exploration of this hypothesis is outside of the scope of this project, and the effect is controlled for in experiments by the absence of pHR5-IE1-ZsGreen and by use of a consistent mass of FF plasmid in each transfection condition.

As discussed, the control results shown in Figure 32 raise no concerns for the experimental results in optimisation experiment 1 (Figure 33).

Experimental results

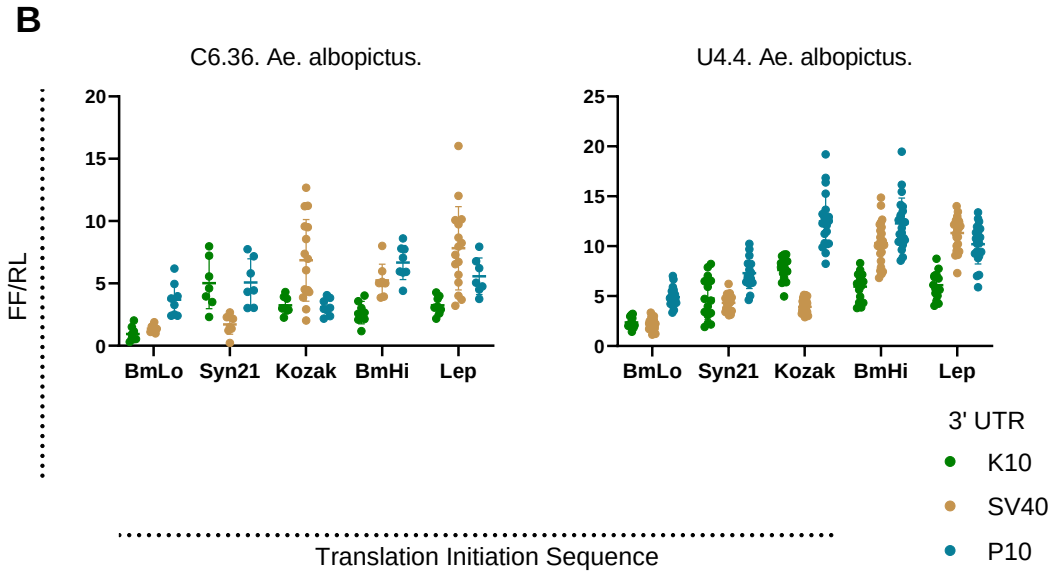


Figure 33: Graphs of experimental results for optimisation experiment 1. Results are shown for transfection of fifteen different FF expressing plasmids (Table 6) in two mosquito cell lines. Data is shown as FF/RL (y-axis) for each sample, which represents translational efficiency in this experiment. FF/RL values cannot be directly compared between cell lines (Analysis), which are therefore shown on independent graphs with independent scales. FF plasmids are grouped by translation initiation sequence (TIS) on the x-axis and by 3' UTR sequence in different colours. Each data point (circle) is an individual sample. Mean and SD are indicated. Replicates were generated using a transfection master-mix. C6.36 N = 7-16. U4.4 N = 15-23.

The luciferase activity for each transfection condition was screened for FF measurements above quenching threshold or RL measurements below background threshold and any such samples were excluded from further analysis. FF activity and RL activity for each sample were then transformed to FF/RL, using the internal control (RL) to standardise for transfection efficiency (and sample handling) between samples. Of note, FF/RL values can only be compared within a cell line, not between cell lines (due to differential activity of promoters HR5-IE1 (FF plasmid) and OpIE2 (RL plasmid)).

Figure 33 shows the results for transfecting fifteen different plasmids (every combination of five TIS and three 3'UTR, Table 9) in two cell lines, C6.36 and U4.4. The data are separated by translation initiation sequence (TIS) on the x-axis and by 3'UTR in different colours. The ratio FF/RL is shown on the y-axis, which is at a different scale in each cell line. Each repeat (N = 7-23) is shown as a solid colour dot with mean and standard deviation (SD) indicated as lines. Although relative activity between different constructs can be compared between cell lines, FF/RL values cannot be compared across cell lines or experiments.

Analysing these results is statistically complicated. Data sets for each construct (N ≤ 23) are too small to support the number of comparisons that would be needed (e.g. interrogating whether one construct is significantly different from each other construct); multiple interrogations of one data set must account for every other interrogation of the same data set (via the degrees of freedom, which depend on N). The solution is to analyse the data using a generalised linear model, which can account for the multiple variables despite the low N. This analysis is itself complicated by the distribution of the data, where variance increases with the mean. Such analysis was referred to a specialist (Phil Leftwich), who kindly analysed the data for "experiment 1", which includes all five cell lines covered in this project. This analysis was not completed for optimisation experiment 1 or 2.

A simple analysis was carried out to describe the difference in FF/RL of the construct with the lowest activity and that with the highest, in each cell line. Table 25 shows summary results for a Mann-Whitney test. There is a significant difference between the highest and lowest expressing constructs in each cell line, represented by an 8-fold change in cell line C6.36 and a 6-fold change in cell line U4.4. Conclusions cannot be drawn from the identity of the construct with the lowest and that with the highest mean FF/RL in each cell line as it is not known that these results are significantly different from other constructs with similar mean activity.

Table 25: Summary results and Mann-Whitney test results

		Mean (FF/RL)	SD	N	P value	P value summary	Approx. fold change
C6.36	**BmLo:K10**	0.941	0.584	8	<0.0001	****	8
	Lep: SV40	7.817	3.347	16			
U4.4	**BmLo:SV40**	2.182	0.573	23	<0.0001	****	6

		Mean (FF/RL)	SD	N	P value	P value summary	Approx. fold change
	Kozak:P10	12.487	2.613	22			

These results confirm that a change can be made to transgene expression activity, in two mosquito cell lines, by altering only the TIS and 3'UTR. This analysis is superficial, but the fold changes achieved are lower than those described in the literature in *D. melanogaster* ("translational enhancers ... can be used to increase protein yields by a factor of more than 20" Pfeiffer et al. (2012)) and in *B. mori* (10 fold change *in vitro* and 47 fold change *in vivo*) (Tatematsu et al., 2014).

Optimisation experiment 2 builds on the work of optimisation experiment 1 and demonstrates further development of the experimental controls (Figure 34). This experiment was carried out in cell line Aag2 (*Ae. aegypti*) and demonstrates that the experimental design can be successful in cell lines representing multiple mosquito species. Optimisation experiment 2 was conducted as a small-scale experiment for final confirmation before proceeding with the full scale (five cell lines) experiment.

Lysate volume controls

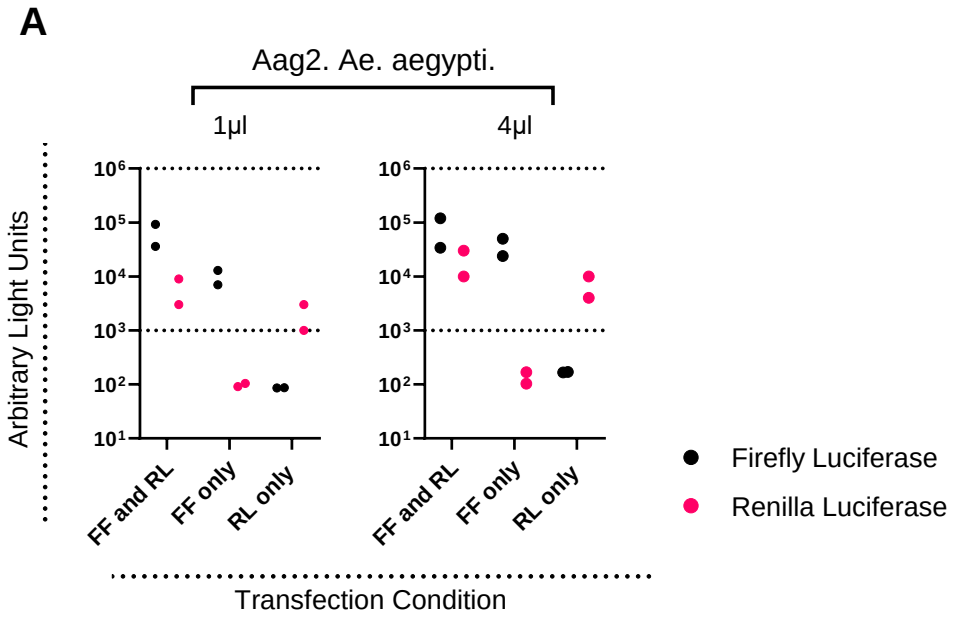


Figure 34: Graphs of lysate controls for optimisation experiment 2. Samples of control transfections were processed by dual luciferase assay at two volumes, 1µl and 4µl, each represented on independent graphs, with arbitrary light units (ALU) on the y-axis and transfection condition (plasmid combination) on the x-axis. Firefly and Renilla luciferase values for each sample are shown interleaved (in different colours). The FF quenching threshold (10^6 ALU) and estimated RL background threshold (10^3 ALU) are indicated. Two samples were processed for each transfection condition. The FF plasmid in transfection condition "FF and RL" is not the same FF construct as in "FF only". These results are used to select a suitable lysate volume to process experimental results.

As discussed around optimisation experiment 1, the dynamic range of this dual luciferase assay set up (reagents at 1 in 10 dilution) is relatively small – from approximately 10^3 ALU to 10^6 ALU. The dual luciferase assay protocol already called for only a portion of the cell lysate sample to be used in the dual luciferase assay reaction; it was decided to screen different volumes of cell lysate (sample) for their luciferase activity before committing to processing an entire cell line of samples. This permits additional intervention to keep samples within the

dynamic range of the assay, without compromising the experiment as FF/RL results can only be compared within a cell line anyway. It accounts for variations in transfection efficiency or culture health (number of cells) between cell lines and experiments.

The volume of passive lysis buffer (Promega, UK) used to harvest transfected cell samples was standardised to 21µl. This permits three uses of the sample – up to 7µl each for screening lysate amounts, processing results and repeat processing of results if needed. This standard (21µl) was used for every transfection for any experiment (in this chapter or future chapters).

Within an experiment, different cell lysate volumes (usually two) were screened for each cell line; these results are shown in Figure 34 and are used to select a lysate volume for the rest of the samples in that cell line. Each single luciferase control is represented, as is a double luciferase sample that is expected to have high expression of firefly luciferase (FF). The same two repeats are used at each lysate volume, but the FF plasmid in the double positive sample ("FF and RL") is not the same as that in the single positive ("FF only") sample.

In Figure 34 the FF and RL measurements of each sample are indicated as different colours with the transfection condition (combination of plasmids) represented on the x-axis and arbitrary light units (ALU) on the y-axis. The two lysate volumes, 1µl and 4µl are represented on independent graphs with the cell line indicated above (Aag2, in this case). Although the specific FF quenching threshold (10^6 ALU) is indicated, an estimated RL background threshold must be used as there is not enough data to calculate the 99.9% confidence interval it is based upon.

In optimisation experiment 2 (Figure 34), both lysate volumes show the desired pattern of results: each luciferase activity present above background threshold only where the corresponding luciferase plasmid was present in the transfection and FF activity present below the FF quenching threshold (10^6 ALU). It was decided to proceed with 4µl lysate volume for the rest of the experiment as RL measurements are higher without corresponding FF measurements becoming close to the quenching threshold (there may be experimental samples with greater FF activity than in the samples screened here).

Single-luciferase controls

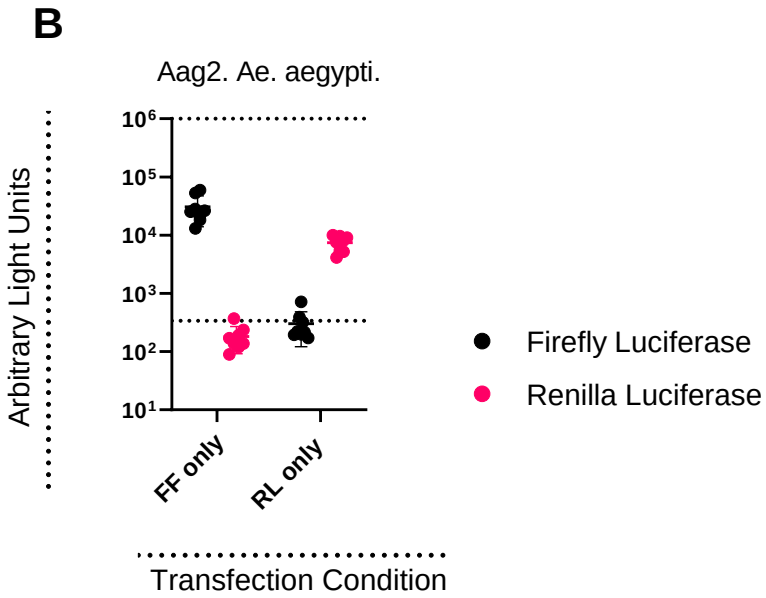


Figure 35: Graphs of controls for optimisation experiment 2. Single luciferase controls are created alongside the main experiment, as in optimisation experiment 1 (Figure 32). These controls facilitate calculation of the Renilla luciferase (RL) background threshold (99.9% CI of RL measurements for transfection condition "FF only") and show whether single luciferase expression is conforming to expectations (RL activity above background threshold, FF activity below quenching threshold) in the event of problems with the experimental data set. Arbitrary light units (ALU) are shown on the y-axis and each luciferase (FF and RL) is distinguished by colour. Transfection conditions are named on the x-axis and the calculated RL background threshold and FF quenching threshold are each indicated. There is no ZsG transfection condition. N = 8.

The single luciferase controls shown in Figure 35 have a shifted purpose from those in optimisation experiment 1 (Figure 32). With the added screening step used to select an optimised lysate volume for processing samples by dual luciferase assay (Figure 34), the predominant purpose of the single lysate controls is in calculating the background threshold for Renilla luciferase measurements. In the event of difficulties with the main experimental results, however, the single lysate control results offer additional information for troubleshooting.

In comparison with optimisation experiment 1, the non-luciferase plasmid control ("ZsG") was dropped as it does not yield information different from that in the single luciferase ("FF only" and "RL only") conditions. A double positive luciferase condition was not included as the data in Figure 35 was generated concurrently with the data in Figure 36 (experimental

results) and the results for each sample in Figure 36 were screened individually for luciferase activity outside of the dynamic range of the assay.

As in optimisation experiment 1, the actual RL background threshold is calculated using the 99.9% confidence interval of the mean of RL activity in the RL-null ("FF only") transfection condition (Appendix Table 11). Looking at the data in Figure 35, the RL measurements are above the RL background threshold and FF measurements are below the quenching threshold, so there are no concerns raised by these controls.

Experimental results

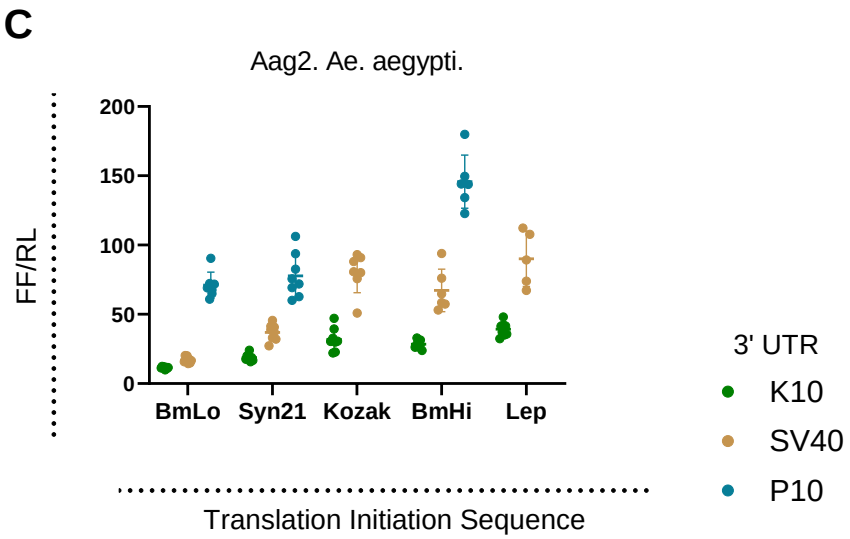


Figure 36: Graphs of experimental results for optimisation experiment 2. Results are shown for transfection of fifteen different FF expressing plasmids in mosquito-origin cell line, Aag2. Data points are shown for each sample (N = 5 - 8) with FF/RL on the y-axis (a measure of transgene expression), translation initiation sequence (TIS) on the x-axis and 3'UTR sequence as different colours. Mean and SD are represented for each combination of TIS and 3'UTR. Raw luciferase (FF and RL) values are screened for under or over-expression before transforming to FF/RL. Two combinations of TIS:3'UTR had over-expression of FF in each repeat; these data sets are absent from the graph and analysis (Kozak:P10, Lep:P10).

After the dual luciferase assay, results for each sample were screened for (and excluded based on) Renilla luciferase (RL) measurements below background threshold (calculated per experiment) or firefly luciferase (FF) measurements above the FF quenching threshold (10^6 ALU). Values were then transformed to FF/RL for each sample to standardise for variation in transfection efficiency between samples. FF/RL values (as presented in Figure 36) represent the efficiency of transgene expression when compared between samples. Actual values (FF/RL) cannot be compared between experiments, only changes in the relationship between two values.

In this experiment, many individual samples had FF expression greater than the quenching threshold. Once these samples were excluded from further analysis, two TIS:3UTR combinations (FF expressing plasmids) were excluded altogether - Kozak:P10 and Lep:P10. As the data set is incomplete, full analysis (generalised linear model) was not carried out.

Simple analysis (similar to that in optimisation experiment 1) was carried out to determine if the constructs with the highest and lowest mean FF/RL were statistically significantly different from each other. In deference to the low degrees of freedom available, only this comparison was done. Summary results for a Mann-Whitney test are shown in Table 26, describing the change in activity from BmLo:K10 to BmHi:P10 as statistically significant. The fold change is 13, which may be an underestimate of what can be achieved in cell line Aag2, as the two high-expression conditions (Kozak:P10 and Lep:P10) could not be analysed. This is a greater fold change than previously obtained in C6.36 or U4.4 (8 and 6 fold), though still in the same order of magnitude. As for optimisation experiment 1, no conclusions can be drawn based on the specific identity of the highest and lowest expressing TIS:3'UTR combinations; we cannot determine statistically that each is different from other high- or low-expressing combinations.

Table 26: Summary results and Mann-Whitney test results

		Mean (FF/RL)	SD	N	P value	P value summary	Approx. fold change
Aag2	**BmLo:K10**	11.48	0.73	8	0.0007	***	13
	BmHi:P10	145.75	19.22	6			

Optimisation experiment 2 corroborates the findings of optimisation experiment 1, showing that a significant fold change in expression can be achieved by altering only the TIS and 3'UTR of a transgene. This effect is now seen in three cell lines, representing two culicine mosquito species (C6/36 and U4.4 *Ae. albopictus*, Aag2 *Ae. aegypti*).

Changes to the experimental protocol have yielded an additional tool for managing the limited dynamic range of the reporter assay, whilst preserving the ability to handle a high number of samples. To elucidate further information, the scale of the experiment was increased to five cell lines (generating additional degrees of freedom without undue additional effort) and is described as "experiment 1".

Experiment 1

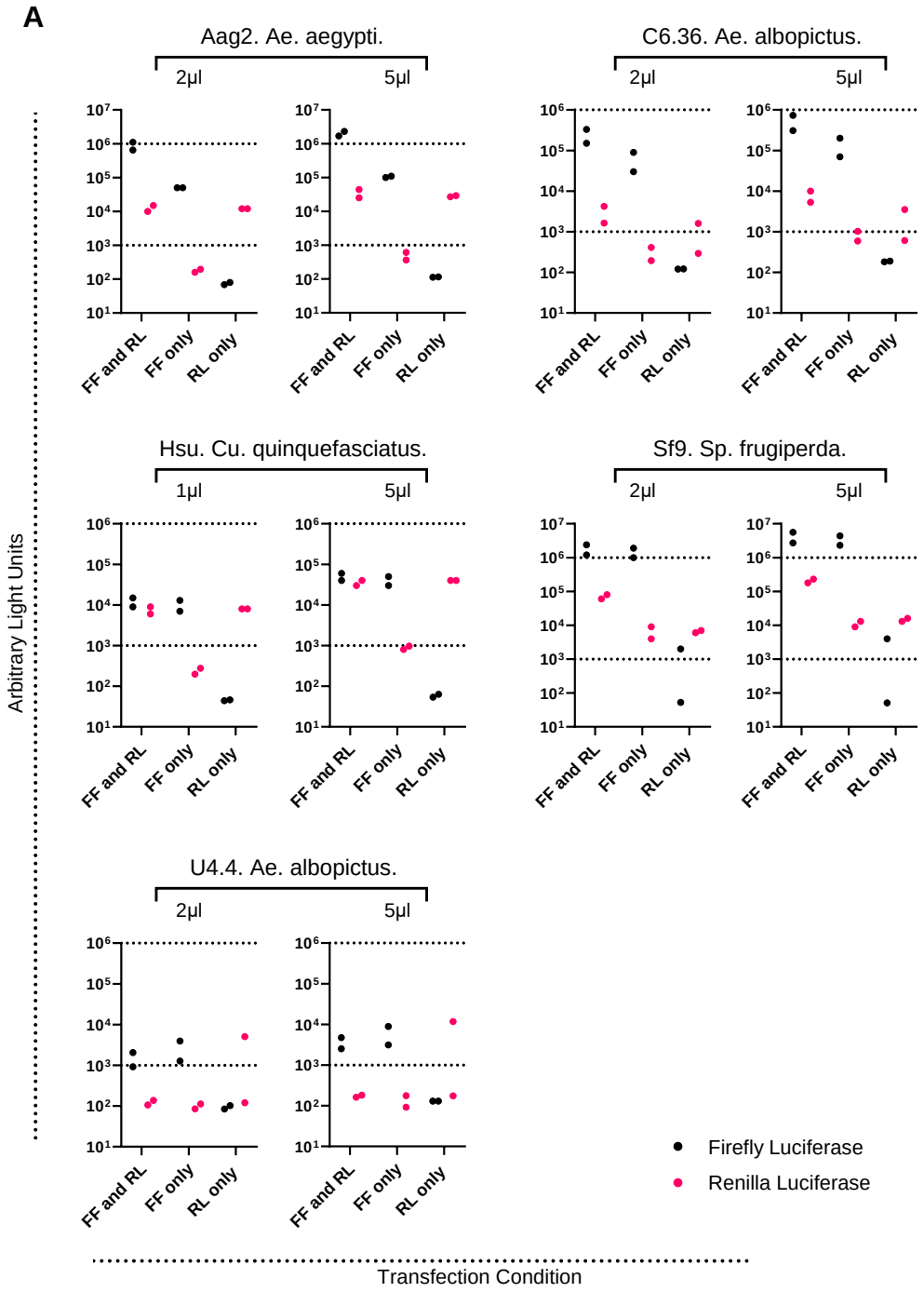


Figure 37: Graph of lysate controls for experiment 1. Samples of control transfections for each cell line were processed by dual luciferase assay twice, at different lysate volumes. This provides an empirical basis for deciding the lysate volume for the full set of samples for each cell line. Each lysate volume for each cell line is represented on an independent graph. Y-axis (arbitrary light units, ALU) scales are the same within a cell line but not between cell lines. The control condition (plasmids transfected) is labelled on the x-axis and different luciferases are denoted by colour. Estimated RL minimum threshold (10^3 ALU) is shown, as is FF quenching threshold (10^6 ALU). The FF plasmid in "FF and RL" is not the same plasmid used for "FF only". The same sample is used for both lysate volumes. N = 2 throughout.

For each cell line in experiment 1 (Aag2 *Ae. aegypti*, C6.36 *Ae. albopictus*, Hsu *C. quinquefasciatus*, Sf9 *S. frugiperda*, U4.4 *Ae. albopictus*), two lysate volumes were screened using control samples, before committing to a lysate volume with which to process the full set of samples. This step is aimed at improving the number of samples that are read within the dynamic range of the dual luciferase assay. The two main concerns are firefly luciferase (FF) measurements above 10^6 ALU and/or Renilla luciferase (RL) measurements below the background threshold. The RL background threshold is calculated from the RL-null control ("FF only") but cannot be found at N = 2, so an estimate RL background threshold is used at 10^3 ALU.

The lysate volume used to process a full set of samples can vary between cell lines as results cannot be directly compared between cell lines anyway (due to differential expression of the internal control plasmid). Each pair of graphs is analysed to determine a lysate volume that best moderates FF activity without losing RL activity below the estimated background threshold. A dual-positive control ("FF and RL") is included in this panel to represent the sample that is expected to have amongst the highest expression, based on preliminary experiments. It also shows if there is an increase in RL activity mediated by the FF plasmid (an effect that is constant between samples, within a cell line).

Table 11 details the lysate volumes used for each cell line and is repeated in summary as Table 27. In brief, the lysate volume for cell line Aag2 was reduced to deal with high FF activity. The same was done for cell lines C6.36 and Sf9. In cell line U4.4 the RL measurements of the RL-positive conditions overlapped with the RL measurements for the RL-null condition, with one exception. The highest lysate volume (7µl) was therefore selected. Cell line Hsu shows the desired pattern of FF and RL activity at both lysate volumes and it was decided to proceed with 1µl as the background RL measurement (that of "FF only") is lower than at 5µl.

Table 27: Lysate volumes for experiment 1

	Cell line				
	Aag2	C6.36	Hsu	Sf9	U4.4
Lysate volume (µl)	1	1	1	1	7

Single luciferase controls

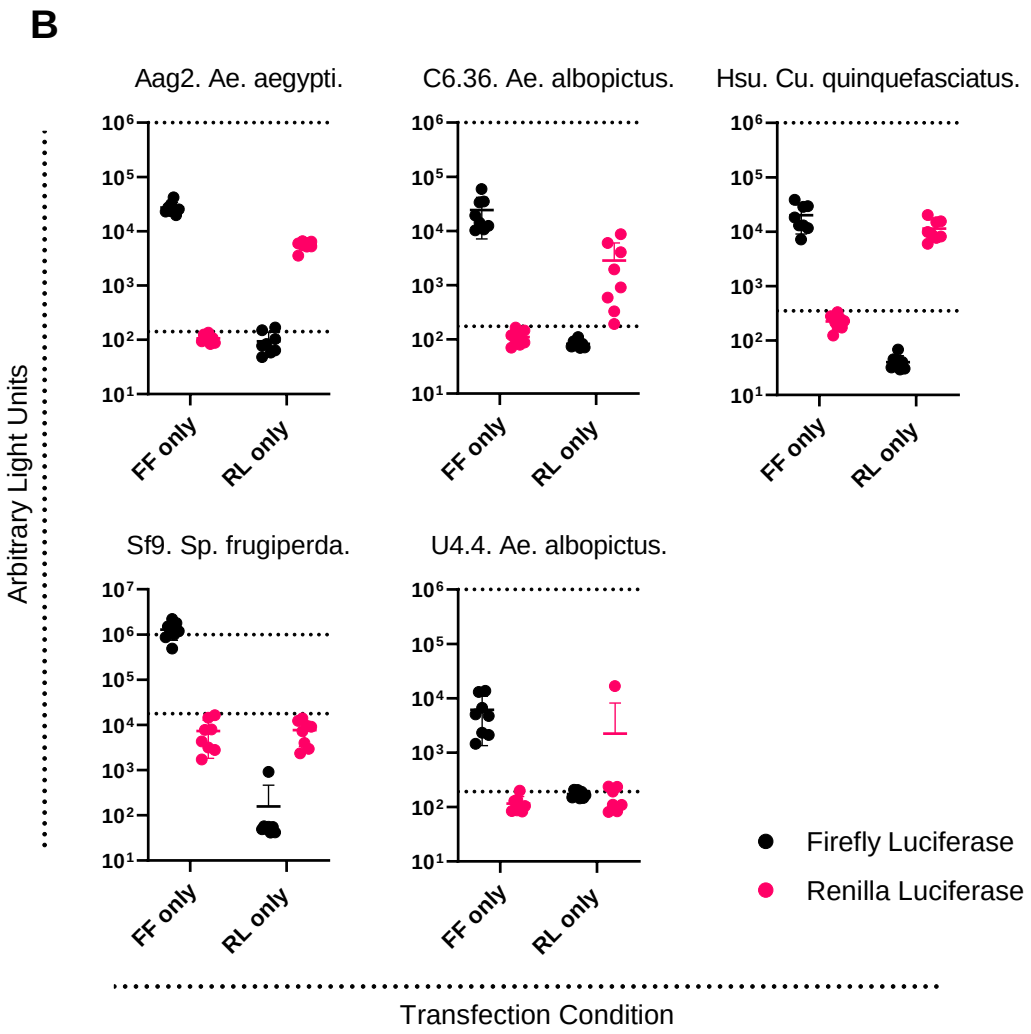


Figure 38: Graphs of controls for Experiment 1. Single luciferase controls are created alongside the main experiment and facilitate calculation of the Renilla luciferase (RL) background threshold (99.9% CI of RL measurements for transfection condition "FF only"). They furthermore offer additional data if problems are identified in the full data set. Firefly luciferase (FF) and RL are represented on independent graphs for each cell line with arbitrary light units (ALU) on the y-axis and control condition on the x-axis. Each luciferase is represented with a different colour and every sample is plotted as a dot with mean and SD indicated. The RL background threshold and FF quenching threshold are marked with horizontal lines. N = 8 throughout.

Luciferase activity results for the single lysate controls were generated concurrently with data for the full set of samples for each cell line. The raw luciferase measurements (arbitrary light units, ALU) for each sample are shown in Figure 38 with independent graphs for each cell line. All eight repeats for each condition are shown, of which two per cell line were previously used in the lysate volume controls (Figure 37).

The single luciferase controls are used to calculate the actual RL background threshold for each cell line (99.9% CI of RL measurements in the "FF only" condition), which are described in Table 28 and indicated within Figure 38.

Table 28: Statistics to determine RL background threshold

	Mean	SD	N	99.9% CI	Upper CI (background threshold, ALU)
Aag2	106	18	8	34	141
C6.36	111	34	8	65	175
Hsu	225	67	8	128	353
Sf9	7308	5487	8	10491	17799
U4.4	116	40	8	76	192

As discussed around Figure 35 in optimisation experiment 2, these results also offer additional information if problems were to occur with the full set of results. Looking at Figure 38 for an indication of what may occur in the main results, the FF activity in cell line Sf9 is often (6 repeats out of 8) above the FF quenching threshold. This will be controlled for in the results through screening and exclusion of samples with results above or below a relevant threshold but indicates that there may be TIS:3'UTR combinations that are then statistically under-represented in cell line Sf9.

Statistics

Appendix Table 11: RL activity of RL-null transfection condition "FF only" in experiment 9.7.18, used to calculate actual RL background threshold (Upper 99.9% CI)

Cell line	Mean	SD	N	99.9% CI	Upper CI (ALU)
Aag2	181	83	8	158	339

Appendix Table 12: RL activity of RL-null transfection condition "FF only" in full experiment, used to calculate actual RL background threshold (99.9% CI)

Cell line	Mean	SD	N	99.9% CI	Upper CI (ALU)
Aag2	106	18	8	34	141
C6.36	111	34	8	65	175
Hsu	225	67	8	128	353
Sf9	7308	5487	8	10491	17799
U4.4	116	40	8	76	192

Appendix Table 13: Summary data for generalised linear model

	Degrees of freedom (Df)	Deviance	Residual Df	Residual Dev	F	Pr(>F)
NULL			591	1407.54		
TIS	4	53.1333	587	1354.4	140.005	<0.001
3’ UTR	2	92.5183	585	1261.88	487.568	<0.001
Cell_Line	4	1014.14	581	247.749	2672.23	<0.001
TIS:Cell_Line	16	15.1661	565	232.582	9.99063	<0.001
3’UTR:Cell_Line	8	161.706	557	70.8765	213.046	<0.001
TIS:3’UTR	8	13.3279	549	57.5486	17.5594	<0.001

Appendix Table 14: Generalised linear model with a log-link function and Gamma family distribution to account for increasing variance with the mean. Model was conceived and produced by Phil Leftwich using a DHARMa package and ggeffects package.

Value			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	42.74	36.08 – 51.03	**<0.001**
Context [Lep]	1.07	0.86 – 1.34	0.531
Context [BmHi]	0.95	0.76 – 1.19	0.642
Context [BmLo]	0.39	0.31 – 0.50	**<0.001**
Context [Syn21]	0.40	0.32 – 0.49	**<0.001**
UTR [K10]	0.59	0.49 – 0.70	**<0.001**
UTR [P10]	3.91	3.15 – 4.87	**<0.001**
Cell_Line [C6.36..Ae..albopictus.]	2.73	2.19 – 3.40	**<0.001**
Cell_Line [Hsu..Cu..quinquefasciatus.]	0.03	0.03 – 0.04	**<0.001**
Cell_Line [Sf9..S..frugiperda.]	0.02	0.02 – 0.03	**<0.001**
Cell_Line [U4.4..Ae..albopictus.]	0.78	0.62 – 0.97	**0.027**
Context [Lep] * Cell_Line [C6.36..Ae..albopictus.]	1.01	0.77 – 1.32	0.955
Context [BmHi] * Cell_Line [C6.36..Ae..albopictus.]	1.28	0.98 – 1.68	0.072
Context [BmLo] * Cell_Line [C6.36..Ae..albopictus.]	1.82	1.39 – 2.39	**<0.001**
Context [Syn21] * Cell_Line [C6.36..Ae..albopictus.]	1.56	1.19 – 2.03	**0.001**

Value			
Predictors	*Estimates*	*CI*	*p*
Context [Lep] * Cell_Line [Hsu..Cu..quinquefasciatus.]	1.02	0.78 – 1.33	0.899
Context [BmHi] * Cell_Line [Hsu..Cu..quinquefasciatus.]	1.26	0.96 – 1.65	0.098
Context [BmLo] * Cell_Line [Hsu..Cu..quinquefasciatus.]	1.24	0.94 – 1.63	0.122
Context [Syn21] * Cell_Line [Hsu..Cu..quinquefasciatus.]	2.48	1.89 – 3.24	**<0.001**
Context [Lep] * Cell_Line [Sf9..S..frugiperda.]	0.74	0.57 – 0.97	**0.030**
Context [BmHi] * Cell_Line [Sf9..S..frugiperda.]	1.21	0.92 – 1.59	0.163
Context [BmLo] * Cell_Line [Sf9..S..frugiperda.]	1.34	1.01 – 1.76	**0.036**
Context [Syn21] * Cell_Line [Sf9..S..frugiperda.]	1.92	1.47 – 2.52	**<0.001**
Context [Lep] * Cell_Line [U4.4..Ae..albopictus.]	0.90	0.69 – 1.18	0.447
Context [BmHi] * Cell_Line [U4.4..Ae..albopictus.]	0.89	0.68 – 1.16	0.390
Context [BmLo] * Cell_Line [U4.4..Ae..albopictus.]	1.32	1.01 – 1.74	**0.042**
Context [Syn21] * Cell_Line [U4.4..Ae..albopictus.]	1.12	0.85 – 1.46	0.421

Value			
Predictors	*Estimates*	*CI*	*p*
UTR [K10] * Cell_Line [C6.36..Ae..albopictus.]	0.95	0.79 – 1.16	0.631
UTR [P10] * Cell_Line [C6.36..Ae..albopictus.]	0.32	0.26 – 0.39	**<0.001**
UTR [K10] * Cell_Line [Hsu..Cu..quinquefasciatus.]	0.83	0.69 – 1.01	0.059
UTR [P10] * Cell_Line [Hsu..Cu..quinquefasciatus.]	0.61	0.50 – 0.75	**<0.001**
UTR [K10] * Cell_Line [Sf9..S..frugiperda.]	1.76	1.45 – 2.13	**<0.001**
UTR [P10] * Cell_Line [Sf9..S..frugiperda.]	9.61	7.84 – 11.79	**<0.001**
UTR [K10] * Cell_Line [U4.4..Ae..albopictus.]	1.03	0.85 – 1.25	0.748
UTR [P10] * Cell_Line [U4.4..Ae..albopictus.]	0.44	0.36 – 0.54	**<0.001**
Context [Lep] * UTR [K10]	1.33	1.10 – 1.61	**0.004**
Context [BmHi] * UTR [K10]	1.35	1.12 – 1.64	**0.002**
Context [BmLo] * UTR [K10]	0.68	0.56 – 0.82	**<0.001**
Context [Syn21] * UTR [K10]	1.47	1.22 – 1.79	**<0.001**
Context [Lep] * UTR [P10]	1.09	0.89 – 1.33	0.412
Context [BmHi] * UTR [P10]	0.69	0.57 – 0.85	**<0.001**
Context [BmLo] * UTR [P10]	0.90	0.74 – 1.11	0.325
Context [Syn21] * UTR [P10]	0.98	0.80 – 1.20	0.822

Value			
Predictors	*Estimates*	*CI*	*p*
Observations	592		
R^2 Nagelkerke	0.990		

Appendix Table 15: Full data Figure 13 panel B

TIS	**estimated mean (FF/RL)**	**SE**	**df**	**asymp.LCL**	**asymp.UCL**
Kozak	15.49	0.468	Inf	14.60	16.44
BmHi	16.05	0.451	Inf	15.19	16.96
BmLo	6.84	0.192	Inf	6.47	7.23
Lep	17.44	0.491	Inf	16.51	18.43
Syn21	10.58	0.298	Inf	10.02	11.18
contrast	**ratio**	**SE**	**df**	**asymp.LCL**	**asymp.UCL**
Kozak / BmHi	0.965	0.0399	Inf	0.862	1.08
Kozak / BmLo	2.266	0.0935	Inf	2.024	2.536
Kozak / Lep	0.888	0.0367	Inf	0.794	0.994
Kozak / Syn21	1.464	0.0605	Inf	1.308	1.639
BmHi / BmLo	2.347	0.0933	Inf	2.106	2.616
BmHi / Lep	0.92	0.0366	Inf	0.826	1.026
BmHi / Syn21	1.517	0.0603	Inf	1.361	1.691
BmLo / Lep	0.392	0.0156	Inf	0.352	0.437
BmLo / Syn21	0.646	0.0257	Inf	0.580	0.72
Lep / Syn21	1.648	0.0655	Inf	1.479	1.837

Results are averaged over the levels of: UTR, Cell_Line

Confidence level used: 0.95

Conf-level adjustment: tukey method for comparing a family of 5 estimates

Intervals are back-transformed from the log scale

Appendix Table 16: Full data Figure 13 panel C

3'UTR	**Estimated mean (FF/RL)**	**SE**	**df**	**asymp.LCL**	**asymp.UCL**
SV40	9.31	0.203	Inf	8.92	9.72
K10	6.61	0.144	Inf	6.33	6.9

P10	32.27	0.735	Inf	30.87	33.75
contrast	**ratio**	**SE**	**df**	**asymp.LCL**	**asymp.UCL**
SV40 / K10	1.409	0.04341	Inf	1.311	1.515
SV40 / P10	0.289	0.00909	Inf	0.268	0.311
K10 / P10	0.205	0.00645	Inf	0.190	0.22

Results are averaged over the levels of: Context, Cell_Line

Confidence level used: 0.95

Conf-level adjustment: tukey method for comparing a family of 3 estimates

Intervals are back-transformed from the log scale

Appendix Table 17: Full data Figure 13 panel D

TIS	**3'UTR**	**Estimated means (FF/RL)**	**SE**	**df**	**asymp.LCL**	**asymp.UCL**
Kozak	SV40	11.62	0.566	Inf	10.56	12.78
	K10	7.33	0.357	Inf	6.66	8.06
	P10	43.68	2.578	Inf	38.91	49.03
BmHi	SV40	12.3	0.599	Inf	11.18	13.53
	K10	10.5	0.511	Inf	9.54	11.55
	P10	32.04	1.56	Inf	29.12	35.25
BmLo	SV40	6.03	0.294	Inf	5.48	6.64
	K10	2.59	0.126	Inf	2.35	2.85
	P10	20.49	0.998	Inf	18.63	22.55
Lep	SV40	11.57	0.563	Inf	10.52	12.73
	K10	9.7	0.472	Inf	8.81	10.67
	P10	47.32	2.305	Inf	43.01	52.06
Syn21	SV40	7.03	0.342	Inf	6.39	7.73
	K10	6.54	0.318	Inf	5.94	7.19
	P10	25.81	1.257	Inf	23.46	28.39
	$contrasts					
TIS	**contrast**	**ratio**	**SE**	**df**	**asymp.LCL**	**asymp.UCL**
Kozak	SV40 / K10	1.585	0.1092	Inf	1.349	1.863
	SV40 / P10	0.266	0.02036	Inf	0.222	0.318
	K10 / P10	0.168	0.01284	Inf	0.140	0.201
BmHi	SV40 / K10	1.171	0.08067	Inf	0.997	1.376
	SV40 / P10	0.384	0.02644	Inf	0.327	0.451
	K10 / P10	0.328	0.02257	Inf	0.279	0.385

BmLo	SV40 / K10	2.333	0.16067	Inf	1.985	2.741
	SV40 / P10	0.294	0.02028	Inf	0.251	0.346
	K10 / P10	0.126	0.00869	Inf	0.107	0.148
Lep	SV40 / K10	1.193	0.08218	Inf	1.015	1.402
	SV40 / P10	0.244	0.01684	Inf	0.208	0.287
	K10 / P10	0.205	0.01411	Inf	0.174	0.241
Syn21	SV40 / K10	1.075	0.07405	Inf	0.915	1.263
	SV40 / P10	0.272	0.01875	Inf	0.232	0.32
	K10 / P10	0.253	0.01744	Inf	0.216	0.298

TIS*3'UTR

Results are averaged over the levels of: Cell_Line

Confidence level used: 0.95

Intervals are back-transformed from the log scale

Conf-level adjustment: Tukey method for comparing a family of 3 estimates

Appendix D: Supplemental Information - Chapter 4

Preliminary experiments

Following initial testing in cell lines Aag2, Hsu and Sf9 with fluorescent reporter plasmid AGG1020, it was decided to develop the CRISPRa assay in the context of a dual luciferase reporter assay.

Initial experiments with fluorescent reporter plasmid AGG1020 gave the following indications of the performance of the CRISPRa assay (results not shown):

- There was no excessive cell death conferred by the transfection at three different total amounts of nucleic acids (553ng/well to 2212ng/well, in a 12-well plate).
- The multi-modal nature of transfected components (two plasmid and one *in vitro* transcribed RNA) was not a barrier to activation of the Tet response element (TRE) and subsequent expression of the reporter protein.
- Expression of the reporter protein was highest when all three CRISPRa components were co-transfected – dCas9-VPR plasmid, reporter plasmid and TetO-specific sgRNA.
- Background expression of the reporter protein occurred wherever the reporter plasmid was transfected, seemingly independent of presence/absence of dCas9-VPR plasmid.
- These results were observed in all cell lines tested: Aag2, Hsu and Sf9. The extent of background expression of the reporter was cell line dependent.
- These results could be seen at 24hrs and at 48hrs (the only time points considered).
- Different TetO-specific sgRNAs appeared to give different amounts of reporter expression.

To enhance utility of the assay, it was decided to replace the fluorescent reporter protein with firefly luciferase. This change to a dual luciferase reporter assay aims to deliver a quantitative resolution that can distinguish between different efficacies of sgRNA and between different amounts of sgRNA (i.e. different efficacies of promoter). A by-product of this format is an increase in throughput capacity conferred by use of a microplate reader (Glo-max multi+) for reporter measurement. Further advantages include the presence of a

transfection control (the control luciferase). Given the robust transfection efficiency seen in preliminary fluorescent experiments, addition of a further plasmid to the transfection was not a concern.

Characterisation of the CRISPRa dual luciferase assay

Following sequence confirmation of the luciferase CRISPRa reporter plasmid AGG1078, a series of optimisation experiments were designed and implemented to characterise the assay. The paramount aim of this optimisation was to determine appropriate constant values for each confounding variable so that any one experimental variable can be reliably tested. Part of this process was to ensure that the results of changing the experimental variable will sit within the dynamic range of the CRISPRa assay; the results of the CRISPRa assay must in turn sit within the dynamic range of the dual luciferase assay (as described in Chapter 3). The secondary aim of this optimisation was to provide data that can be used to understand and troubleshoot unexpected outcomes in experimental use of the CRISPRa dual luciferase assay.

The CRISPRa assay has three components: dCas9-VPR protein, target sequence specific (TetO) sgRNA and a reporter transgene/plasmid. The dual luciferase reporter assay has two components: the experimental luciferase (from the reporter transgene) and the control luciferase. In practical terms, this amounts to four nucleic acid components that must be co-transfected (Figure 42). The main variable for these components is the amount of nucleic acid that is transfected as part of the assay (discussed as ng/well). For the sgRNA component there are additional considerations: the sequence of the proto-spacer and whether the sgRNA is transfected as RNA (*in vitro* transcribed (*iv*)) or as plasmid DNA that transcribes the sgRNA *in situ*. The backbone structure of the sgRNA impacts the binding activity of sgRNA (Jinek et al., 2012, Dong et al., 2015) and therefore the reporter output of the CRISPRa assay. This is discussed further in Chapter 5.

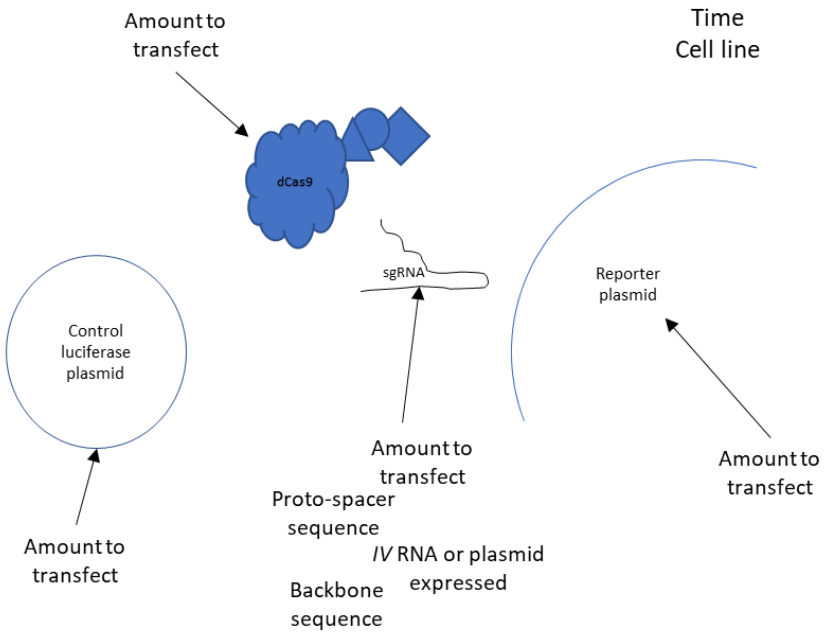


Figure 39: Cellular components of the CRISPRa assay. Variables for each component of the CRISPRa assay are shown in this cartoon representation. Overarching variables, "time" and "cell line", are listed in the top right.

The other variables of concern are time and cell line. The preliminary (fluorescent) experiments indicated that the CRISPRa assay could produce results from before 24 hours post transfection and after 72 hours. These assays were not quantitative, however, and it is not known whether the time point for harvesting samples will affect the ability of the assay to discriminate between different amounts or types of sgRNA.

Based on the preliminary experiments, the CRISPRa assay was functional in all three insect cell lines tested – *Ae. aegypti* derived Aag2; *C. quinquefasciatus* derived Hsu and *S. frugiperda* derived Sf9. Differences were noted, however, in the background expression of reporter protein from the unstimulated reporter plasmid. Characterisation of the assay was therefore carried out in all three cell lines of interest, wherever possible.

Design principles

Optimisation experiments are carried out for three cell lines and results are extrapolated for other cell lines in later experiments. Serial dilutions are typically done for at least two variables as the pairwise interactions between variables are reasonably expected to affect outcomes. For sgRNA where amount and 'strength' (binding affinity) are both variables, two 'strengths' of sgRNA (different TetO target sequences) are used. Multiple time points are considered in the first optimisation and are not re-visited. Early optimisation experiments make use of *in vitro* transcribed sgRNA before moving to plasmid-expressed sgRNA in later experiments.

Optimisation experiment 1: Reporter plasmid and sgRNA (*in vitro* transcribed)

The first optimisation experiment with the CRISPRa dual luciferase assay was carried out in three cell lines of interest (Aag2, Hsu and Sf9) at two time points, 24 hours and 48 hours post-transfection. This optimisation experiment primarily looks to examine:

- Whether samples can be collected at 48 hours post transfection (in deference to practical preferences)
- The characteristics of background reporter expression from the unstimulated reporter plasmid
- The relationship of different amounts of the reporter plasmid with different amounts of sgRNA, for two 'strengths' of sgRNA (different affinities of the sgRNA for the TetO sequence as they each have different target sequences within TetO)
- What amount of reporter plasmid and what amount of *in vitro* transcribed sgRNA are most suited to use in the CRISPRa assay?

Figure *40*, Figure *41* and Figure *42* show the results of optimisation experiment 1. Data is split by cell line (different figures), then by time point at which the data was collected – at 24 hours or 48 hours post transfection. To represent the three variables, 'strength' of sgRNA is shown in separate categories on the x-axis with "sgRNA1" and "sgRNA3" having different proto-spacer sequences (Table 16) within the TetO repeats of the tetracycline response element (TRE). Different transfected amounts of each sgRNA (ng/well) are shown as different x-axis points within each 'strength' of sgRNA. The third variable, amount of reporter plasmid transfected, is shown in different colours for each x-axis group (noted in the legend).

Controls are shown at the far right of each x-axis and were done for each cell line and time point. CRISPRa controls were done for three amounts of reporter plasmid and dual luciferase assay controls were done without reporter plasmid. All data are shown as firefly luciferase activity on the y-axis. This representation was chosen as not all controls contain both firefly and Renilla luciferase. Graphs of FF/RL are included in supplemental information (Appendix Figure 3). Each data point is the mean of 1 to 4 repeats and standard deviation is shown where present.

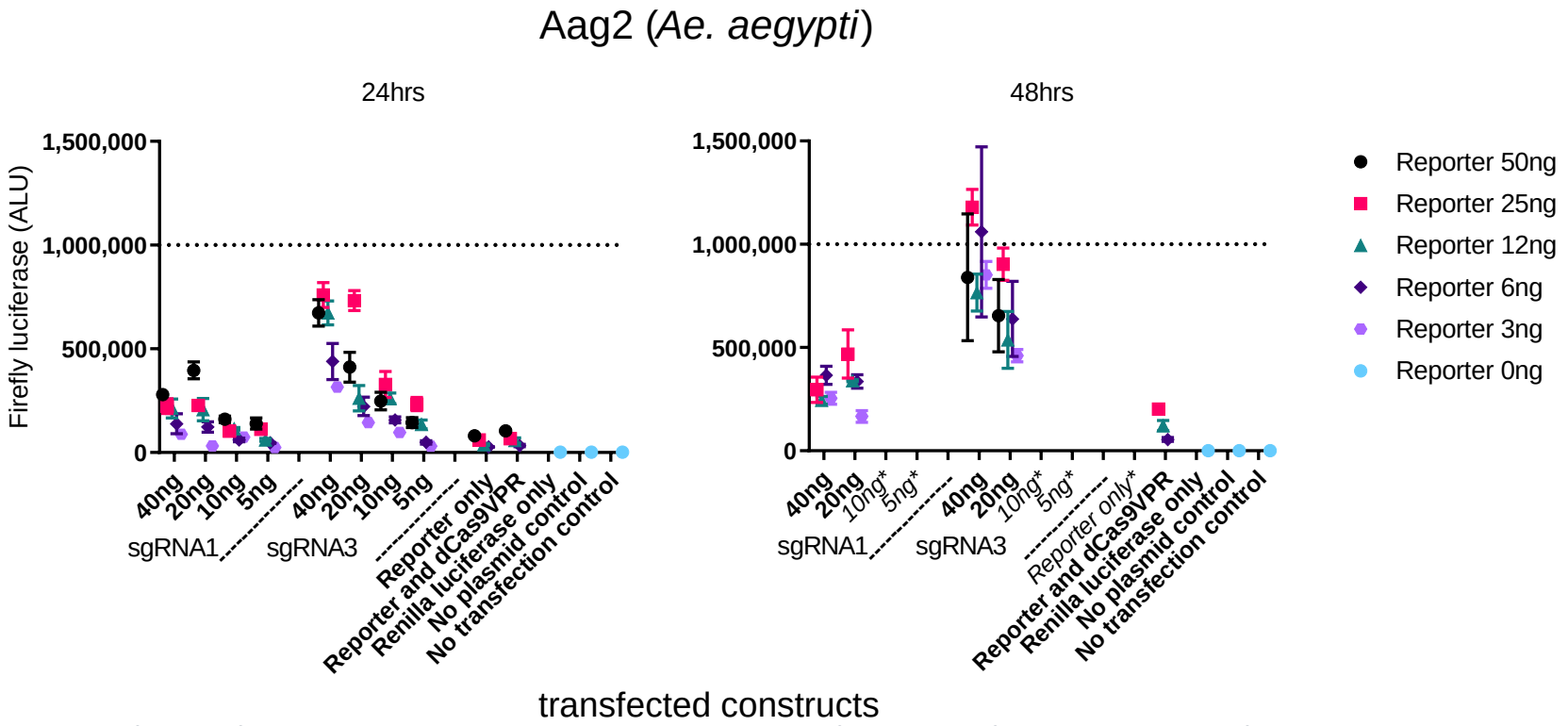


Figure 40: Graphs of results from optimisation 1 in cell line Aag2. Two graphs of the same information are shown for samples harvested at 24hrs post transfection (left) and for samples harvested at 48hrs post transfection (right). Both graphs show activity of reporter firefly luciferase (FF) on the y-axis with units in arbitrary light units (ALU). A dotted line is marked horizontally at 10^6 ALU, which is determined as the FF quenching threshold (as discussed in chapter 3). The ng amount of reporter plasmid transfected in each group is indicated with different colours (listed at far right). The x-axis shows which constructs were present in a transfection. Unless indicated otherwise, this includes the CRISPRa components (dCas9-VPR expressing plasmid, reporter plasmid, *in vitro* transcribed TetO-specific sgRNA) and dual luciferase assay component (Renilla luciferase). Some groups have no results (user error) and are indicated with italic font and an asterisk on the x-axis. “sgRNA1” and “sgRNA3” refer to two different TetO-specific sgRNA sequences. Dashed lines are present on the x-axis for visual clarity. Results are shown as mean with standard deviation. N= 1-4.

Where needed, the firefly luciferase (FF) quenching threshold is indicated (10^6 ALU), and any missing data groups are noted with italics and an asterisk in the x axis label. Some samples were excluded based on user error. N = 1 to 4 for each condition and samples with excluded results were not re-analysed. Analysis is carried out independently for each cell line tested and results across cell lines are then discussed.

Results of optimisation experiment 1 in cell line Aag2

Although it is difficult to draw conclusions in the comparison of the two time points with a reduced data set at 48 hours, we can see in Figure *40* that reporter activity (firefly luciferase, ALU) is higher at 48 hours than at 24 hours. There appears to be more distinction between the negative control categories at 48 hours ("Reporter and dCas9-VPR" at different reporter amounts) and a greater distinction between reporter activity from different sgRNAs (sgRNA1 and sgRNA3). We can also see that there is less distinction between different amounts of sgRNA3 at different reporter amounts than there is in the 24 hour data set.

Looking more closely at the characteristics of background expression from the unstimulated reporter plasmid, we can see at both time points in Figure 40 that reporter activity of "Reporter and dCas9-VPR" (the unstimulated reporter plasmid) is directly related to the amount of reporter plasmid transfected. There is little difference between the presence/absence of the dCas9-VPR plasmid at 24 hours ("Reporter only" vs "Reporter and dCas9-VPR"), but the comparison cannot be made at 48 hours where "Reporter only" is missing. From a practical perspective, it is observed that the distinction between TetO-sgRNA containing groups (experimental conditions) and the unstimulated reporter groups (negative controls) appears to be greater at 48 hours than at 24 hours.

Moving on to the sgRNA containing groups (all *in vitro* transcribed (*iv*)), the data in Figure 40 allows an examination of the relationship of different amounts of reporter plasmid with different amounts of sgRNA, for two 'strengths' of sgRNA (different binding affinities). Singling out sgRNA1 and looking at both time points, there is little distinction in reporter activity (ALU) between different amounts of sgRNA1 with a constant amount of reporter plasmid (i.e. comparing sgRNA1 40ng with 20ng and 10ng, within a single colour). Where present (24 hours graph) there is a small distinction observed between sgRNA1 40ng and 5ng (not quantified) and at both time points there appears to be a trend of higher reporter activity associated with greater amounts of reporter plasmid (looking at different colours within a single x-axis group).

For the 'stronger' sgRNA3, there is a greater distinction in reporter activity between different amounts of sgRNA present in the transfection (e.g. between sgRNA3 40ng and 20ng, within a single amount of reporter plasmid (colour)). Where more data is available at 24 hours, it can be seen generally that more sgRNA3 in the transfection (x-axis) leads to more reporter output and that more reporter plasmid (colours) leads to more reporter output. The trend for increasing amounts of reporter plasmid leading to greater reporter activity plateaus from 25ng/well to 50ng/well (black and pink) and for several x-axis groups there is greater or equal activity from 25ng/well reporter plasmid as for 50ng/well. Before this plateau, from 3ng/well to 25ng/well reporter plasmid, there is not a linear effect of 2x reporter plasmid resulting in 2x reporter activity.

Comparing the two sgRNAs (sgRNA1 and sgRNA3), there is no or minimal difference in reporter activity at lower amounts of sgRNA (5-10ng/well) but a clear trend of greater reporter activity from sgRNA3 than sgRNA1 at the higher amounts of sgRNA (20-40ng/well). This is true for each amount of reporter plasmid (colours) and both time points (graphs).

The design of optimisation experiment 1 with several experimental variables and a low number of repeats (N = 1 to 4) precludes meaningful statistical analysis without a significant investment of time and expertise to produce a custom analytical model. Nonetheless, working from graphical representations such as Figure *40*, we were able to make estimates of appropriate amounts of reporter plasmid and *iv* sgRNA to use in further optimisation experiments and ultimately in the protocol for the CRISPRa dual luciferase assay. Further limited in cell line Aag2 by the partial dataset at 48 hours post transfection, we can make the following practical conclusions:

- 24 hour vs 48 hour time point: 48 hours has greater variation between repeats than 24 hours, but better distinction between the positive groups (sgRNA1 and sgRNA3) and better distinction between the positive groups and the negative controls ("Reporter plasmid and dCas9-VPR").
- Amount of reporter plasmid: greater amounts of reporter plasmid confer greater reporter activity in the stimulated (sgRNA present) and unstimulated (sgRNA absent) groups. At greater amounts of reporter plasmid, there is a greater ability to distinguish between stimulated and unstimulated groups. This effect plateaus between 25ng/well and 50ng/well.
- Amount of *iv* sgRNA: At 20-40ng/well of *iv* sgRNA, reporter activity is differentiated between different TetO specific sgRNAs. There is indication that transfecting more

iv sgRNA correlates with increased reporter activity, but this is less clear at lower amounts of sgRNA (5-10ng/well) and with the 'weaker' sgRNA (sgRNA1).

Results of optimisation experiment 1 in cell line Hsu

In cell line Hsu (Figure *41*), the full dataset was collected at both 24 hours and 48 hours post transfection. It is notable that there is less distinction between the amounts of reporter activity recorded at 24 hours and at 48 hours than was seen in cell line Aag2 (Figure *40*). There are areas where luciferase activity at 48 hours has decreased from those recorded at 24 hours (e.g. Reporter 12ng (green), sgRNA3, all transfection amounts), which is not seen at all in cell lines Aag2 or Sf9 (Figure *40*, Figure *42*). This trend in cell line Hsu is not present in all transfection groups: where there is 50ng/well reporter plasmid the luciferase activity does not appear to have decreased and for all reporter plasmid quantities, there is greater luciferase activity at 48 hours than at 24 hours for the unstimulated reporter groups ("Reporter only" and "Reporter and dCas9-VPR").

Focusing on the characteristics of background expression from the unstimulated reporter plasmid, we can see that reporter activity appears to increase with increased amount of reporter plasmid ("Reporter only" and "Reporter and dCas9-VPR") up to a plateau between 25-50ng/well reporter plasmid. This is seen for both time points, though more clearly at 48 hours. Looking at both time points, there is no clear difference in reporter activity between the presence and absence of dCas9-VPR. The differentiation between reporter activity from experimental groups (containing sgRNA) vs the unstimulated reporter groups is good for all conditions at 24 hours, for sgRNA amounts greater than 5ng/well. This differentiation is greatly diminished at 48 hours.

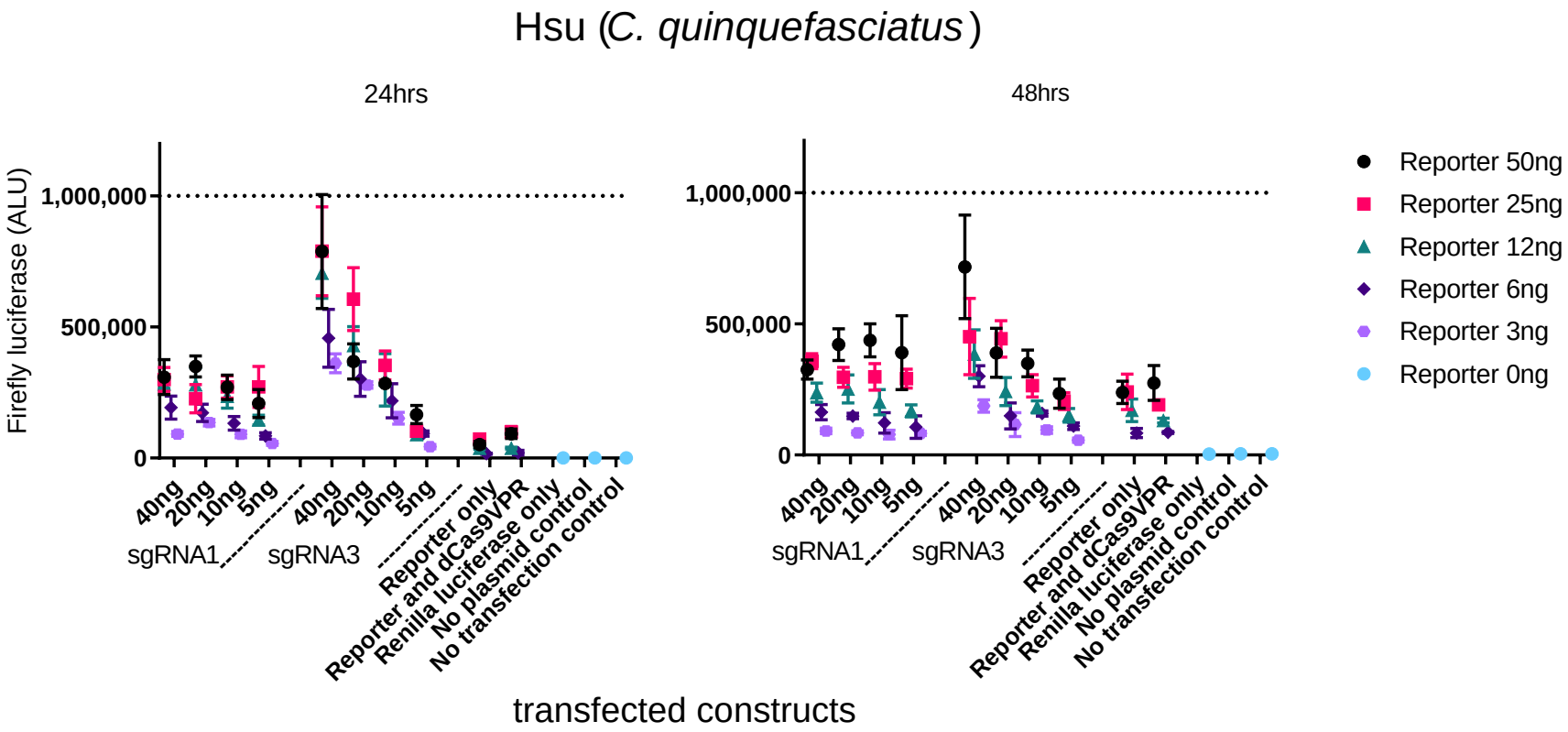


Figure 41: Graphs of results from optimisation 1 in cell line Hsu. Two graphs of the same information are shown for samples harvested at 24hrs post transfection (left) and for samples harvested at 48hrs post transfection (right). Both graphs show activity of reporter firefly luciferase (FF) on the y-axis with units in arbitrary light units (ALU). A dotted line is marked horizontally at 10^6 ALU, which is accepted as the FF quenching threshold (as discussed in chapter 3). The ng amount of reporter plasmid transfected in each group is indicated with different colours (listed at far right). The x-axis shows which constructs were present in a transfection. Unless indicated otherwise, this includes the CRISPRa components (dCas9-VPR expressing plasmid, reporter plasmid, *in vitro* transcribed TetO-specific sgRNA) and dual luciferase assay component (Renilla luciferase). "sgRNA1" and "sgRNA3" refer to two different TetO-specific sgRNA sequences. Dashed lines are present on the x-axis for visual clarity. Results are shown as mean with standard deviation. N= 2-4.

Looking next at the relationship of different amounts of reporter plasmid with different amounts of sgRNA, we can see similar trends as those in cell line Aag2 (Figure 40). In the 24 hours graph (Figure 41) there is a distinct trend of increasing quantities of sgRNA3 resulting in increased reporter activity. This trend is arguably present between the 20ng/well and 5ng/well groups for sgRNA1, but it is a much flatter line. At 48 hours this trend is gone for sgRNA1 and is much flattened for sgRNA3. In all cases there is increased reporter activity as the amount of reporter plasmid per well increases; this plateaus from 12ng/well to 50ng/well (reporter plasmid, colours). As the amount of sgRNA per well increases for sgRNA3 at 24 hours, the increase in reporter activity conferred by increasing amounts of reporter plasmid accelerates, then plateaus at 40ng/well sgRNA3 where 12, 25 and 50ng/well reporter plasmid are indistinguishable. This suggests that there is a saturation point that the CRISPRa dual luciferase assay can reach. This saturation is not of concern as the amount of reporter plasmid per well will be fixed below this point once the assay has been optimised.

For the 24 hours graph (Figure 41) there is a difference in reporter activity (ALU) between sgRNA1 and sgRNA3 (for matched amounts of reporter plasmid) once the amount of sgRNA transfected is greater than 10ng/well. This is not clearly seen at 48 hours. Higher transfection amounts of *iv* sgRNA resulting in greater differentiation between the two "strengths" of sgRNA is consistent with data from cell line Aag2 (Figure 40).

As for cell line Aag2, this dataset can be used to make estimates that inform practical choices for the further optimisation of the CRISPRa assay:

- 24 hour vs 48 hour time point: There is no immediate explanation for the incomplete pattern of decreased reporter activity at 48 hours vs 24 hours, particularly as the reporter protein (firefly luciferase) is stable (Promega, 2015) and should accumulate with time regardless of degradation of the sgRNA or other assay components. For this reason, these results in cell line Hsu were considered to be anomalous and not part of the decision to select a time point for the CRISPRa dual luciferase assay.
- Amount of reporter plasmid: Increased quantity of reporter plasmid present in the transfection confers increased reporter activity (ALU) in all conditions. This effect plateaus between 12 – 50ng/well of reporter plasmid.
- Amount of *iv* sgRNA: As in cell line Aag2, greater amounts of sgRNA confer unchanged or increased reporter activity (never decreased). Where there are differences detectable between matched sgRNA1 and sgRNA3 groups, they are

greater as the amount of sgRNA increases. 5ng/well of sgRNA is too low to differentiate sgRNA containing groups from unstimulated (negative control) groups.

Results of optimisation experiment 1 in cell line Sf9

In cell line Sf9 (Figure *42*), the full dataset was collected at both 24 hour and 48 hour time points post transfection. There is a considerable increase in luciferase activity between the 24 hour and 48 hour time points, with increased definition between different samples at 48 hours. This is visually emphasised by the decision to maintain a common y axis scale between the two graphs. The increase in luciferase activity between the 24 hour and 48 hour time points is not of a consistent amount for every data point – where there is more reporter plasmid or sgRNA, there are greater increases in luciferase activity (e.g. sgRNA1 has a steeper trend from 40ng/well to 5ng/well for any fixed amount of reporter plasmid than at 24 hours. 48 hours furthermore has greater distinction between amounts of reporter plasmid for 40ng/well of sgRNA1 than for 5ng/well of sgRNA1). This trend is seen for negative control "reporter only" but not "reporter and dCas9-VPR".

Although the "reporter and dCas9-VPR" negative control for Sf9 behaves in the same fashion as the CRISPRa controls for cell lines Aag2 and Hsu, the "reporter only" group shows reporter activity on par with that of the most active 'stimulated' (containing sgRNA) data points. This is the case for all amounts of reporter plasmid (colours) and for both time points. Of note, there is poor differentiation between the stimulated groups at 24 hours and the "reporter and dCas9-VPR" control. The difference in activity between "Reporter only" and "Reporter and dCas9-VPR" increases as the amount of reporter plasmid increases, at both time points. There is better differentiation of stimulated groups from "Reporter and dCas9-VPR" at 48 hours. As noted in previous cell lines, greater amounts of sgRNA confer better differentiation from the negative control.

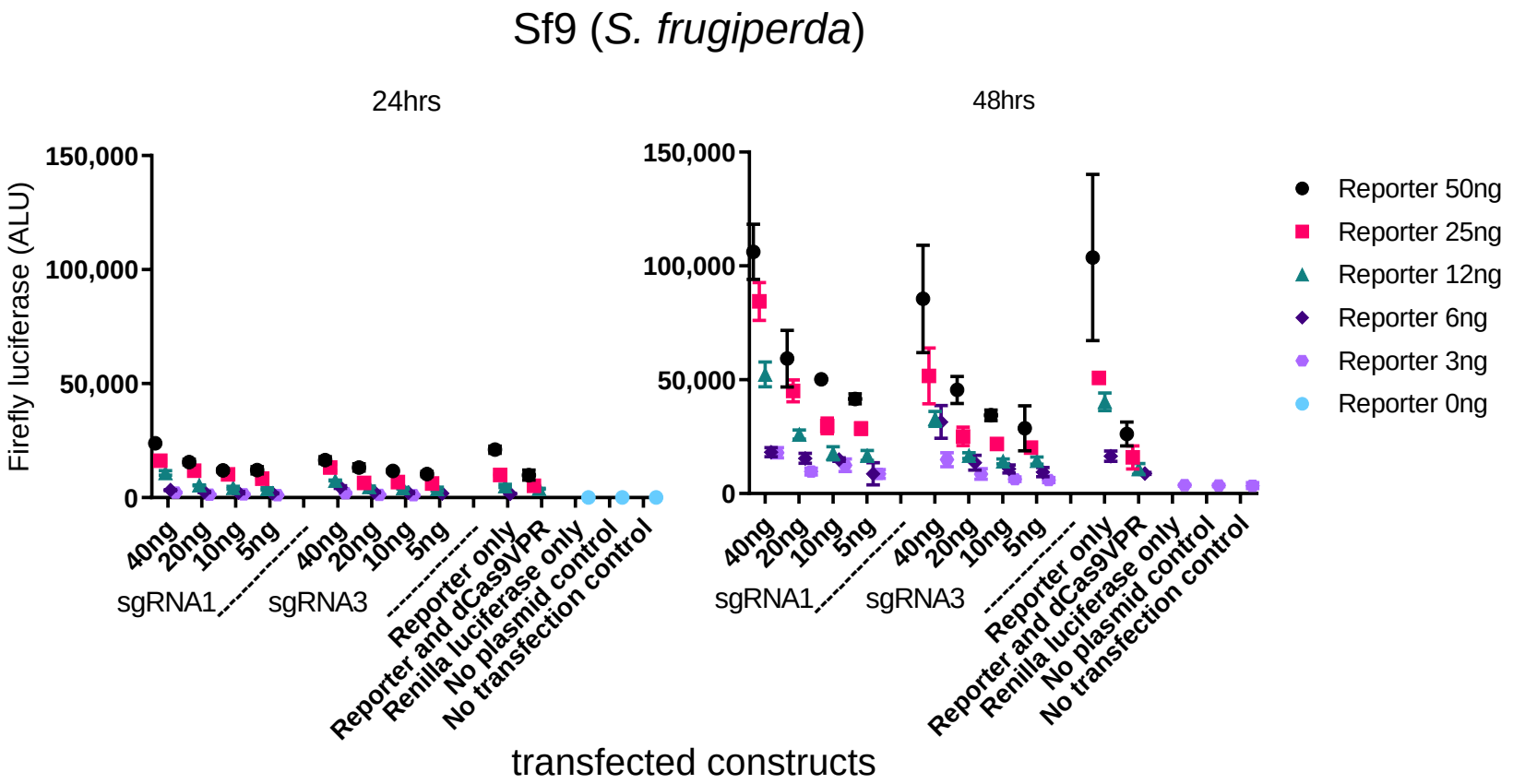


Figure 42: Graphs of results from optimisation 1 in cell line Sf9. Two graphs of the same information are shown for samples harvested at 24hrs post transfection (left) and for samples harvested at 48hrs post transfection (right). Both graphs show activity of reporter firefly luciferase (FF) on the y-axis with units in arbitrary light units (ALU). The ng amount of reporter plasmid transfected in each group is indicated with different colours (listed at far right). The x-axis shows which constructs were present in a transfection. Unless indicated otherwise, this includes the CRISPRa components (dCas9-VPR expressing plasmid, reporter plasmid, *in vitro* transcribed TetO-specific sgRNA) and dual luciferase assay component (Renilla luciferase). "sgRNA1" and "sgRNA3" refer to two different TetO-specific sgRNA sequences. Dashed lines are present on the x-axis for visual clarity. Results are shown as mean with standard deviation. N= 1-4.

Putting aside the "reporter only" conditions, it is noted that there is little difference between the two sgRNA 'strengths' at either time point. Within each sgRNA group there is a gentle trend of increasing sgRNA per well conferring increased reporter activity; this trend is more pronounced at 48 hours. Different amounts of reporter plasmid show a direct relationship with reporter activity, once more increasing as the amount of sgRNA per well increases (only clear at the 48 hour time point). No plateau is observed.

Coming back to the unexpectedly high activity of negative control "Reporter only", there is no obvious user error that could have resulted in these results. It should also be considered that the activity of sgRNA1 and sgRNA3 was compromised or is very weak in cell line Sf9 in this experiment; the resolution of the dual luciferase assay is such that expected trends can be seen in data otherwise considered 'background noise'. No explanation is presented for reduced activity of the CRISPRa assay in cell line Sf9 vs Aag2 or Hsu, if that is the case.

With the understanding that no results in Sf9 can be accepted to be above background, it is nonetheless noted that Figure 42 does not contradict the conclusions drawn from Figure 40 and Figure 41.

Conclusions from optimisation experiment 1:

Acknowledging that there is no one complete data set from optimisation experiment 1 without points of concern (missing data in cell line Aag2; unexpected decrease in activity at 48 hours in Hsu; experimental groups not significantly different from background in Sf9), it was decided to proceed with caution in further optimisation experiments. Caution is especially given to the potential for pair-wise interactions between different variables confounding results sought for a specific variable (e.g. the amount of reporter plasmid influencing the amount of activity seen from an sgRNA at different amounts of sgRNA or different amounts of dCas9-VPR plasmid). Where possible, experiments were carried out in multiple cell lines.

To manage the scope of this work, the 48 hour time point was selected based on there being better differentiation between sgRNA1 and sgRNA3 than at 24 hours. The amount of reporter plasmid transfected was reduced to a maximum of 25ng/well and the amount of *iv* sgRNA used was set to 40ng/well. These decisions were based on plateaus of increased reporter activity as the reporter plasmid was increased past 12-25ng/well and the increased differentiation between different sgRNAs seen at higher amounts of sgRNA. Increasing sgRNA quantity past 40ng/well was considered but ultimately discarded as further iterations of the CRISPRa assay were expected to use plasmid expressed sgRNA.

Optimisation experiment 2: dCas9-VPR expressing plasmid and reporter plasmid

Building on the work of optimisation experiment 1, optimisation experiment 2 looks at the effect on reporter activity of varying the amount of dCas9-VPR expressing plasmid (dCas9-VPR plasmid) transfected. With the expectation of pairwise interactions between CRISPRa components, this experiment was done at two fixed amounts of reporter plasmid. Two different 'strengths' of sgRNA are used, but sgRNA3 is replaced with sgRNA2, which is expected to have greater reporter activity (based on preliminary, fluorescent experiments). A plasmid expressed sgRNA, AGG1094, is included in this experiment at two amounts (ng/well). Plasmid AGG1094 was created for a different experiment; for the purposes of this optimisation work, it consists of a strong *Ae. aegypti* U6 promoter driving expression of TetO_sgRNA2 with 5' modifications. The 5' modifications are thought to hamper dCas9-VPR and dsDNA binding efficiency.

Although suitable CRISPRa negative controls were carried out with the control luciferase (Renilla luciferase (RL)), data is once more presented as untransformed firefly luciferase (FF) activity in arbitrary light units (ALU). This was done to represent the transfection groups where reporter activity was well in excess of the 10^6 ALU FF quenching threshold, beyond which the dual luciferase assay chemistry cannot be guaranteed to quench all FF activity before measuring RL activity (discussed in Chapter 3). Excessive FF activity would invalidate samples reported as FF/RL but is not limiting where data is presented untransformed.

The data for optimisation experiment 2 is presented for cell line Aag2 and is split across two graphs by the amount of reporter plasmid that was transfected (Figure *43*). N = 8 for all sgRNA containing groups and N = 7 for some control groups; data is shown as mean with standard deviation. Different transfection groups (sgRNA conditions) are shown on the x-axis and spaces have been added (dotted lines) for visual clarity; controls are shown at the far right of the x-axis for each graph. Three amounts (ng/well) of dCas9-VPR plasmid were tested and this data is represented in interleaved colours (described in the legend).

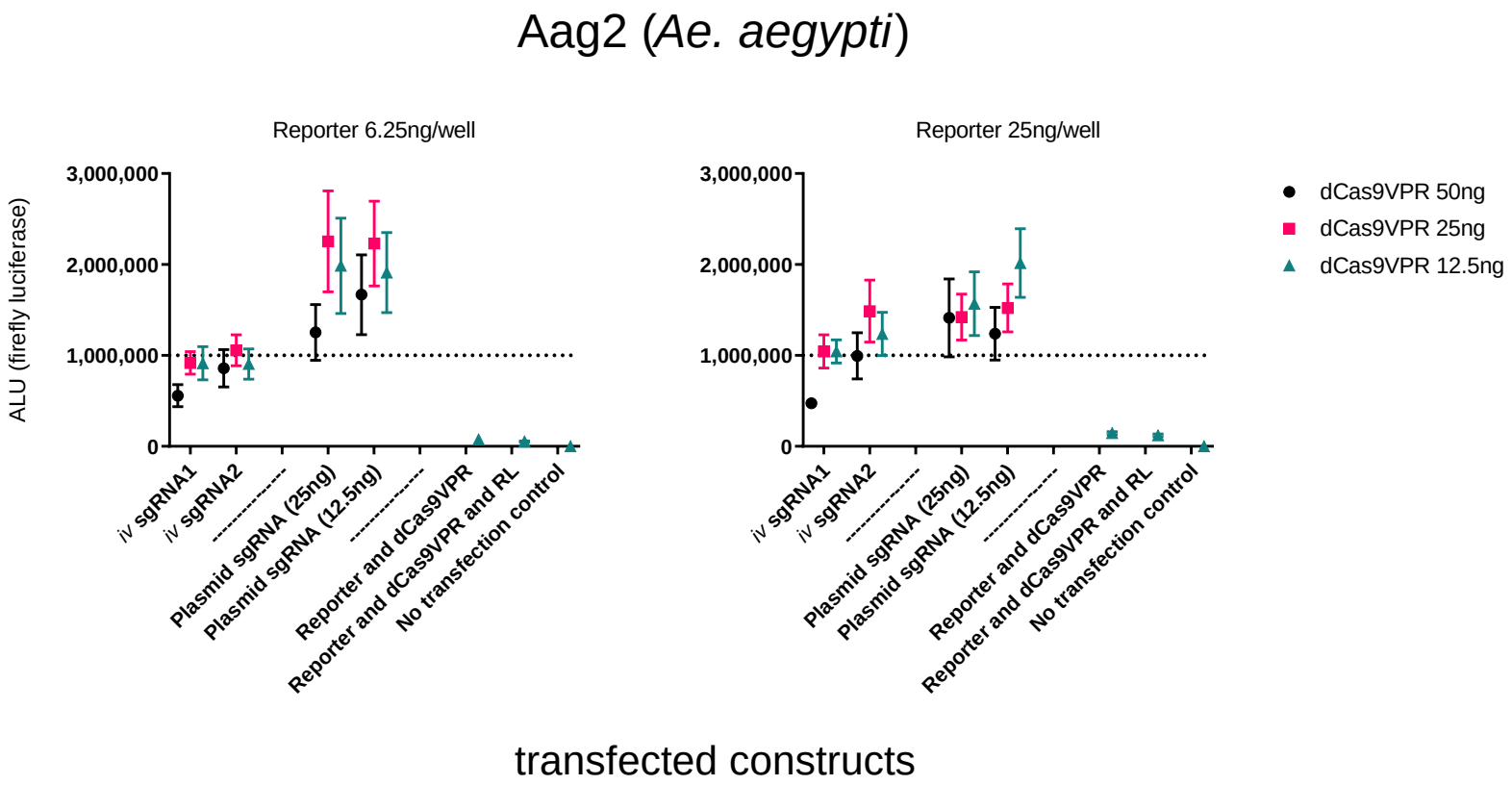


Figure 43: Graphs of results from optimisation 2 in cell line Aag2. dCas9-VPR expressing plasmid was transfected at three different amounts per well (indicated by colour) for two different amounts of reporter plasmid, shown on separate graphs. A single y-axis scale is used, showing reporter luciferase activity (firefly luciferase) in arbitrary light units (ALU). The firefly quenching threshold (discussed in chapter 3) is shown at 10^6 ALU for context. Each data point is shown as mean and standard deviation for transfection of different sgRNA components (listed on the x-axis) or for the control groups (at the far right of each x-axis). All sgRNA groups include TetO-specific sgRNA. sgRNA1 and -2 have different TetO spacer sequences. RL indicates the Renilla luciferase control. N = 7-8

In Figure 43 the amount of reporter plasmid used (25ng/well or 6.25ng/well) does not appear to interact with the other experimental variables (the relationship of different sgRNAs or different amounts of dCas9-VPR plasmid to one another). There is an increase in reporter activity as reporter plasmid increases for the *iv* sgRNAs, which is not proportional – a 4x increase in reporter plasmid does not cause a 4x increase in luciferase activity. The same is seen for the negative control (no sgRNA) groups. For groups with plasmid expressed sgRNA, there is an increase in luciferase activity as the amount of reporter plasmid is decreased. Variation is also increased, and this effect is confounded by the post-hoc understanding that the 5' modifications to the plasmid sgRNA do not have predictable interactions with other CRISPRa assay elements.

Focusing now on the change in amount of dCas9-VPR plasmid transfected, we can see that there is little or no change in reporter activity when all other factors are held constant. In several instances (*iv* sgRNA1; plasmid sgRNA) there is a decrease in luciferase activity associated with the highest amount of dCas9-VPR plasmid (50ng/well) as compared to the lower amounts (25 and 12.5ng/well). It is straightforward to conclude that the dCas9-VPR element is not a limiting factor of the CRISPRa assay from at least 12.5ng/well of plasmid.

Having noted that the CRISPRa assay is subject to saturation from various components (dCas9-VPR plasmid in Figure 43, reporter plasmid in Figure 40 and Figure 41), we can see that saturation is also occurring with the plasmid sgRNA. Where the amount of plasmid sgRNA transfected is halved from 25ng/well to 12.5ng/well, there is no decrease in luciferase activity. It is possible that this is also occurring between *iv* sgRNA1 and *iv* sgRNA2, where sgRNA2 is expected to be 'stronger' (have higher binding affinities) but does not necessarily reflect this in additional luciferase activity.

From optimisation experiment 2 it is concluded that the amount of dCas9-VPR plasmid transfected is not a limiting factor for the CRISPRa assay and that as little as 12.5ng/well is sufficient. It is also noted that any pairwise interaction between the amount of reporter plasmid and the amount of sgRNA does not have an observed effect on luciferase activity, nor is there an observed interaction of the amount of reporter plasmid and the amount of dCas9-VPR plasmid. Further work is needed to discern parameters within which the CRISPRa dual luciferase assay can reliably discriminate between different amounts of sgRNA expressing plasmid, which is a proxy for later experiments characterising the relative activity of different RNA pol III promoters.

Optimisation experiment 3: plasmid expressed sgRNA

Cell line Sf9

Having investigated the relationship of amount of *in vitro* transcribed sgRNA with reporter output (luciferase activity) in the CRISPRa assay in optimisation experiment 1, optimisation experiment 3 looks at the effect of different amounts of plasmid expressing sgRNA. Preliminary work was done in cell line Sf9 with a plasmid using RNA pol III promoters from the diamondback moth, *P. xylostella*. This plasmid, AGG1092, includes three TetO_sgRNA expressing cassettes with three *P. xylostella* U6 promoters (Huang et al., 2017) and was a kind gift from Tim Harvey-Samuel.

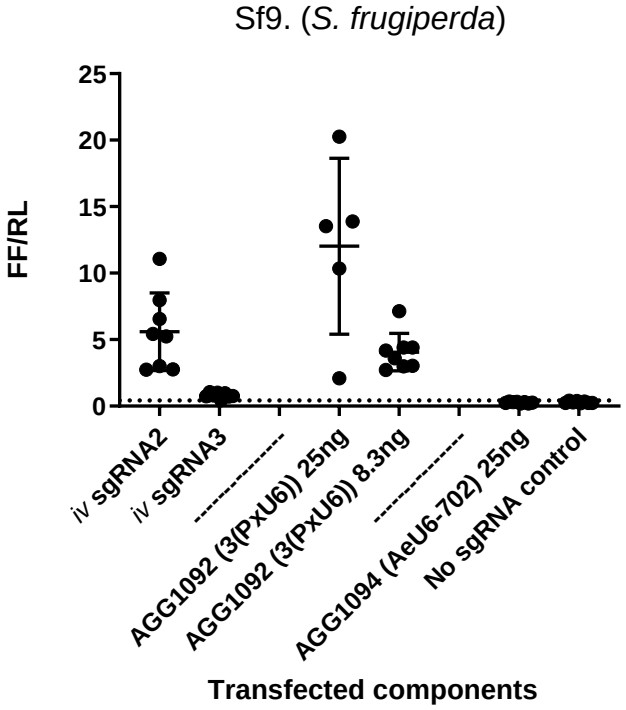


Figure 44: Graph of results for optimisation experiment 3 in cell line Sf9. Results are reported as FF/RL on the y-axis and are shown as individual dots for each sample with mean and standard deviation indicated. *iv* sgRNAs are included at 40ng/well each as a positive control for the CRISPRa assay. Plasmid AGG1092 contains three TetO_sgRNA expressing cassettes, each with a different *P. xylostella* U6 promoter. This plasmid was transfected at 25ng/well and at 0.3x, 8.3ng/well. Plasmid AGG1094 (as discussed for Figure 43) has an *Ae. aegypti* U6 promoter driving expression of a TetO_sgRNA with 5' modifications. The "No sgRNA control" includes all CRISPRa elements and Renilla Luciferase, but no sgRNA. The horizonal dotted line (along the x-axis) indicates the background threshold for the CRISPRa assay, set to the upper 99.9% CI of "No sgRNA control". N = 5 – 8.

Looking at a plasmid expressing sgRNA (AGG1092) in cell line Sf9 at two amounts (ng/well), there appears to be a decrease in reporter activity (FF/RL) corresponding with the decrease in plasmid amount (Figure 44). The positive controls with *in vitro* transcribed (*iv*) sgRNA

indicate that the CRISPRa assay is working with TetO_sgRNA2 but does not appear to have activity distinct from background with the 'weaker' TetO_sgRNA_3 ((referring to binding affinity) as determined in previous experiments). Both *iv* sgRNAs are transfected at 40ng/well. A further sgRNA expressing plasmid was included in this assay, AGG1094. This plasmid uses an *Ae. aegypti* RNA pol III promoter (U6-702) to express TetO_sgRNA with a 5' modification (as discussed on page 181). In moth cell line Sf9, this plasmid does not confer reporter activity, which is taken as an indication that the mosquito (*Ae. aegypti*) promoter is not effective in cells from such an evolutionarily distant species. This explanation is supported by the measurable activity of plasmid AGG1094 in optimisation experiment 2, where the promoter is species matched to the cell line.

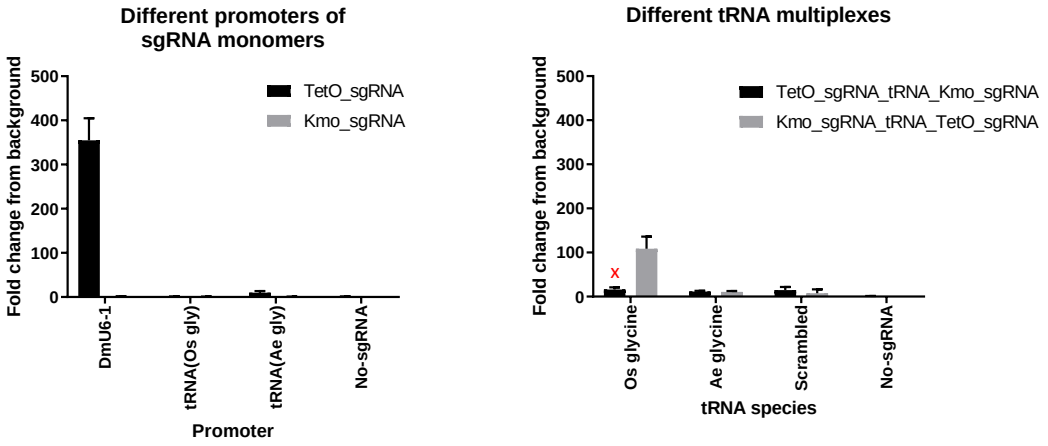
In Figure 44, results are presented as a FF/RL value, making full use of the dual luciferase reporter assay format. With results standardised to an internal control, it is not suitable to set the background threshold to the absence of luciferase activity. A "No sgRNA control" group is therefore included as a negative control. The background threshold for the CRISPRa dual luciferase assay is thus set to the upper 99.9% CI of the "No sgRNA control", in line with practices established in Chapter 3.

Positive control data, tRNA-sgRNA 'operon' in *D. melanogaster* cell line S2

Although investigation was made into the merits of using methods of expressing multiple sgRNAs from a single promoter, in the fashion of a bacterial operon, these avenues were discontinued due to negative results (Yan et al., 2016, Xie et al., 2015, Xie et al., 2017, Wang et al., 2017, Port and Bullock, 2016).

Outcomes of tRNA - sgRNA 'operons' were mixed and indicated moderate sgRNA activity in the absence of efficient processing – sgRNAs with tRNA sequences or sequence fragments were noted to be active in CRISPRa assays in mosquito cell lines. There was a differential effect depending on the position of the tRNA sequence (5' or 3'), though this too was inconsistent.

An sgRNA - tRNA experiment in *D. melanogaster* cells was more informative.



Appendix Figure 2: Graphs showing results of a CRISPRa assay in cell line S2. Each graph shows FF/RL, transformed to background as 1, on the y-axis and a consistent scale is used. Two constructs are shown for each category on the x-axes, these are noted in each legend. "Kmo_sgRNA" was the irrelevant sgRNA used as a negative control. The left graph shows constructs that used different promoter sequences – a *D. melanogaster* U6 promoter "DmU6-1"; no promoter, only tRNA(*Os* glycine); no promoter, only tRNA(*Ae* glycine); CRISPRa negative control "No-sgRNA". The right graph shows constructs with a consistent promoter (DmU6-1) and different tRNA sequences used in the sgRNA – tRNA – sgRNA 'operon'. A red cross indicates the construct with a sequence error. Mean and SD are indicated for each bar and N = 6 – 8.

To examine the hypothesis of promoter activity arising from the tRNA sequences themselves, the results on the left graph were generated. Promoter activity of a rice (*O. sativa*) tRNA (glycine) and that of an *Ae. aegypti* tRNA (glycine) were tested using the CRISPRa assay and the TetO-sgRNA. An irrelevant sgRNA (Kmo447) was used as a negative control for each

promoter. A positive control for the CRISPRa assay was included, using the *D. melanogaster* promoter U6-1. A CRISPRa negative control ("No sgRNA") was also used. Data is reported as fold change in FF/RL from "No sgRNA".

Looking first at the CRISPRa controls, it is noted that there is no expression from the CRISPRa negative control (either sgRNA) nor from the irrelevant sgRNA in the CRISPRa positive control. The TetO_sgRNA condition with the positive control promoter (DmU6-1) produces mean activity around 350 arbitrary light units (ALU). These results indicate that the CRISPRa assay performed as expected. The negative control sgRNA (Kmo447) also performed as expected.

Looking at the two middle conditions, with tRNA sequences in place of promoters, there was a very small but potentially non-zero activity from the *Ae. aegypti* tRNA (glycine). This is seen only for the TetO_sgRNA, not the negative control sgRNA (Kmo447). For the *O. sativa* tRNA (glycine), there is no activity measured for either sgRNA. These constructs were developed and tested as tRNA sequences in their native genomic context can have promoter elements within their coding sequence.

The results from the left graph suggest that there is no or minimal promoter activity from the tRNA sequences examined; it confirms too that the negative control sgRNA (Kmo447) does not produce measurable activity in the CRISPRa assay.

The right graph was designed to examine the relative activity of TetO_sgRNA when expressed 5' of a tRNA-sgRNA conjugate (black) or 3' of an sgRNA-tRNA conjugate (grey), i.e. in either the first or second position of the $sgRNA^{A}$-tRNA-$sgRNA^{B}$ structure. This was done with a consistent promoter (DmU6-1) and three tRNA sequences: *O. sativa* glycine, *Ae. aegypti* glycine and a 'scrambled' nonsense tRNA sequence. The data shown for monomeric TetO-sgRNA on the left graph can be considered a positive control for the data on the right graph (all data was generated in the same experiment).

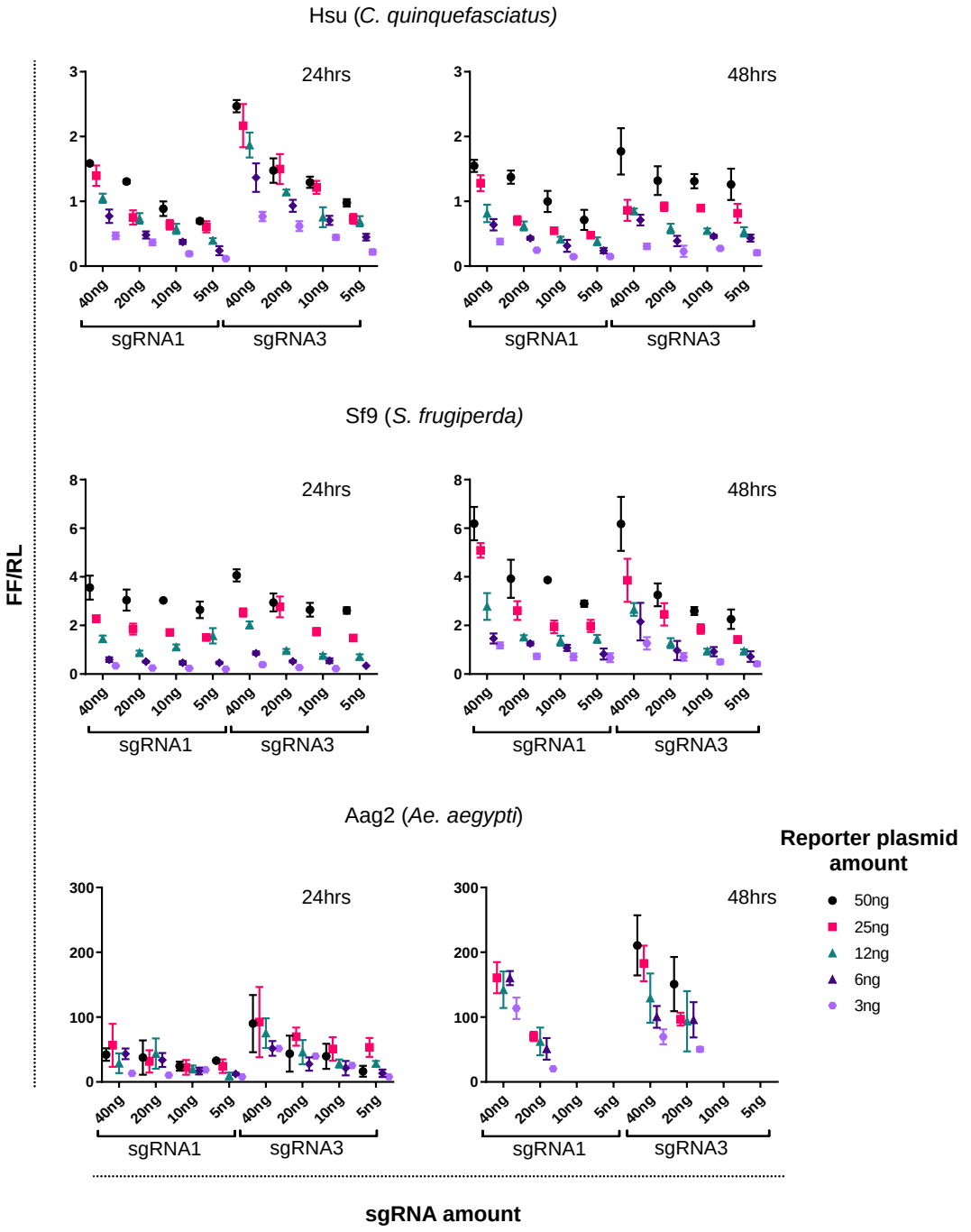
The right graph shows non-zero activity for both constructs with all three tRNAs. Looking at the scrambled tRNA sequence, which should not induce any cleavage of the sgRNA-tRNA-sgRNA conjugate, it is seen that each of the 5' and 3' conjugates greatly impede sgRNA activity (as compared to the monomer on the left graph), but do not silence it. There is greater activity with the TetO_sgRNA 5' of the tRNA-sgRNA conjugate.

With *Ae. aegypti* tRNA glycine there is activity above zero, but not above that of the scrambled tRNA. This was true for both arrangements of conjugate, i.e. whether the TetO_sgRNA was 5' or 3' relative to the tRNA. This suggests that there is no processing activity of the 'operon' conveyed by the *Ae. aegypti* tRNA sequence.

The *O. sativa* tRNA glycine has activity equal to that of the scrambled tRNA for TetO_sgRNA 5' of the tRNA-sgRNA conjugate, which is not unexpected as there was a production error with this plasmid and the tRNA sequence is truncated and expected to be non-functional (marked in Appendix Figure 2 with a red "x"). There is greater activity (circa 100 ALU) for the *O. sativa* tRNA glycine with the TetO_sgRNA 3' of the sgRNA_tRNA conjugate (grey), which was synthesised correctly. That either arrangement of conjugate shows activity greater than that of scrambled tRNA supports the hypothesis that the *O. sativa* tRNA sequence conveys processing of the TetO_sgRNA from the operon/its conjugate, in a *D. melanogaster* cell culture context (Port and Bullock, 2016). Such an effect was not seen in a mosquito cell context (*Ae. aegypti*).

Standardised data, companion figures for optimisation experiment 1

Due to CRISPRa controls not containing pRL-OpIE2 (which expresses the control (Renilla) luciferase), data for optimisation experiment 1 is presented as firefly luciferase (FF) values only in Figure 40, Figure 41 and Figure 42. For completeness, the standardised data (FF/RL) is presented in Appendix Figure 3.



Appendix Figure 3: Graphs of results from optimisation experiment 1 presented as standardised FF/RL values. Each graph shows results from a different cell line, at different time points.

Table of putative U6 promoters

Each of these U6 genes has an RNA polymerase III terminator sequence. The 'promoter' is taken to be the sequence up to 600nt 5' of the beginning of the U6 gene. Where a TATA-like sequence is not found, a proximal sequence element (PSE) is not sought. A local name is given only to promoter sequences with both TATA-like sequence and PSE.

Appendix Table 18: Mosquito putative U6 genes

Species	**U6 gene accession**	**TATA-like sequence**	**Presence of PSE**	**Notes**	**Local name**
Ae. aegypti	AAEL017702	TATATATA	Yes	(Konet et al., 2007)	AeU6-702
Ae. aegypti	AAEL017763	TATATAA	Yes		AeU6-763
Ae. aegypti	AAEL017774	TATATAA	Yes	(Konet et al., 2007)	AeU6-774
Ae. aegypti	AAEL017905	TATATAAA	Yes		AeU6-905
Ae. aegypti	AAEL028846	No			
Ae. aegypti	AAEL028848	TATATAA	Yes		AeU6-848
Ae. aegypti	AAEL028972	TATATAAA	Yes		AeU6-972
Ae. aegypti	AAEL029000	TATATAA	Yes		AeU6-000
An. albimanus	AALB015132 /AALBO15274	TATATATA	Yes	Identical (2 single nt changes)	AalbU6-132
Ae. albopictus	AALF029195	No			
Ae. albopictus	AALF029465	No			
Ae. albopictus	AALF029489	No			
Ae. albopictus	AALF029578	No			
Ae. albopictus	AALF029580	No			
Ae. albopictus	AALF029625	TATAAAA	No		
Ae. albopictus	AALF029628	No			
Ae. albopictus	AALF029629	No			
Ae. albopictus	AALF029637	TATATAT	No		

Species	U6 gene accession	TATA-like sequence	Presence of PSE	Notes	Local name
Ae. albopictus	AALF029700	No			
Ae. albopictus	AALF029725	TATATAA	Yes		AbU6-725
Ae. albopictus	AALF029726	TATATAAA	Yes		AbU6-726
Ae. albopictus	AALF029727	TATAAATATATAA	Yes		AbU6-727
Ae. albopictus	AALF029743	No			
Ae. albopictus	AALF029744	TATATATA	Yes		AbU6-744
Ae. albopictus	AALF029757	3, none near -30bp	No		
Ae. albopictus	AALF029791	TATAAAC	No		
Ae. albopictus	AALF029820	TATAAAT	No		
Ae. albopictus	AALF029826	No			
Ae. albopictus	AALF029842	No			
Ae. albopictus	AALF029845	No			
Ae. albopictus	AALF029870	No			
Ae. albopictus	AALF029888	No			
Ae. albopictus	AALF029903	No			
Ae. albopictus	AALF029910	TATAAAC	No		
Ae. albopictus	AALF029917	No			
Ae. albopictus	AALF029955	TATATAA	Yes		AbU6-955
Ae. albopictus	AALF029956	TATATAA	Yes		AbU6-956
Ae. albopictus	AALF029957	TATATAAA	Yes		AbU6-957

Species	U6 gene accession	TATA-like sequence	Presence of PSE	Notes	Local name
An. arabiensis	AARA015171	TATATATA	Yes		AaraU6-171
An. arabiensis	AARA015449	TATATATA	Yes	Weak PSE match	AaraU6-449
An. funestus	AFUN015538	TATATATA	Yes		AfunU6-538
An. funestus	AFUN015704	TATATATA	Yes		AfunU6-704
An. gambiae	AGAP013557	TATATA	Yes	(Konet et al., 2007)	AgU6-557
An. gambiae	AGAP013695	TATATA	Yes	(Konet et al., 2007)	AgU6-695
An. stephensi	ASTEI11842	TATATATA	Yes		AsteiU6-842
An. stephensi	ASTEI11858	TATATATA	Yes		AsteiU6-858
An. stephensi	ASTEI11917	TATATA	Yes		AsteiU6-917
C. quinquefasciatus/pipiens	CPIJ039543	TATATAA	Yes		CqU6-543
C. quinquefasciatus/pipiens	CPIJ039596	TATATAA	Yes		CqU6-596
C. quinquefasciatus/pipiens	CPIJ039653	TATATAA	Yes		CqU6-653
C. quinquefasciatus/pipiens	CPIJ039728	TATATAA	Yes		CqU6-728
C. quinquefasciatus/pipiens	CPIJ039801	TATATAA	Yes		CqU6-801

U6 gene promoter sequences

Each putative U6 promoter sequence is presented as up to 600bp 5' of the given Accession number. **U6 gene sequence and poly-T terminator are noted.** Asterisks (*) are used to indicate previously published promoters.

AeU6-763 (AAEL017763)

GGCAACTATAGAGTTTCCATGTTTCCAGACTTTCCCTCCCGGTAAACGGAGACAAAACGA
CAGACGTAAGTAGGTACATATGCATACCGCACGGACAAATCAAATTTGTCTGGCAGCTCC
AATTAGAGTCGTTAAAAATTTAACGATGCGCTAAATAACTTCAAGCTATTTGTCTCGCTG
GATTGGCTTCGAGTGGTAAGATCCTATCAAATGCCGAAAACAAAAAACTTTCTTCTTAAT
TGTTCGTTCTTCAACACCTCTCCATGGTGATAACGGATACGGTTTCATTGTCAGCATCCA
TCCTCCGAAAAATACATTACGCCTTGAAATATGCAATCGCAAACACGGATCTGTTTGGAA
CATTTATTTTACTATGAAGAGATGCGATAGGTAATATTTATTTGAGCGTTTAAGATACTC
ATTGTTCTCTCAAAGAATGTCATTGAAAGCCAACGAGGTCAAATCAAATATTATAATAAA
AAGGTCAAAGAGGACTAACTTAAAGCTCTCTTTATGGATAGGAAAAAATATTTTCGCCCA
TCGCTAGAACTTTTACCGTTTCCATTGAGTATATAACTAAGATGAATGAGGCTAATTGAT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTT

*AeU6-774 (AAEL017774)

TTCCCGTAGACAAAAATCAAATTTCATGGTATTTTTCGGCGATAAACAGTAAGTTAAAGG
GAACCCGAATTTAGTGCTATATAATTTAATTCCACTAGAGTTTGTATCCTTTGATAGATA
CGCGTATTTCGACCTCAACTGCAAGGCCGTCGTGTACTAGACTTGACTAATCCAGACTGG
TCTTTTAGTTATGACTTCTGTCCACATCTCCATACATTCAACGCACTGTGCGGCTGTGCT
GTGCGACTCCGTCGAGTCGACCAACATAGTTGAAACAAATTGAATATTTAATTGATCGTT
ATAGGAATGGTGTTAGATGAGTCATCCTTTACAGTAAGCACATACAGTATTATAATTGAA
GATCGTCGGCAGATAGGTGTGTAGGGTAGAGTATCAGCAATAAGTTGGGACGTTTGACTT
TTTGTAGGTAGACAAAAACTAAACTTTTTTTCGCTTCTCTATGTGTGCCCCTGGGTAGCG
TTCCGTTCCGATTGGGGTGCGAACGAATGAAATCGCCCATCGAGTTGATACGTCCATCCA
TCGCTAGAACCGCGTTCGCTGTAGAAGACTATATAAGAGCAGAGGCAAGAGTAGTGAAAT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTT

AeU6-905 (AAEL017905)

CCTATCGCATCTCTTCATAGTAAAATAAATGTTCCAAACAGATCCGTGTTTGCGATTGCA
TATTTCAAGGCGTAATGTATTTTTCGGAGGATGGATGCTGACAATGAAACCGTATCCGTT
ATCACCATGGAGAGGTGTTGAAGAACGAACAATTAAGAAGAAAGTTTTTTGTTTTCGGCA
TTTGATAGGATCTTACCACTCGAAGCCAATCCAGCGAGACAAATAGCTTGAAGTTATTTA
GCGCATCGTTAAATTTTTAACGACTCTAATTGGAGCTGCCAGACAAATTTGATTTGTCCG
TGCGGTATGCATATGTACCTACTTACGTCTGTCGTTTTGTCTCCGTTTACCGGGAGGGAA
AGTCTGGAAACATGGAAACTCTATAGTTGCCAGGTAGACCATCTGCCTCCGTCGGCTGGC
TGGATTCCAATTTGAATATTGGCTAATTGGAAGAGATGGAAGTTTTTGAATGGATGATTG
AATAATTGAAGCGACTCCGGGTACCTGTTTGTAAGCTCTGCAACAGTGCCATAGATTCGT
GTCAGTCCATCACTAGAATCAAATCAACTTGTACTTGATATATAAATGGCTTGGGTTAAT
GTGCTTCGGCAAGACATATACTAAAATTGGAACGATATAGAGAAGATTAGCATGGCCCCT
GCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTT

*AeU6-702 (AAEL017702)

GTCATTGACTTGACTACTTTTATTGAACTAATAATTCTATTAACAGAAATTGATAACAAC
AGTTTGCTTATGAAATAACATGATAGCCAAGTCAATAACAGGTCCAGTTTAGACTTTCGG
GTTTTCCATAGTAAATATACTACAAATAATATTAATGTTCCATGAAAAAAGGAGTAAGAG
TCTGGTAACCCTAGTGCACGCAAATATCTCGCGGGCATATTTGGTTGCTGAGGTATATTT

ATATTTGAACGCCATGAGAAAAAGCGGAAGAAATTGGCTCATGGCCGATTTTAAGGATAT
TTAAAAATTGTACAATGTACATATAATTAAACATCCGTTCCTTCAATGTGTTCTTTTTTT
TAAGCGTGTGTTAAAAGTTTGCTCTGGTGGTGAATTCACGCTCTACCCGTTCAGGCAGCA
TTCATCGAAAAGCCCTATCTGCTCGCACACATTTACAAAATGCTGATTGCGTTGTGTGCT
GAATGGGTCACTCGTCCGTCACTGCTTGCTGTGTACACTGTACAGTTACGCAGTCTGTGC
ATCGCTAGAATCATATTTACGGAAAAGTATTATATATACCCAATGCGTTGCTCATCGGTT
GTCCTAGCTTCGGCTGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTT

AeU6-848 (AAEL028848)

GTTAAAATCATATATTTTATGATCTTAACACCGTGTTTTAGCAGTTTTCCCTTAGGTGCA
CCACTAATGGTACACTTGCCCTATATAATTTAATTCCACTAGAGTTTGTATCCTTTGACA
GATACGCGTATTTCGACCTCAACTGCAAGGCCGTCGTGTACTAGACTTGACTAATCCAAA
CTGAAGGTTGTCTTTCAGTTATGACTTCTGTCCAACATCTCCATACATTCAACGCACTGT
GCGGCTGTGTGACTCCGTCGAGTCGACCAACATAGTTGAAACAAATTGAATATTTAATTG
ACCGTTATAGGAATGGTGTTAGATGAGTCATCCTTTACAGTAAGCACGTACAGTATTATA
ATTGAAGATCGTCGGCAGATAGGTGTGTAGGGTAGGGTATCAGCAATCAGTTGGGACGTT
TGACTTCTTCTAGGTAGACAAAAACATTTTTCGCTTCTCTATGTGTGCCCCTGGGTAGCG
TTCCGTTCCGATTGGGGTGCGAACGAATGAAATCGCCCATCGAGTTGATACGTCCATCCA
TCGCTAGAACCGCGTTCGCTGTAGAAGACTATATAAGAGCAGAGGCAAGAGTAGTGAAAT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTT

AeU6-972 (AAEL028972)

AATGTTCCAAACGAATCCGTGTTTGCGATTGCATATTTCAAGGCGTAATGTATTTTTCGG
AAGATTGTACTTATTATGATTGTGGATGCTGACAATGAAACCGTATCCGTTATCACCATG
GAGAGGTGTTGAAGAACGAACAATTAAGAAGAAAGTTTTTTGTTTTCGGCATTTGATAGG
ATCTTACCACTCGAAGCCAATCCAGCGAGACAAATAGCTTGAAGTTATTTAGTACATCGT
TAAATTTTTAACGACTCTACTTGGAGCTGCCAGACAAATGTGATTTGTCCGTGCGGTATG
CATATGTACCTACTTACGTCTGTCGTTTTGTCTCCGTTTACCGGGAGGGAAAGTCTGGAA
ACATGGAAACTCTATAGTTGCCAGGTAGACCATCTGCCTCCGTCGGCTGGCTGGATTCCA
ATTTGAATATTGGCTAATTGGAAGAGATGGAAGTTTTTGAATGGATGATTGAATAATTGA
AGCGACTCCGGGTACCTGTTTGTAAGCTCTGCAACAGTGACATAGATTTGTGTCAGTCCA
TCACTAGAATCAAATCAACTTGTACTTGCTATATAAATGGCTAGGACTAGCGGAAGATTT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTT

AeU6-000 (AAEL029000)

TCCAGACTTTCCCTCCCGGTAAACGGAGACAAAACGACAGACGTAAGTAGGTACATATGC
ATACCGCACGGACAAATCACATTTGTCTGGCAGCTCCAAGTAGAGTCGTTAAAAATTTAA
CGATGTACTAAATAACTTCAAGCTATTTGTCTCGCTGGATTGGCTTCGAGTGGTAAGATC
CTATCAAATGCCGAAAACAAAAAACTTTCTTCTTAATTGTTCGTTCTTCAACACCTCTCC
ATGGTGATAACGGATACGGTTTCATTGTCAGCATCCACAATCATAATAAGTACAATCTTC
CGAAAAATACATTACGCCTTGAAATATGCAATCGCAAACACGGATTCGTTTGGAACATTT
ATTTTCCTATGAAGAGATGCGATAGGTAATATTTATTTGAGCGTTTAAGATACTCATTGT
ACTCTCAAAGAAAGTCATTGAAAGCCAACGAGGTCAAATCAAATATTATAATAAAAAGGT
CAAAGAGGACTAACTTAAAGCTCTCTTCATGGATATTCAAATTCAAATTGTTTTCGCCCA
TCGCTAGAACTTTTACCGTTTCCATTGAGTATATAACTAAGATGAATGAGGCTAATTGAT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTT

AbU6-727 (AALF029727)

ATTGATGAACTTACAGTAAAAATTTGGGAATATTTGATGCATTTTTCGAAAAGTTACAGC
TAGTTGAACATTTTTTTGAAAAGTGAAAATTTTGCCTGTCCCGATCATTTTGTCTATCCC
CTGTACATGTATAACGATGTACTAAATAGCTTCGGGAGATATGTCTTCGCTTTTACATCT

TAGCGACATCAAATTCCAAATAATCAATTTGAATGTTGTGTTCTACGAACACTTCTCCAA
CAATAACGTTCCAGATGATAACGAAACCAATCGTTATTTGGCTTTTTGTCGACTAATTCA
GCTTCCATACATAAGAATTACCACCGTCTGCCTGTGATCAGCACTTCACCTGAAGTTAGG
CAACATCGGGCATCGAACGATTTGAAATTAGAAGGAAACGCACATTGGATTGGTTGACCT
TTTGATAGGTAAATTTTCCACCAACTCTTAATATTTGAGGTACCTTTCGCATGAAAAAAG
TTAAATCGAGCAAAAGCAAAATAACTTTACCTGATAAGCCTTTCAATCAAGATTCGTCCA
TCGCTAGAACCTAACCGGTATACTATAAATATATAAGAGGGGAAGCAACGGCAATGAAAT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTT

AbU6-744 (AALF029744)

AATCCTTTGTTTTCTTACTCAGTTTCCGTCAAAAGTGGGACNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGTAAGAAAACAAAGGATTTTTCG
ATTTATGAGTAAGGCCAACTCAGTCATTGAGTAGAAATTTTTCTCGGTGTTGATTTGCAG
AATTCCTATTGGAATATTTTGGGTGCACCCATGTCATTCAAATTCTTACAGCCAACACAC
GCGCACGCGGATGTTCACTAACAGGTATACGTTTATCGCTTCATGCATTCATTCTCGTCG
AGTACCCGGACTGTTTGCCTGGACAGCAGTGTTCGGTAAGTGGCATATCGCCGTCTATTC
ATCGCTAGCACCATTTTGATGGCTGCCTGTTATATATACTGGGTTTTTCGTTCTCCGCTC
GTCCTAGCTTCGGCTGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTGT

AbU6-725 (AALF029725)

CGTCTAGTCTGATGGAATTGGACTCGTCATTTGCATTTCACTATTGCTTGCATGTTCTAC
AAAGCATGCAGCAGACAGACGACTGAGGTTCGAATATGCAGAGTCACAAAGTAAAGGCAT
AGAAGAGTAAAAAGTTGACATTCCATCCAAAAGACTCTAATTATTCTTCGCTTGACTGCG
GGAAAACAGTCGACAGCTCAGCTCGAGTGTGGTCTTAGCTCAGAGTTGTGTCGAGTTCGA
CTCGACCATACGGAGCTAGAACAAATTGAATATTTAAGTGTCCGTTATAGGAATGGTGTT
AGATGTTCGATGAGTCATCCTTCGGAGTAGCACGAACAGTATACTATAATTGAAGATTGT
CGGCTGATAAGTGTGAAGCTCTATAGGTATACATATCAGCAATCAGTAGGGACGATTGAC
TTCTTCGAGGTAGACAACGAATTCAATGTTTTTATTTTGTATGGGTATGTATATGTCTCG
CATCAAGCCGTTCGGGGGTGCGAACGAACGAAAACGTTTATTCAGTTTGATTGTCTATCCA
TCGCTAGAACCGTAGTCGTCGCAGCAGAATATATAAGAGTGGAAGCACCGGCAATGAAAT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTGATA

AbU6-726 (AALF029726)

GCTGAAAGTTGGCTATCGGATAATTTTACCTTTTTCAAGAATCTTTATAACTTCGTAAAA
AACGGTCCGATCTTTTCTAAATTTGGACCACTGATACACAATGAGTTGATGAACTTACAG
TAAAAATTTGATTGCACTTTTCAAAAAGTTACAGCTAGTTGAACTTTTTTTTTGAAAAGT
GAAAATTTTGCCTGTCCCGATCATTTTGTCTATCCCCTGTATTCAATGACTCTAATTGGA
GCGTCCAGACAGAATTGATTTGTACGTGTGCAGTGTATGGCTACACTGGGTAGGTACATT
TTTGTCGCTTTTTGTCTCCGTCGCCGCCGGCAGACAGAGACAGTCCAAAACTAGTTGCCA
GGACGATCCAACTGTGCTGCTGGTTGGCTGGAATCCAATTTGAATGTTTGCTAATTGGAA
AAGATGGAAGTTTTTTTTTCTTGAATGGATGATTTAATAATTGAAACGATTCCAGGTATC
TGATATAGATATGCGAACATATGTAGATTACTTACGAACGTGAGATGAAGATTCTACGAA
TCGCTAGAATCAAATCAACTGGTGGTTGATATATAAACGATGTGGGTCGACCCGATGCTT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTTTTATTGA

AbU6-955 (AALF029955)

GATGGAATTGGACTCGTCATTTGCATTTCACTATAGGTACTGCTTGCGAGCTCTACAAAG
CATGCAGCAGACAGACGACTCCGGATCGAATATGCAGAGTCACAAAGTAAAGGCATAGAG
GAGTAAAAAGTTGGCATTCCATCCAAAAGACTCTAATTATTCTTCGCTTGACTGCGGGAA
AACAGTCAACAGCTCAGCTCGAGTGTGGTCTTAGCTCAGAGTTGTGTCGAGTTCGACTCG
ACCATACGGAGCTAGAACAAATTGAATATTTAAGTGTCCGTTATAGGAATGGTGTTAGAT
GTTCGATGAGTCATCCTTCGGAGTAGCACGAACAGTATACTATAATTGAAGATTGTCGGT
TGATAAGTGTCTAAGCTCTAGGTATACATATCAGCAATCAGTAGGAACGATTGACTTCTT
CGAGGTAGACAACGAATTCAATATTTTCATTTTGTATAGGTTAGGTATGTATATGTCTCG
CATCAAGCCGTTCGGGGTGCGAACGAACGAAAACGTTTATTCAGTTTGATTGTCTATCCA
TCGCTAGAACCGTAGTCGTCGCAGGAGAATATATAAGAGCGGAAGCAACGGCAATGAAAT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTGGTAG

AbU6-956 (AALF029956)

TTCAATAATGATAGGTAAATTTTTCATCTACTCAAAATATGTGAGGTACCTTTCGAAAGA
GAAATAAAAATAAATCGAGCAAAAGCAAAACAGTTTTACCCAAGGTTGCAACCATTCATT
TGCAAAGATTTACATTCATTTGCTATTATCTCAGTTCAGAAGCATGCTATCGAAAAACAA
TGTATGGACGAATTTAACCTTGTGGTTTCATCTGAAAGTTTGCCTATTAACATTGGGGTC
GCACACGCATTCCAAAGTCGTGAGCGAGCTGTGAAGGCAACTTTCGACAGCGTGAATGAA
ATTTACATTCACCGCGTGGAGAGTTGCCTTCACAGCTCGCTCACTACTTTGGTATACGTT
TGCGACCCCAATGTTATTCGGCAAACTTTCAGATAAAACTACAAGGTTAAATTCATCCAT
ACATTGTTTTTCGATAGCATGCTTCTGACCTGAGATAACAGCAAATGAATGTAAATCTTT
GCAAATGAATGGTCGCAACCTTGGTTTTACCTGATAAGCCTTTCAATCAAGATTCGTCCA
TCGCTAGAACCTAACCGGTATACTATGAATATATAAGAAGGGAAGCGACAGCAATGAAAT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTTACAT

AbU6-957 (AALF029957)

CTAACTTCTGGCGTAATGCTGTTAATAAGCAGACAGCAATAATCCGAATGTTGAATTAAT
CGACAAAAAACCAAATAACGGTCCATCTGAAACGTTATCGTCGTGTACGTAGAACACAAA
ATTCAATTGTATTTTTTTTAATTTGATGTCGCTAAGATGCAAAAGCGGAGATACATCCCG
TGAAGCTATTTAGTACATCGTTATATATGTATTTTATGACTCTAATTGGAGCTGCCAGAC
AGAATTGATTTGTACGTGCAGTGTATGTACGGTGTACGGTTAGGTACATTTTTGTCGCTT
TTTGTCTCCATCGCCGCCGGCAGGCAGAGACAGTCTGAAAACATGGAAAACTAGTTGCCA
GGACGATCCATCTGTGCCGTTGGTTGGCTGGAATCCAATTTGAATGTTTGCTAATTGGAA
GAGATGGAAGTTTTTTTTTCTAGAATGGATGATTTAATAATTGAAACGATTCCAGGTATC
TGATATAGATATGCGAACATATGTAGATTACTTACGAACGTGAGATGAAGATTCTACGAA
TCGCTAGAATCAAATCAACTGGTGGTTGATATATAAACGATGTGGGTCGACCCGATACTT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTTTATTGAGA

CqU6-653 (CPIJ039653)

TGAGCAGAAGCAGGCATCTTATCTCGCGAGATTCTCATTGCCAGACGACGACGACGACTT
GCAAGAAGAAAAATTGAAACAAATTGTGTGGGATTAAATGATTTGAGCCGTTGCTTGAGG
GGAGTTGCGATGGAGGGCGGGGGTGATCCCCCATTACTGACGCCTCGTTTTTACTCCAGA
CAACGAGGGGAGGGGAGTAAAACCGCTTACGGTTATTTAGGAGTTGTTGAAGCACTGCGA
GTCGTAATGGGTATAATTGAAAATGCCAGACAAATTTGATTTGCTATGCATTTTGTTGCG
GAGCAGACGCGTCGACTGAGATTAATCTGCATATTTTGAGAGTCGCATTTAGTCACAAAC
GGCAATTCTGATTCAATTAGCGAAATCTTGTTTTAAATTCAATCGCAGGTACATAACCTA
ACCTCGAACTGACGGGTAATCGATTACTAAACGTTTACTTGCCCCCCATTGCTGTGAACC
TTACAAATGTTCTCACTTTTTCGCGTTTCTTTCGCTCACTTCGCACTCCGCACTCCGCCA
TCGCTAGGACCGTTTAACGATTCACGAACTATATAAGCAACGAGGGCACCCTCCGAACTT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTTT

CqU6-543 (CPIJ039543)

TTTCTTTTTCATGACCCTTGCCTTGACCATCTCTTGAGGATTTTTTTAAATGGATTTATC
GCTTTCATTCTTTTGGAGCCTTTAAACTTAAAACGAGTTTTTTTTAATCAAATACTTTCT
GAATTAGTTTTTGCGCGGTTTCGGATTTCTGGTCGGCGCGGATTTGGCGCGGATTTTTTT
TTCGACTCTCCCGTAACAACCCTGGAGTTCGTGTCGGAGATTTCGATGGATTGATCGGGT
TAAGCTATAGAGAGAGCTATTCAATCGGTGTTTGCGTTTCCGTATTTGCATTTCGGAGAT
TGTCATCGATAGGATTGTGGGCTTTCGGGGGAAGGAATTTTGATTATATTGGAGAGTACT
CATTGTTGCATTGTTTTTTTTGGGACGTGGACGAATTTTGTTTGCCTAACTCAATTCTAG
GATTGAACTGTTTATAAATATTCAAGTAGGATATGCGGGGCATCAAAACTGATGATACCA
CCACTCAAAGCAAACACAATATCCAATAAGATTGGTTTGATTGTCCACGTAGCTTACGTA
TCGCTAGAACTGTTTTGTCAGCCACGAACTATATAAGCAACGAGGGCACCCTTGGAACTT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAAAAGATTAGCATGGC
CCCTGCACAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTT

CqU6-596 (CPIJ039596)

AAGATAATAAGGTCTACTGTTACATTTAAGAAATATTCTTATTTAATTTGAGCTGTGAGC
AAAACCGTAAAATTAGTGCGTCCATCCGGGGAATGCCCGGGACAAAATTGCCGGGATTTT
TTCAGATTTTTTATATTTTATCCAAGGATTTCCCGAAATTGATTAAAAAAATCAAATCTC
TTTAAAAATTAGAAACAACCTGTTAAAAATAAGCTTATAGAAAAAAAAATAAAGTGTTGC
TACAAGGACAATCATCATAACGAAAAACTTTAAGGTTCTTGAAAAGCACGTCCCAAGATG
GGAAAGTGTGTTTCATTTCCAGGCGATCCTTTTCGCTTCACGAAATTTTTGATTGCTACC
TCGTTTAGAGCCCTAAGACAAAATACTTCGTCAAAAAAAGCGAACTCGAACCCAAACTTT
GTACTCCTATGCTAAAATGTCTGCATTCATGGGATGTCGCGCAAACGAGTCCAATAAGAA
ATCAAACTGCTCGAATTCAACGTGACGCTACCCGGATTGCTATACCCATAACTTTATACA
TCGCTAGAACCGCGTTCGCGCTCGCTTACTATATAAGCACTTTTCCGGCCCTACAAACTT
GTCTTAGCTTCGGCTAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTT

CqU6-728 (CPIJ039728)

ATCTCGCGAGATAAGATGCCTGCTTCTGCTCATCAGCATGATCATCATCACTGGCGAGGA
AGGAGGCTCATTTCCTTAAACTCGTGGGAGGACGACACTGCAATTTAAAGGCGAGAGGAG
GTTATGTTGCTTTTTTTCCGGGCCCTTGCCTTAATGGGTTCTGAATCCCTCTGTCTGTGG
TTGTTGTTGTTGTTCTTGGGGGTCGATGTCGTCGAGCCCACTCCTTGCCCGGCCTGTTTG
TGTGTATGTTTATGTGATAAAATATTAATGATGCAACCCCGGTGGCAGACACTGCTGGCG
AATGTTGCGGACCAGTTAAGGTTCTGGTAGACAGGGTGACGAGTGCTCCGGTTGACATTG
CTGCATTTTAGGGGTGGTTCAGAGTTCTGGAAAAAAAAAGTTTTATAAATAAAATTCCAG
TTGGCATGAAATCGACGTTGCAAATTGTTTTGAGACAGATCTAATAGGACTTCTTTTTGA
TATTGCAGGTAATTTACCTGTTATTACCTGTGCATTGAACTTGGAAACAAATATTAAATA
TCGCTAGAACTGAATTGAAGTGCTATAACTATATAAGCAAAATTTGAGCCCACGGTACTC
GTCCTAGCTTCGGCTGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTT

CqU6-801 (CPIJ039801)

GACTTTTGAAGTTTATCTTGATTTGTTCCAAAATTATTTTGAATATCATAATATTTTTGT
TTAATTTTATCCATCAAAAAGATTTTTTCTCTTTCTTTAAACGACGCTGCTTGATTTTTT
TTAATAAATTACACATCATTGATTTTGAAAATTAATTTTGGTTTAATGTATTGGTTTTTT
GCTTTACAGAAATTAATTTTAAAAATATTGTTATTCACAAAACAATTTCTTTTAAATATA
GAAGTGATCTACCTTCTAATATTTGATCATCATTTTATTGGTAGCTATTCAAATTGGATG
AAGTCTTTATGAGGCAATTCAATAAGGTTGAATACTTTCTTTAGGCTTATGATACGCAAT
GCCAAAGATTCACAAAAAATAAATGTTCAATCAATGTTGCCTTAAACTGCAGCGAGTTAT
GTCTATAACGAAATAAATCCTTTTACATGTTTTTTTTCTAGTCTTATAAGAATGTAATCA
GTCAGTCGTTCTAAAAATAAGTTTTACCTGGATTAATTTCTAAGCTAATTTTCTAACTCA
TCGCAAGAACTGTTTTGAAATGTGCTAACTATATAAGCAAATATGTAGACAACTTAACTT
GTCCTAGCTTCGGCTGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTT

*AgU6-557 (AGAP013557)

CGTAGAGACATTTTGTACCAGAATATTTAAGTAAAATAAACATTATTTTTACCGGAATGT
ATACATAATTACGAAAAATGTTTTATTTTTCAAAAAAAATGTTCAAATTTCTGCCATAAT
GGTATTTGCTTGAAATCTAAACCCGAGCACACACTTTCTTGCTCCGTGCTCTCACCCTTT
GCCGGCTCTTTCTTCTCTTTCTCACTGCGATACGAACCCTACGGTACGCGACGTTCGGCT
GGGAATGAGTTCCAGCATTAGATGGTGAGTGCTTGTTGGCCACATATACACTGAAGCGGT
TTGGCTCTTGGTTATTCTAGAAAACTCTGTTTTTGAATTAATCATAGGAAAACTGTTTAT
TCTTTTTTGATATTTCAACGTATTACAATGAATAGTAACTTTTGATTAAAAAAGAGATGA
AGCAATACAAGAAATGTTGAAAAGTTTATGAAACATCACATTAGCGTGAGTTACGGCAGG
ATCAAAACCTTATAACAGTCACACTCAGGCAAAAAATCCTTCTTGAATATCCTTTATGCA
TCGCTAGAGCAAGGATTGAAAGCGCAAAGTATATATACAACCTTTTTTCCCCTCGTCCTT
GTCCTTGCTTCGGCAGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTGTC

*AgU6-695 (AGAP013695)

GAAATAGCAAATAGAACTTTATGTATGTCCTTCAATTCTATTTTGTACACGCATTATCTT
GCTCCACAACCGCAGAACTTTCCGGCAGTGATGGACAAGGCCAAAAATCCCACTCGTCAT
CAACAGCGAACGACGACGGTTCCAAAGTTCGGCAAAACTATTAACAAAAAACAAACAAAC
CAATTGGCTGGCTTATTAATGTGTAGTAATGGGAGATAGAAATCCATTAGTCAGTTTTCC
ATCCATTTCGTTTTGCTTTGCGCTATTGAAGCAATAATAATCGTATGAAATTAAAATGAC
AACCGTGGTAATTTGCTGATTGAGAATGTTTACTCACCGCAAATGCCACCCCACATATAG
CGAAGGATATATGGACGTTAGTCCTGCCACCGATGCCACCGACCTTTTGTGTTTTTCATC
GTGCAGGTACACACAACTGTGCTATCTTTCAGCCCTTTTGTATGCGTGCGCTTGAAGGGT
TGATCGGAACCTTACAACAGTTGTAGCTATACGGCTGCGTGTGGCTTCTAACGTTATCCA
TCGCTAGAAGTGAAACGAGCGTGCGTAGGTATATATATGAAATGGAGTTGCTCTCTGCTT
GTCCTAGCTTCGGCTGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTGACA

AalbU6-132 (AALB015132)

GTAGGCAAACCAACCACTCTCACGATCAAGATGAACACTTTCCACTACTAAAGATGAACA
CTATCCGATGGTTTATTTTACATGCGTCCGGCTCTCGGCCGCAGATTTCTGTGCTGTCTG
GAGCTGTTTGTGAGGTAGTCCATCGGTTCCCAGATGGCTGCAAAAGGATTCACAATCCTT
TTTCATAGTTTCAGCTTCATTGTATTGTGGTTCTAGACGTCCCGGTTCTGGAATTCGCAT
CTGTAGTTAGGTCCTCCGTAATTTTATCGTTTTTCTTGCTTCTTGCTTCTAGGAAGCGTA
CACGACGATCTTTCGTGTGGCATTTGTTCGTTGTCCTCCTGAGCAGAACTGAACGCTAAG
GCGGTTGGGTATCGTGGGTTTCTGCTAACTTCTGGGGCACCGAGTTAAGAATGGGTGACA
GCTGTTCATATGGCGCTGAGTGTAGCGCAGAACTCTACAACGTCACTTAAGGTTCTGCGT
TGAAACAGTTCCGAATCTTTCGCTAGATGGCAGTGTATCGCAGAGGACGAATATTATGTA
TCACCAGAATAAAAATGAAATGAGTGTAGGTATATATAATGTGTAGAGGCCGCTCCGCTT
GTCCTAGCTTCGGCTGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTTT

AaraU6-171 (AARA015171)

AGAGTGTGTGAAATAGCAAATAGAACTTTATGTATGTCCTTCAATTCTATTTTGTACACG
CATTATCTTGCTCCACAACCGCAGAACTTTCCGGCAGTGATGGACAAGGCCAAAAATCCC
AACTCGTCATCAACAGCGAACGACGACGGTTCCAAAGTTCGGCAAAACTATTAACAAAAA
AACAAACAAACCAATTGGCTTATTAATGTGTAGTAATGGGAGATAGAAATCCATTAGTCA
GTTTCCATCCATTTCGTTTTGCTTTGCGCTATTGAAGCAATAATAATCGTATGAAATTAA
AATGACAACCGTGGTAATTTGCTGATTGAGAATGTTTACTCACCGCAAATGCCACCCCAC
ATATACCGAAGGATATATGGACGTTAGTCCTGCCACCGATAACCGTTCGTGTTTTTCATC
GTGCAGGTACACACGACTGTGGTACCTTTTAGCCCTTTTGTATGCGTGCGCTTGAAGGGT
TTATCGGAACCTTACAACAGCTGTAGCTGTACGGTTTCGTGTGGCTTCTAACGTTATCCA
TCGCTAGAAGTGAAACGAGCGTGCGTAGGTATATATATGAAATGGAGTTGCTCTCTGCTT
GTCCTAGCTTCGGCTGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTT

AaraU6-449 (AARA015449)

CAAAGACACATTTTGTACCAGAATATTTAAGTAAAATAAACATTATTTTTACCGGAATGT
ATACATAATTATGTAAAAATGTTTTATTTTTCAAAAAAATGTTCAAATATCTGCCATAAT
GGTATTTGCTTGAAATACAAACCGCAGCACACACTTTCTTGCTGCGAGCTCTCACCTTTT
GCCGGCTCTTTCTTCTCTTTCTCACTGCGATACGAACCCTACGGTACGCGACGTTCGGCT
GGGAATGAGTTCCAGCATTAGATGGTGAGTGCTTGTTGGCCACATATACACTGCTGCAGT
TTGGCTCTTGGTTATTCTAGAAAACTCTGTTTTTGAATTAATTATATGAAAACTGTTTAA
AATTTTGAGATATTTCAACGTGTTACAATGAATAGTAACTTTTGATTAAAAAAGAGATGA
AGCAATACAAAAAATGTTGAAAAGTTGATGAAACATCACATTAGCGTGAGTTACGGCAGG
AGCAAAACCTTACAACAGTCACACTCAGGCAAGGAAACCATCCTGAATATCCTTTATGCA
TCGCTAGAGCACGGATTGAAAACGCAAAGTATATATACAACCTTTTTTCCCCTCGTCCTT
GTCCTTGCTTCGGCAGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTT

AfunU6-704 (AFUN015704)

TTATTATAATAAAAAAATCTTTACATTAAGTGTTTGGGATGAAATTATGTGAAACATTAT
TAAAATAATACGATTTTTAATATTATTTTTAAATTTGTTTGCCGAGTGGAATTGCTTGAA
ATCGTAACGCTCGTTTAAGTTTTCTTTTACAACGAGCGCTCAGCTCTGTTCGCTCTTTCC
GCTCTCTTTATCTTCTCTCTTGGATCGACCGCCGAACGCATTCGTGTTGCGATGAGTTCC
AGCATGCTCTCTTCTGAGTGCTCTTAGCACGTGTATACACGTGCTAATGCTTTATCTGTT
TGTTGAAACATATAACCAAAAATATGCGCATATATGCCGGTGTAATTTATTTTTACTTAA
ACTATTAGCATAAAGCTTTTCAAACTTATTTGACATTTAAAAAAGAGAATATAAATAATA
AAGACGAAAATATGTACAATGATAAGTTAAATTCCAATGGTGAGAAAAAGGTTTCGCTTC
AATGAAGCCTTACAACAGTTGCATAGGTTATCCAGCAGCTTTGTGAATATCCTTTATCCA
TCGCTAGAACAAGGATTCAAAACGGAATGTATATATATTGACTGAGTGGCACACATCCTT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTT

AfunU6-538 (AFUN015538)

TATGTCCTTGAAATCTATTTGTATTCATCTGCGTTGTCGACTCTATCAACGGTTTGCACC
GTCAGCCAGGAATACATTTCGATTCGTCATCAATGAGGGATGTGTCGTGAATTCATAGTC
TTAGAGTCTAACTTACCCACAAACTAGTTTCCATGTTTGGCAGAACCGTTTCAGTTATGC
TACACAAATATTCTGAGCCATGCGATCAAAACTGTTCATTCTACTACGCACATCGCGCTA
AAATAGAAATCCATTACCATTTCCTTCGCCGGCTAGTCAGAAAACCGAATAATCAACAAT
ACAATTAAAATGACAGCCATGGTAATTTGTTGATTGAGAATGTTTACGCATCACAAATGC
TTTAACTAAAGCCCTAAAGCGATGGTGCCCTTTGTTAAAAAAGAAGCCTGAAATACGACC
TATCAAACAATCGATTTTCGCTTTGCTAACACGCTACACAAAACAGTGATGCGTGCGCTC
GAAGGGTTTAACGGAACCTTACAACGTTGTAAGGTTCCGTATGAAGCACAACACTATCCA
TCGCTAGAAGTAAAACGAGCGAACGTTTGTATATATACAAACGGGATGGTGTCTTCACTT
GTCCTAGCTTCGGCTGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTT

AsteiU6-842 (ASTEI11842)

TACCAACCCATTTGCCCCCTCCCTTGTCTTTGCTTCGGCAAGACATATACTAAAATTGGA
ACGATACAGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAAATCGTGAAGCG
TTCCACATTTTTTTGTGGAAATTTGATTCACTTGTTTTAGGAAAATAATTACCTTCCTTT
GAAACAGGTATTTACAATGATGGACGATAGAAAAAGACCACCTTAAAATGCTTTATTCAA
GCTATTGCTTTCGATTCTTCATTGAGATCTAATCTCAAATTTGTGTATTTAGAAAAAAAA
CAGCGAATATTGTACTATAACAATTCTTTGGTAATCTCAGGCTTCCAGTAATGATAGAAT
CCTTAAGACTGTAAGAATATACAAGTTTTAAAAAAAAGGAAAACAGGTATATCAAATCAG
AAAACAACAACAGCTCTTTTCCGTTGAATTCTACCTGTACGATGGCTTAAGGATGAGCTG
CAACTGAAACGGAACCTTACAACAGCCACGCAGAACGAGCAAGGACTTTGTTTTATCCAT
CGCTAGAACTAAAACGAAGGGTCAGATTGTATATATACCAACCTGTTTGCCCCCTTCCTT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTTT

AsteiU6-917 (ASTEI11917)

CTTCTTAGACAGTGTTGGAGGAGAAGTTTGTATAAACATAATGTCAATAAGGGTAATTTT
TACTATTATTTTAAACATTTTGTGTCGCGTGTTGATTTGCTTGAAAGGGAAACGCCTGTT
TGTATTTGCTTGCTGCGAGCGCTCAGCTTCGCTTGCTCTTGCTCTCCTCTCTTTCTCATC
TCTTCTAACGCATGGACTCAACTGCCGACGAGTTGCTCGGGTTTCGATGAGTTCCAGCAT
ACGGGCATGCTGCTGAGTGCTTCTTGGTCGTGTATACACTCATGCCCTTTTTTCCTCGTG
AGAAATTTTTTTCTACTCAATGCAAATTTCAAATTTGATGGCAATTTTCGTGCTTATGTT
TAATATTCTATAAAAGAAAATTTAGAACTAAAAACTCATGTACAGCAGTAGATGAAGCTT
TAACCGTTCTTCTCACTTAAATTCTACCTGCACAATGGCTGGTGGAAGAGCTATAATTGA
AGCAGAACCTTATAACAGTCACGCGGAAGGATCAAGAGCTTTGTGAATATCCTTTATCCA
TCGCTAGAACTAAAACGGATGACAGACGGGATATATACCAACCCATTTGCCCCCTCCCTT
GTCTTTGCTTCGGCAAGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTTTT

AsteiU6-858 (ASTEI11858)

ATGTATTTGTGTATTCATCTGCGTCGCCGATTTCCGCGGGAATTCGTCATCGGCGCTGAG
TGACATTTCGGGAAACGGTTAAAATAGCCAAAAGGGGGGGGTTTTCCAGGTTTTTCGAAC
GGAAGAAAACGGAACAGTTATGAGCTATACTCGCATAAAAAAAACACACGATGCGTCGAA
ATAGAAATCCATTAACGTTTCTACGCCCTGCTCCCCACCAGGCAGGCACAGGCAGGCAAG
GTCTATTCCGAAAGGAACACGAGTAATAATCAACAATACAATTAAATCAGACAGCCATGG
TAATTTGTTGATTGAGCATGTTTACGCATCACAAAAAGAAACGCTCAAAACCCCAAAGCA
CCAACACCAACAACTGTAACAACGGTTTTATTGTGCGTTTGTTCAACCAGCACAATCGAT
TTTCTCGCCTCGCTAACACATGCGGACGCATGACTGCGGTGTGATGCGTGCGCTTGAAGG
GTTAAGACGGAACCTTACAAAGTCAGCTGCTTGTCGTGTGCGATGGTACGGAGTTATGCA
TCGCTAGAAGTAAAACGGACGAGCGTCGGTATATATACAGTGCGCGATTGCTCCTTACTT
GTCCTAGCTTCGGCTGGACATATACTAAAATTGGAACGATACAGAGAAGATTAGCATGGC
CCCTGCGCAAGGATGACACGCAAAATCGTGAAGCGTTCCACATTTTT

CLUSTAL multiple sequence alignment (by MUSCLE) of a selection of U6 promoters active in mosquitoes


```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


Ae_U6-905       ----------------------------CCTATCGCATCTCTTCATAGTAAAATAAATGT
Ae_U6-774       -----------------------------------TTCCCGTAGACAAAAATCAAATTT
Ae_U6-702       -------------------GTCATTGACTTGACTACTTTTATTGAACT------AATAAT
Ag_U6-695       -GAAATAGCAAATAGAACTTTATGTATGTCCTTCAATTCTATTTTGTACACG--CATTAT
Ae_U6-763       GGCAACTATAGAGTTTCCATGTTTCCAGACTTTCCCTCCCGGTAAACGGAGACAAAACGA
Cq_U6-728       ------------ATCTCGCGAGATAAGATGCCTGCTTCTGCTCA---------------
Cq_U6-596       ----------------AAGATAATAAGGTCTACTGTTACATTTAAGA------AATATT
Cq_U6-801       ---GACTTTTGAAGTTTATCTTGATTTGTTCCAAAATTATTTTGAATATCAT--AATATT
                                                          *

Ae_U6-905       TC----------------------CAAACAGATCCGT-------------GTTTGCGATT
Ae_U6-774       CATGGTATTTTTCGGCGA------TAAACAGTAAGTTAAAGGGAACCCGAATTTAGTGCT
Ae_U6-702       TCTA--------------------TTAACAGAAATTGATAACAACAGTTTGCTTATGAAA
Ag_U6-695       CTTGCTCCACAACCGCAGAACTTTCCGGCAGTGATGGACAAGGCCAAAAATCCCACTCGT
Ae_U6-763       CAGACG-----------------TAAGTAGGTACATATGCATACCGC--ACGGACAAAT
Cq_U6-728       ------------------------TCAGCATGATCAT----------------CATCACT
Cq_U6-596       CTTATTTAATTTGAGCTG------TGAGCAAAACCGTAAAATTAGTGC--GTCCATCCGG
Cq_U6-801       TTTGTTTAATTTTATCCA------TCAAAAAGATTTT-------------TTCTCTTTCT
                                        *

Ae_U6-905       GCA----------TATTTCAAGGCGTAA------TGTATTTTTCGGAGGATG--------
Ae_U6-774       ATA----------TAATTTAATTCCACTAGAGTTTGTATCCTTTGATAGATA-------C
Ae_U6-702       TAACATGATAGCCAAGTCAATAACAGGTCCAGTTTAGACTTTCGGG--------------
Ag_U6-695       CAT----------CAACAGCGAACGACGACGGTTCCAAAGTTCGGCAAAACT--------
Ae_U6-763       CAA----------ATTTGTCTGGCAGCTCCAATTAGAGTCGTTAAAAATTTAACG-----
Cq_U6-728       ---------------------GGCGAGGAAGGAGGCTCATTTCCTTAAACTCGTGGGAGG
Cq_U6-596       GGA----------ATGCCCGGGACAAAATTGCCGGGATTTTTTCAGATTTTT--------
Cq_U6-801       TTA------------------AACGACGCTGCTTGATTTTTTTTAATAAATTACACATCA
                                     *           *

Ae_U6-905       --GATGCTGACAATGAAACCGTATCCGTTATCACCATGGAGAGGTGTTGAA---------
Ae_U6-774       GCGTATTTCGACCTCAACTGCAAGGCCG-------------TCGTGTACTA---------
Ae_U6-702       ----TTTTCCATAGTAAATATACTACAA----ATAATATTAATGTTCCATG---------
Ag_U6-695       ----ATTAACAAAAAACAAACAAACCAATTGGCTGGCTTATTAATGTGTAG---------
Ae_U6-763       ----ATGCGCTAAATAACTTCAAGCTAT-------------TTGTCTCGCT---------
Cq_U6-728       ACGACACTGCAATTTAAAGGCGAGAGGAGGTTATGTTGCTTTTTTTCCGGGCCCTTGCCT
Cq_U6-596       --------------TATATTTTATCCAA-------------GGATTTCCCG---------
Cq_U6-801       TTGATTTTGAAAATTAATTTTGGTTTAATGTATTGGTTTTTTGCTTTACAG---------
                          *                             *

Ae_U6-905       GAACGAACAATTAAGAAGAAAGTTTTTTGTTTTCGGCA-TTTGATAGGATCTTACCACTC
Ae_U6-774       GACTTGACTAATCCAGACTGGTCTTTTAGTTATGACT--TCTGTCCACATCTCCATACAT
Ae_U6-702       AAAAAAGGAGTAAGAGTCTGG------------------------TAACCCTAGTGCAC
Ag_U6-695       TAATGGGAGATAGAAATCCATTAGTCAGTTTTCCATCCATTTCGTTTTGCTTTGCGCTAT
Ae_U6-763       GGATTGGCTTCGAGTGGTAAGATC-------------------------CTATCAAATGC
Cq_U6-728       TAATGGGTTCTGAATCCCTCTGTCTGTGGTTGTTGTTG--------TTGTTCTTGGGGGT
Cq_U6-596       AAATTGATTAAAAAAATCAAATC------------------------TCTTTAAAAAT
Cq_U6-801       AAATTAATTTTAAAAATATTGTTATTCACAAAACAATTT----------CTTTTAAATAT


Ae_U6-905       GAAGCCAATCCAGCG--AGACAAATAGCT-TGAAGTTATTTAGCGCATCGTTAAATTTTT
Ae_U6-774       TCAACGCACTGTGC----------------GGCTGTGCTGTGCGACTCCGTCGAGTCG--
Ae_U6-702       GCAAAT-ATCTCGCGGGCATATTTGGTTGCTGAGGTATATTTATATTTGAACGCCATGAG
Ag_U6-695       TGAAGCAATAAT-------------------------AATCGTATGAAATTAAAATGAC
Ae_U6-763       CGAAAACAAAAAACT--TTCTTCTTAATT-GTTCGTTCTTCAACACCTCTCCATGGTGAT
Cq_U6-728       CGATGTCGTCGAGCC--CACTCCTTGCCCGGCCTGTTTGTGTGTATGTTTATGTGATAAA
Cq_U6-596       TAGAAACAACCTGTTAAAA-----------ATAAGCTTATAGAAAAAAAAATAAAGTGTT
Cq_U6-801       AGAAGTGATCTACCTTCTAATATTTGATC-ATCATTTTATTGGTAGCTATTCAAATTG--


Ae_U6-905       A----ACGACTCTAATTG----------------GAG--------------------CTG
Ae_U6-774       -----ACCAACATAGTTG----------------AAA--------------------CAA
Ae_U6-702       AAAAAGCGGAAGAAATTGGCTCATGGCCGATTTTAAG--------------------GAT
Ag_U6-695       A----ACCGTGGTAATTTG------------------------------------CTG
Ae_U6-763       A----ACGGATACGGTTTCATTGTCAGCATCCATCCT--------------------CCG
Cq_U6-728       ATATTAATGATGCAACCCCGGTGGCAG-------ACA--------------------CTG
```

```
Cq_U6-596       GCT--ACAAGGACAATCATCATAACGAAAAACTTTAAGGTTCTTGAAAAGCACGTCCCAA
Cq_U6-801       --------GATGAAGTCTTTATGAGGCAATTCAATAA----------------------G


Ae_U6-905       CCAGACAAATTT------GATTTGTCCGTGCGGTATGCATATGTACCTA-----------
Ae_U6-774       ATTG--AATATT------TAATTGATCGTTATAGGAATGGTGTTAGATGAGTCATCCTTT
Ae_U6-702       ATTTAAAAATTG------TA--------CAATGTACATATAATTAAACA----------T
Ag_U6-695       ATTGAGAATGTT------TA------CTCACCGCAAATGCCACCCCACA-----------
Ae_U6-763       A-----AAAATA------CATTACGCCTTGAAATATGCAATCGCAAACA-----------
Cq_U6-728       CTGGCGAATGTTGCGGACCAGTTAAGGTTCTGGTAGACAGGGTGACGAG-TGCTCCGGTT
Cq_U6-596       GATGGGAAAGTG------TGTTTCATTTCCAGGCGATCCTTTTCGCTTC-----------
Cq_U6-801       GTTGAATACTT-------TCTTTAGGCTTATGATACGCAATGCCAAA-------------
                      *  *


Ae_U6-905       ------CTTACGTCTGTCGTTTTGTCTCCG------TTTACCGGGAGGGAAAGTCTGG--
Ae_U6-774       ACAGTAAGCACATACAGTATTATAATTGAA------GATCGTCGGCAGATAGGTGTGTAG
Ae_U6-702       CCGTTCCTTCAATGTGTTCTTTTTTTTAAG-----CGTGTGTTAAAAGTTTGCTCTGGTG
Ag_U6-695       --------------------TATAGCGAAG------GATATATGGACGTTAGTCCTGCCA
Ae_U6-763       ---------------CGGATCTGTTTGGAACATTTATTTTACTATGAAGAGATGCGATA
Cq_U6-728       GACATTGCTGCATTTTAGGGGTGGTTCAGA------GTTCTGGAAAAAAAAAAGTTTTATA
Cq_U6-596       ACGAAATTTTTGATTGCTACCTCGTTTAGA------GCCCTAAGACAAAATACTTCGTCA
Cq_U6-801       -----------------------------------GATTCACAAAAAATAAATGT-TCA


Ae_U6-905       --------------------AAACATGGAAACTCTATAGTTGCCAGGTAGACCATCTGCC
Ae_U6-774       GGTAGAGTATCAGC---AATAAGTTGGGACGTTTGACTTTTTGTAGGTAGACAAAAACTA
Ae_U6-702       GTGAA------------TTCACGCTCTACCCGTTCAGGCAGCATTCATCGAAAAGCCCTA
Ag_U6-695       C---------CGATGCCACCGACCTTTTGTGTTTTTCATCGTGCAGGTACACACAACTGT
Ae_U6-763       GGTAATATTTA------TTTGAGCGTTTAAGATACTCATTGTTCTCTCAAAGAA------
Cq_U6-728       AATAAAATTCCAGTTGGCATGAAATC-GACGTTGCAAATTGTTTTGAGACAGA-------
Cq_U6-596       AAAAAAGCGAACTCGAACCCAAACTTTGTACTCCTATGCT------------AAAATGTC
Cq_U6-801       ATCAATGTTGC------CTTAAACTGCAGCGAGTTATGTCTAT------AACGAAATAAA


Ae_U6-905       TCCGTCGGCTGGCTGGATTCCAATTTGAATATTGGCTAATTGGAAGAGATGGAAGTTTTT
Ae_U6-774       AACTTTTTTTCGCT---------TCTCTATGTGTGCCCCTGGGTAGCGTTCCGTTCCGAT
Ae_U6-702       TCTGCTCGCACACA---------TT------------------------TACAAAATGCT
Ag_U6-695       GCTATCTTTCAGCC---------CTTTTGTA------------------TGCGTGCGCTT
Ae_U6-763       ----------------------TGTCATTGAAAGCCAACGAGGTCAAATCAAATATTAT
Cq_U6-728       ---TCTAATAGGAC---------TTCTT---------------------TTTGATATTGC
Cq_U6-596       TGCATTCATGGGAT------------------------GTCGCGCAAACGAGTCCAAT
Cq_U6-801       TCCTTTTACATGTT---------TTTTT---------------------TCTAGTCTTAT


Ae_U6-905       GAATGGATGATTGAATAATTGAAGCGACTCCGGGTACCTGTTTGTAAGCTCTGCAACAGT
Ae_U6-774       TGGGGTGCGAACGAA----------------------------TGAAATCGCCCATCGAG
Ae_U6-702       GATTGCGTTGTGTGCTGAATGGGTCACTCGTCCGTCACTGCTTGCTGTGTACACTGTACA
Ag_U6-695       GAAGGGTTGATCGGA-------ACCTTACAACAGTTGTAGCTATACGGCTG--CGTGTGG
Ae_U6-763       AATAAAAAGGTCAAA----GAGGACTAACTTAAAGCTCTCTTTATGGATAG------GAA
Cq_U6-728       AGGTAATTTACCTGT--------------------TATTACCTGTGCATTGAACTTGGAA
Cq_U6-596       AAGAAATCAAACTGC-------TCGAATTCAACGTGACGCTACCCGGATTG----CTATA
Cq_U6-801       AAGAATGTAATCAGTCAGTCGTTCTAAAAATAAGTTTTACCT---GGATTAATTTCTAAG


Ae_U6-905       GCCATAGATTCGTGTCAGTCCATCACTAGAATCAAATCAACTTGTACTTGATATATAAAT
Ae_U6-774       TTGATACGTCC------ATCCATCGCTAGAACCGCGTTCGCTGTAGAAGACTATATAAGA
Ae_U6-702       GTTACGCAGTCT-----GTGCATCGCTAGAATCATATTTACGGAAAAGTATTATATATAC
Ag_U6-695       CTTCTAACGTT------ATCCATCGCTAGAAGTGAAACGAGCGTGCGTAGGTATATATAT
Ae_U6-763       AAAATATTTTC------GCCCATCGCTAGAACTTTTACCGTTTCCATTGAGTATATAACT
Cq_U6-728       ACAAATATT--------AAATATCGCTAGAACTGAATTGAAGTGCTATAACTATATAAGC
Cq_U6-596       CCCATAACTTT------ATACATCGCTAGAACCGCGTTCGCGCTCGCTTACTATATAAGC
Cq_U6-801       CTAATTTTCTA------ACTCATCGCAAGAACTGTTTTGAAATGTGCTAACTATATAAGC
                                *** * ****                    ******


Ae_U6-905       G-----------------GCTTGGGTTAATGTGCTTCGGCAAGACATATACTAAAATTGG
Ae_U6-774       GCAGAGGCAAGAGTAGTGAAATGTCTTT----GCTTCGGCAAGACATATACTAAAATTGG
Ae_U6-702       CCAAT-GCGTTGCTCATCGGTTGTCCTA----GCTTCGGCTGGACATATACTAAAATTGG
Ag_U6-695       GAAATGGAGTTGCTCTCTGCTTGTCCTA----GCTTCGGCTGGACATATACTAAAATTGG
Ae_U6-763       AAGATGAATGAGGCTAATTGATGTCTTT----GCTTCGGCAAGACATATACTAAAATTGG
```

```
Cq_U6-728       AAAATTTGAGCCCACGGTACTCGTCCTA----GCTTCGGCTGGACATATACTAAAATTGG
Cq_U6-596       ACTTTTCCGGCCCTACAAACTTGTCTTA----GCTTCGGCTAGACATATACTAAAATTGG
Cq_U6-801       AAATATGTAGACAACTTAACTTGTCCTA----GCTTCGGCTGGACATATACTAAAATTGG
                                 *   *     ********  ******************

Ae_U6-905       AACGATATAGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAAATCGTGAAGC
Ae_U6-774       AACGATACAGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAAATCGTGAAGC
Ae_U6-702       AACGATACAGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAAATCGTGAAGC
Ag_U6-695       AACGATACAGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAAATCGTGAAGC
Ae_U6-763       AACGATACAGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAAATCGTGAAGC
Cq_U6-728       AACGATACAGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAAATCGTGAAGC
Cq_U6-596       AACGATACAGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAAATCGTGAAGC
Cq_U6-801       AACGATACAGAGAAGATTAGCATGGCCCCTGCGCAAGGATGACACGCAAAATCGTGAAGC
                ******* ************************************************

Ae_U6-905       GTTCCACATTTTTTT---
Ae_U6-774       GTTCCACATTTTTTT---
Ae_U6-702       GTTCCACATTTTTTT---
Ag_U6-695       GTTCCACATTTTTTGACA
Ae_U6-763       GTTCCACATTTTTT----
Cq_U6-728       GTTCCACATTTTTTT---
Cq_U6-596       GTTCCACATTTTTTT---
Cq_U6-801       GTTCCACATTTTTTTT--
                **************
```

7SK gene promoter sequences

Each putative 7SK promoter sequence is presented as up to 600bp 5' of the given Accession number. **7SK gene and terminator sequence are indicated**.

>AAEL018514_7SK

TTATGGGAAACCCGAAACAGATTTTATTTTATGCTCCATTCTCCGCCACTTGTTGATGCGGACCCTAACCAC
GTGGTCGCTCCTCTGCTCACCGGAGCACGTTTCATACAGCCTGACGACGACGAGCAATCAGAGGTATGGTGA
GCATGCGCATGGAGAGTGGACAGCAGTGCACCCTAAAATCAATTCACACATCATGTGTCAATAGCTGTGTCA
ATGTTGCACAGCCTTTTCTTATTAAATTTACTCCTTTTGTGACCATTTCTCTTTCATCCACCGTTATTTTAA
TGAGTTTTGTGTTCCGGTGGACGAACGTTCACACAAAAAATGTGTAAATCTTAATCAACCAGAACACAAAGT
ATAGTGAAAAAATTAAAGTGTGTGGCTTTTATACATCCTAACTGTAAATTATTTTTAGAGTGCGTGCGATCG
TTCTCTCGAACCACGCTCTCCGCTACACATTCGCAGCGAATGGCGTGAATGGATGAAAGAACAAACTAAAGT
TTATTTTTAGATTCGTCTCAAAACAACTGCTGTGCATCGCTAGAACCAAGAAATACGCCACTCAGTATATAT
AGCACTTCCAACCCCGCTTTCCTC**GGAGGTGTGTGTCTTCGTCTGTGATGGCACGATAACTGAACATTGATC
GCTTTACGTGTTAGTTTGCAGATCTGCTCAGTGGCAACCCGTCACACCTAAAATACAACTCTGGCAGTCCGG
ATCTGGTATCACGGGTGAACTCTCGCTGCACGGCGCCGGGCCGAACGCACGATTGATGTCATTTGTGATACA
AGACTACTGCCGTTCTTACCCAACTCTTTCCAAATTGTTGAGTATAAAATCGTAATTTAATACAGATAGCTT
AGCTTCGGATTAAAATTACATTGTTCAGAACGCTTCCATATCACTAGGGCACCGCCGAGCGGTCGGCCCATT
CTTTTG**

>AALB015206_7SK

CGATTGTGAGACCTTCCAATCCATGTTCATGTCCTCGATAGTATAGTGGTCAGTATCCCCGCCTGTCACGCG
GGAGACCGGGGTTCGATTCCCCGTCGGGGAGGGTTGAGTTTTTTAAATTTTCAATAAGTACTCTGAACATTA
CCTATTTTGCTTCATTTATGTTAAGTTAATAACTAAGTCGCTGTCAAGAAGGTGGCCATTGATGGAAAAGTG
AAAAGCTGGTGACACACGAAGCGTAAACTGAAATGTAATGCCAAAAAGTTTAGGCCGCAAAAAGCTTGCCGT
TTACACGCTGTAAAAATGACTATAAGTCCGCGGAGCGTCAAAGTAAGTTTCTATACTATTTATTTCTAACAA
AAGTTAATGAATGTTCAGTACAATCAATTTTACGCGTCTCAACATGACCAATTAAACACAGCCAAGCGTAAA
ATTCTTGACATTTATTTTATATTTATGTTATAGTTTATGCTTCGTCTGTCGCCAGCTAAAAATGTTGAGAAC
AAGTTTATGCGTATGCGCAGCACTGCCGTTAGCAATCGCCAGAACAAACATGACAGCTGGTAGGGTATATAG
GAACATGTGCCATCACTGAGCTTT**GGAGGTGTGTGTCAAATCAGTATGTGATGGCACGATAACTGAACATTG
ATCGCCAAAAACAGTTTAGTTTGCAGATCTGTCCAGTGGCATGCGTCACACTTCTAATGGTAGTCTTTCTTC
TGTGTCATCGGTGATCTCTCGTTGCACAGCGACGGGCCGTACGCACGATTGATGTCATCTGTGACACAAGAT
TCTACCGCATTACGAAATTAGTTGAAGTTGTAATTATACAAGTTAGCTTAGCTTCGGATTAAAATTACATTG
TTCAGAACGCTTCCATATAACTCGGGCACTGCCGAGCAGTTGGCCCATTCTTTTT**

>AALF029648_7SK

TAGCATAGTCAGTCAGTGAATAAGCACCCAAGCCAAGCGACCATCCACCGAACAACTGAGGAGGGAGATTGT
TACAAGCACAGCTGATTCGATCTCGCTCTCGGTACCGCATGGCTTGCGAGGCAAGAGCATCAAGCTACGAGC
AGGCAAAACAACACCCCTCATAAACCGTAAACATACCCTTACGCTTACCCCATCCATTCCCTACCGTAGCGA
CCAGCTCATTATGGGAAACCCGAAACAGATTTTATGTTATGCTTCTTTCTCTGCCACTTGTTGATGCCGCAT
CAAGAGACCACGCGGTCGCTCGACCACGTTCCATGCGGCACAGGTAGAGGAGGTAGGCAATCGAGTGCGCGC
CAGAGGCATGGTGAGCATGCGCACAGAGAGACGACCGCACCATGTGAGGCGATCGTTCTCTCCGTTCTGAAA
GCTCTCCTCGCTCCGTCGTTTTGAATTAATTTGTATGAGTAAAGGTAGGCAAAAGTTATTTTTAGCCACTCG
ACTCGAGACGTTGAATTCATAGCAACTGCCATCCATCGCTAAAACCGAAATTTTCGCAGTCTACTATATATA
CGACTTCCACCACCGGATATCTTC**GGAGGTGTGTGTCTTCGTCTGTGATGGCACGATAACTGAACATTGATC
GCTTTACGTGTTAGTTTGCAGATCTGCTCAGTGGCAACCCGTCACACCTTGATACAATCGTCTGGCAGTCCG
GATCTGGTATCACGGGTGAACTCTCGCTGCACGGCGCCGGGCCGAACGCACGATTGATGTCATTTGTGATAC
AAGACTACTGCCGTTCTTACCCAACTCTTTCCAAATTGTTGAGTATAAAAATCGTAATTTAATACAGATAGC
TTAGCTTCGGATTAAAATTACATTGTTCAGAACGCTTCCATATCACTAGGGCACCGCCGAGCGGTCGGCCCA
TTCTTTTG**

>AARA015292_7SK

GTTTTAACTCCCTCCCCCTCCCCTAACGCGACGATTAAATCGAAACGGAGAATAAAAACACAGCCCGAGTCG
CACTCGGCCAAACAGCTGAACAGGTTAATATTAAGATTCGAATTCCTAAACGGCAAAATAAGGCTGAACAAC

GGCTTTAGAGTCTCTTATTACTCTGTCTGACTATATAGAGCTGTTAACATTGTTCGTTCCTAATTCAACTGA
AAAATTCGCAAAATCACCAATTTAACAGCATCTTTTGTAGGACGTTTTTAACCAGAAAGCACATTGTTACAG
ACAGAGCGAAAGAAAATACACCATAGCGGTTGCTCTCACATTCTCTCTCCTTCTTTAAACGTGCGGCCTTTT
GCATCCCCTCTCTAACGCAGTCGACTGTTATAAGGTTCTGCCGCAGACCGTTTGGAAGGGATGCTAAAAATA
GAACACAAAAGCGACGAAGGGAAAAGTGCGTAGCATGAGAGCATGCGCACTCGCAGCATCGGTGTTTGGTGT
GTGCGTGAATGAGATGGAAGACCATTTTTATACATCGCTAGAACTCGGTTGAAGTTAGCGTGGTATATAATA
GCAAACAGCATGCAGAGGTTACTC**GGAGGTGTGTGTCTTCGTATGTGATGGCACGATAACTGAACATTGATC**
GCCAAAAACAGTTTAGTTTGCAGATCTGTCCAGTGGCATGCGTCACAACTCTAATGGTAGTCTTTCTTCTGT
GTCATCGGTGATCTCTCGCTGCACGGCGACGGGCCGTACGCACGATTGATGTCATCTGTGACACGAGATTCT
ACCGCCATACGAAATTAGTTGAAATTGTAATTATACAAGTTAGCTTAGCTTCGGATTAAAATTACATTGTTC
AGAACGCTTCCATATCACTCGGGCACTGCCGAGCAGTTGGCCCATTCTTTTT

>AFUN015339_7SK

TTAATTTTCCCTAGCTCATACACTTTTTCTTCATTATCATTGAAAGTTATGTTCTATAGAAATACCCTTCCA
AATACCATATTTTTACAATATTTTTTACTAAAAAATACAAAATTTTCTTATAATATAAGACATACTTAATTT
CGTTTTTGCACAGTTTATGTGATATATAATGAGAATTATTTTATTCTACTCAAGTATCAACCCAATAAAGAG
TTTATTTTTCAGGTGCACATTTTTTAGAAATGTAATCCAAATCGCTTTATTCAAAACAATCGTTGTAACAAT
GCTCCGAAACATAACCAGATTGTTGGTAAAATAAACGAGTAAAACAATACACGCGCCCATTCATAAGAAATA
CGCATTTGCAGCAAATGTTTCCTCGAACACTGTTGTAGGGTTCCGTTGTATGGTACCGATGCTATAAATAGA
ACACCAATCCAAACGACGAAAGACCGCACGTGACGTGTACGCATGCGCGAGCATAGCATAGATGCGTAGCAA
GCGTATGAACGAGATGGAAGTTATGCTTTATGCATCGCTAGAAGTCGGTTGTGTTCGCAGTGGTATATAATA
GCAAACAGCTTCCCTAGGTATCTT**GGAGGTGTGTGTCTTCGTATGTGATGGCACGATAACTGAACATTGATC**
GCCAAAAACAGCTTAGTTTGCAGATCTGTCCAGTGGCATGCGTCACAACTCTAATGGTAGTCTTTCTGCTGT
GTCATCGGTGATCTCTCGCTGCACGGCGACGGGCCGTACGCACGATTGATGTCATCTGTGACACGAGATTCT
ACCGCCATACGAAATTAGTTGAAGTTGTAATTATACAAGTTAGCTTAGCTTCGGATTAAAATTACATTGTTC
AGAACGCTTCCATATAACTCGGGCACTGCCGAGCAGTTGGCCCATTCTTTTG

>AGAP028235_7SK

GTTTTAACTCCTTCCCCCTCCCCTCACGCGACGATTAAATCGAAACGGAGAACAAAAACACAGCCCGAGTCG
CACTCGGCCAAACAGCTGAACGGGTTAATATTAAGATTCGAATTCCTAAACGGCAAAATAAGGCTGAACAAC
GGCTTTAGAGTCTCTTATTACTCTGTCTGACTATATAGAGCTGTTAACATTGTTCGTTCCTAATTCAACTGA
AAAATTCGCAAAATCACCAATTTAACAGCATCTTTTGTAGGACGTTTTTAACCAGAAAGCACATTGTTACAG
ACAGAGCGAAAGAAAATACACCATAGCTGTTGCTCTCTCATTCTCTCTCCTTCTTTAAACGTGCGGCCTTTT
GCATCCCCTCTCTAACGCAGTCGACTGTTATAAGGTTCTGCCGCAGACCGTTTGGAAGGGATGCTAAAAATA
GAACACAAAAGTGACGAAGGGAAAAGTGCGTAGCATGAGAGCATGCGCACTCGCAGCATTGGTGTTTGGTGT
GTGCGTGAATGAGATGGAAGACTATTTTTATACATCGCTAGAACTCGGTTGAAGTTAGCGTGGTATATAATA
GCAAACAGCATACAGAGGTTTCTC**GGAGGTGTGTGTCTTCGTATGTGATGGCACGATAACTGAACATTGATC**
GCCAAAAACAGTTTAGTTTGCAGATCTGTCCAGTGGCATGCGTCACAACTCTAATGGTAGTCTTTCTTCTGT
GTCATCGGTGATCTCTCGCTGCACGGCGACGGGCCGTACGCACGATTGATGTCATCTGTGACACGAGATTCT
ACCGCCATACGAAATTAGTTGAAATTGTAATTATACAAGTTAGCTTAGCTTCGGATTAAAATTACATTGTTC
AGAACGCTTCCATATCACTCGGGCACTGCCGAGCAGTTGGCCCATTCTTTTT

>ASTEI12173_7SK

GTTCTAGCAGTTAATAGATAGCTTACTCTTAACATGGTAACATGGTATAAACGCAATGCTTAACCTTTTTAA
TTAATTTGCAACGATATGGCGATGCTCCTTTCAAGAGTTGTTACCCTTACCAGGATAATTGAATCAGAAAAC
AAAAAAGTCTGCATTTCAATCAATGAATGATTTCATTCATCAGCAACAAAACAATCTCCATCACTTTCTTCA
TTGAAGATTTTACAAGAATCAACTGTGCGTACCATTATTCCTAAACAAAGAACGTTTCCCAAACAAAATAAT
GGATGAGGCGCCAATCGCTGTGAAGAATCCATTCATAACCGCATGCAATGAGAGACTACGCACCGAGCATCG
ACTCTCCGCGACTGTTGTAAGGTTCCGGCGTTCGGTGGAAGCGAGCGGAGCTAAAAATAGAACTCAACGGCA
GGCAGCAGAAAAGCGATGCGTCGCGTGAACGCATGCGCGACATCGCAGCAGCATAGCCGCGTAGCATAAGTG
CGACCGAGAGAGAAGGGGTAGCAAATGCTTATCCATCGCTAGAGCTAGGTTGGGCTGGATGCGGTATATATA
GCGAACGGTGGCGCACCGTTCCCT**GGAGGTGTGTGTCTTCGTATGTGATGGCACGATAACTGAACATTGATC**
GCCAAAAACAGCTTAGTTTGCAGATCTGTCCAGTGGCATGCGTCACAACTCTAATGGTAGTCTTTCTGCTGT
GTCATCGGTGATCTCTCGCTGCACGGCGACGGGCCGTACGCACGATTGATGTCATCTGTGACACGAGATTCT
ACCGCCATACGAAATTAGTTGAAGTTGTAATTATACAAGTTAGCTTAGCTTCGGATTAAAATTACATTGTTC
AGAACGCTTCCGTATGACTCGGGCACTGCCGAGCAGTTGGCCCATTCTTTTG

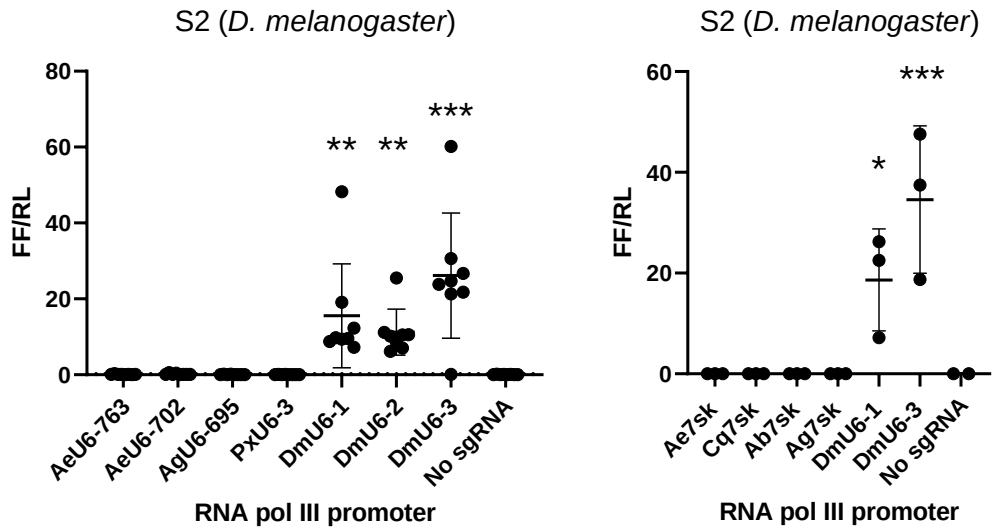
>ASTEI015331_7SK

AGACCAGACGCAGGCATCTCCATTATAAAAATAATATATAGTTCTAGCAGTTAATAGATAGCTTACTCTTAA
AATGGTAACATGGTGTAAAACGATTGCTTATCCTTTTTTATTAATTAGCAACGATATGGCGATGCTCCTTTC
AAGAGTTGTTGCCCTTACCAGGATAATTGAATCAGAAAACAAACAAAAAATATGCATTTCAATCAATGAATG
ATTTCATTCATCAGCAACAAAACAATCTCCATCACTTTCTTCATTGAAGATTTTACAAGAATCAACTGTGCG
TACCAGTATTCCTAAACAAAGAATGCTTTCCCGAACAAAAGAATGGATGAGAGCAATGAGAGACTACGCATC
GAGCATCGACTCTCCGCGACTGTTGTAAGGTTCCGGCGTTCGGCGAGCGGAGCTAAAAATAGATCTCAATGG
CAGGCAGCAGAAAAGCGATGCGTCGCGTGAACGCATGCGCGACATCGCAGCAGCATAGCCGCGTAGCATAAG
AGCGACCGAGGGAGGGGGTAGCAAATGCTTATCCATCGCTAGAGCTAGGTTGGGCTGGATGCGGTATATATA
GCGAACGGCGGCACACCGTTCCCT**GGAGGTGTGTGTCTTCGTATGTGATGGCACGATAACTGAACATTGATC**
GCCAAAAACAGCTTAGTTTGCAGATCTGTCCAGTGGCATGCGTCACAACTCTAATGGTAGTCTTTCTGCTGT
GTCATCGGTGATCTCTCGCTGCACGGCGACGGGCCGTACGCACGATTGATGTCATCTGTGACACGAGATTCT
ACCGCCATACGAAATTAGTTGAAGTTGTAATTATACAAGTTAGCTTAGCTTCGGATTAAAATTACATTGTTC
AGAACGCTTCCGTATGACTCGGGCACTGCCGAGCAGTTGGCCCATTCTTTTG

>CPIJ039933_7SK

AAGATGAAGAGCATTTAGCCATACCACTTAGAATGTAAATGAAATGTAATTATTATAAGAAAGTTCAATAAA
GACATATTTAATTTCAAAAAAAAATTTCATACATCCTCTCTCACGTTTGCTTCTATATCTACCCGCTAAGGT
AACAAGCGGAACATAAACACCACTGAGAGTCTGTTCCCTTCCCTGGCTACATGACGGTATTGCACAGCCAGC
TGACTCTCCTTTCAATCCTCCTTTTGAAGAGAGTCTCCTCATCCACTTGTTGCTTTGATTCAATACCGCTAG
CAAAAAGGTAGTCTGACACACGACTCCGAACAGACAAAAGAGAACAAGAGAGAGAAATTCTCATTCGAGAGA
CGGAGAGAGCAGGCAGTCTCTCTGTTTACAAAACAAGATTGAACATGTTCAAAGGGGAGAATACGATTCTCA
TTGAAACTATACCCAGCAGCTTGTTACGAACGGGTAGAAAACGAAGTGACTGCATTTGTAAGGCTCCACTAC
TGAAAAGAGAGCCAAAGCACCCTCGTTTTCATCCATCGCTAGAACTGCGTTGCTCGCCGCGCACTATATATA
CACGTTCCACAACCTCGGTTCTTC**GGAGGTGTGTGTCTTCGTCTGTGATGGCACGATAACTGAACATTGATC**
GCTTTGTTAGTTTGCAGATCTGCTCAGTGGCAACCCGTCACACTCTTGATAAACGGCAGTCCGGATCTGGTA
TCACGGGTGAACTCTCGCTGCACGGCGCCGGGCCGAACGCACGATTGATGTCATTCGTGATACAAGACGCTG
CCCAGACCCAACTATTTCTCAAAATTGTTGAGTATATCGTAATTTAATACAGATAGCTTAGCTTCGGATTAA
AATTACATTGTTCAGAACGCTTCCATATCACTAGGGCACCGCCGAGCGGTCGGCCCATTATTTTG

Positive control tests of RNA pol III promoters inactive in mosquito cell lines

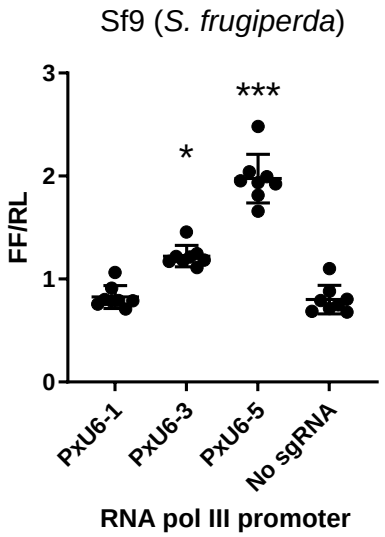


Appendix Figure 4: CRISPRa dual luciferase assay in *D. melanogaster* cell line S2. Each graph represents a different experiment and shows results as FF/RL on independent y-axis scales. A selection of RNA pol III promoters were tested in cell line S2 to confirm activity from D. melanogaster promoters not active in mosquito cell lines. Mosquito RNA pol III promoters are shown to not have activity significantly different from background. Kruskal-Wallis analysis was used to determine significant difference in results between groups and Dunn's multiple comparison was used to determine which groups are significantly different from background ("No sgRNA"). Where present, significance is denoted with asterisks: "*" $P < 0.05$; "**" $P < 0.01$; "***" $P < 0.001$. N = 2 – 8.

D. melanogaster cell line S2

Appendix Figure *4* demonstrates that *D. melanogaster* promoters are active, acting as a positive control for Figure 24 where no activity is seen from DmU6-3 in *An. gambiae* cell line Sua5.1. Appendix Figure *4* furthermore demonstrates a reciprocal relationship where mosquito RNA pol III promoters are not active in *D. melanogaster* cells and vice versa.

S. frugiperda cell line Sf9



Appendix Figure 5: CRISPRa dual luciferase assay in *S. frugiperda* cell line Sf9. *P. xylostella* RNA pol III promoters were tested independently (each plasmid uses one promoter with otherwise identical TetO_sgRNA2 cassettes) in a moth cell line, Sf9. Activity is shown as FF/RL on the y-axis with promoter identity on the x axis. Results (N = 8) are shown with mean and SD. A Kruskal-Wallis test was used to determine significant difference between groups and Dunn's multiple comparison was used to determine which groups are significantly different from background ("No sgRNA"). Where present, significance is denoted with asterisks, using the same key as Figure 14.

Appendix Figure *5* demonstrates that the *P. xylostella* promoters are active, which was not seen in mosquito cell lines. Although this is a small dataset, the rank order of promoter activity corroborates the findings of (Huang et al., 2017) in a *P. xylostella* cell line. It is notable that *P. xylostella* and *S. frugiperda* are not closely related moth species, presenting an opportunity for cross-species use of RNA pol III promoters, for work in Lepidopterans without previously described RNA pol III promoters.

Table of U6 promoter activity in species of interest

Since the completion of the work discussed in Chapter 4 and published as Anderson et al. (2020) there have been further publications in the realm of U6 promoters, particularly in relation to the development of CRISPR gene drive systems. Appendix Table 19 sets out the state of the field for reported activity of U6 promoters in various species. The manner of assay (RNAi, CRISPR endonuclease or dCas9-VPR (CRISPRa)) is indicated, as is the cellular context of the experiment – cell line (species indicated) or whole insect. The table is grouped alphabetically by species, then by assay type.

The rank orders stated are judged by eye from the figures of the referenced paper; statistical analysis is typically limited to 'presence' or 'absence' of activity. Zero is indicated to distinguish promoters that are not different from background. U6 promoter identities are denoted by species ("Ae" *Ae. aegypti*; "Ag" *An. gambiae*; "As" *An. stephensi*; "Cq" *C. quinquefasciatus*; "Dm" *D. melanogaster*; "Px" *P. xylostella*; "Bm" *B. mori*; "Sf" *S. frugiperda*) and by the last three digits of U6 gene accession number or by local identifier in the case of lepidopteran U6 promoters (Px, Bm and Sf).

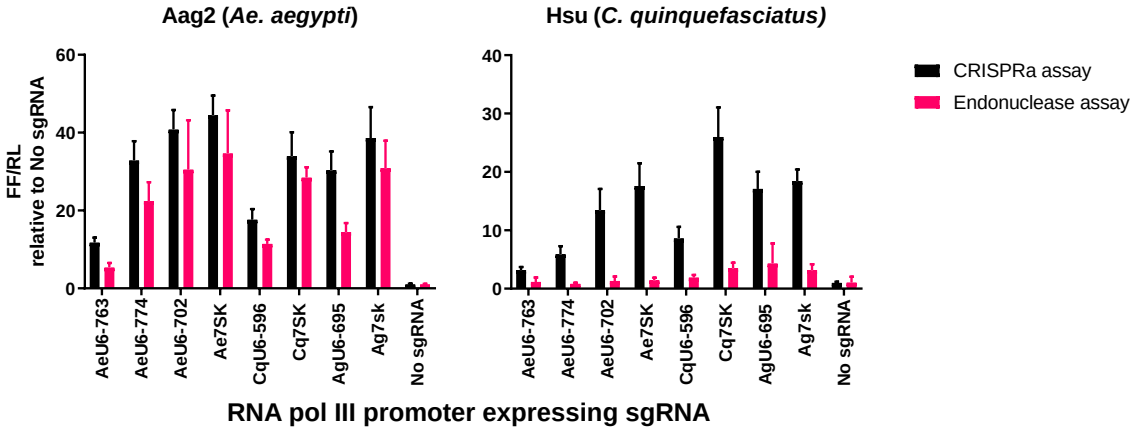
Appendix Table 19: Literature review of activity of U6 promoters in species of interest.

Ref	Species	Assay	Context	Rank order
(Li et al., 2020)	*Ae. aegypti*	CRISPR	whole insect	0 = Ae-578 = Ae-574 < Ae-763 <= Ae-774 < Ae-702 < Ae-905
(Anderson et al., 2020)	*Ae. aegypti*	dCas9VPR	Ae cell line	0 < Ag-557 < Ae-763 = Ae-905 = Ae-000 < Ag-695 <Ae-774 = Ae-972 < Ae-702 = Ae-848
(Konet et al., 2007)	*Ae. aegypti*	RNAi	Ae cell line	0 = Ag-557 < Ae-774 < Ae-702 < Ag-695
Figure 24	*An. gambiae*	dCas9VPR	An cell line	0 = Ae-702 < Ag-557 < Ag-695
(Konet et al., 2007)	*An. gambiae*	RNAi	Ag cell line	0 = Ae-774 < Ae-702 < Ag-557 < Ag-695
(Gantz et al., 2015)	*An. stephensi*	CRISPR	whole insect	0 < As-697 (from AsteS1.8)
(Feng et al., 2021)	*C. quinquefasciatus*	CRISPR	Cq cell line	0 = Cq-693 = Cq-801 < Cq-728 < Cq-653 < Cq-596
(Feng et al., 2021)	*C. quinquefasciatus*	CRISPR	Embryo injections	0 = Cq-801 < Cq-728 < Cq-596 < Cq-653

Ref	Species	Assay	Context	Rank order
(Anderson et al., 2020)	*C. quinquefasciatus*	dCas9VPR	Cq cell line	0 = Cq-728 = Cq-801 < Cq-653 < Cq-543 < Cq-596
(Port et al., 2014)	*D. melanogaster*	CRISPR	whole insect	0 < Dm-2 < Dm-1 < Dm-3
Appendix Figure *4*	*D. melanogaster*	dCas9VPR	Dm cell line	0 < Dm-2 = Dm-1 < Dm-3
(Wakiyama et al., 2005)	*D. melanogaster*	RNAi	Dm cell line	0 < Dm-2 = Dm-3
Appendix Figure *5*	*P. xylostella*	dCas9VPR	Sf cell line	0 = Px-1 < Px-3 < Px-5
(Huang et al., 2017)	*P. xylostella*	RNAi	Px cell line	0 = Bm1 < Px1 = Px5 < Px3
(Huang et al., 2017)	*P. xylostella*	RNAi	whole insect	0 < Px-3
(Mabashi-Asazuma and Jarvis, 2017)	*S. frugiperda*	CRISPR	Sf cell line	0 = Dm-3 = Bm-1 < Sf-A

Appendix E: Supplemental Information - Chapter 5

Validation of endonuclease assay



Appendix Figure 6: Graphs showing RNA pol III promoter activity in cell lines Aag2 (left graph) and Hsu (right graph), for two CRISPR based assays. Data is shown as relative activity with the negative control set to 1 in each instance. Results from each assay type are shown grouped by cell line (by graph) and then as interleaved coloured bars. The CRISPRa assay (black) is a transcriptional activation assay based on dCas9-VPR, whereas the endonuclease assay (pink) uses a modified firefly luciferase reporter plasmid to report activity of Cas9.

The data in Appendix Figure 6 is presented to validate the activity of the endonuclease assay. RNA pol III promoters described in Chapter 4 are shown in a side-by-side comparison for two assay types (CRISPRa and endonuclease) in two mosquito cell lines (Aag2 and Hsu). Data from cell line Aag2 shows that there is a reduced activity (relative to background) in the endonuclease assay as compared with the CRISPRa assay; this is expected, based on the design of each assay. In spite of this difference in magnitude of expression between the two assays, the rank order and relative expression between RNA pol III promoters appears to be conserved. In cell line Hsu there is very little activity from the Endonuclease assay, for all of the promoters tested. The endonuclease assay was not used for further experiments in cell line Hsu.

Relative activity of a panel of sgRNA variants – reproduced from Noble et al. (2019)

This work to validate sgRNA variant sequences in Chapter 5 was designed to closely mimic results published by Noble et al. (2019). The experiments are not a direct repeat of the reported work, but there was no expectation of gross differences between results achieved by each group.

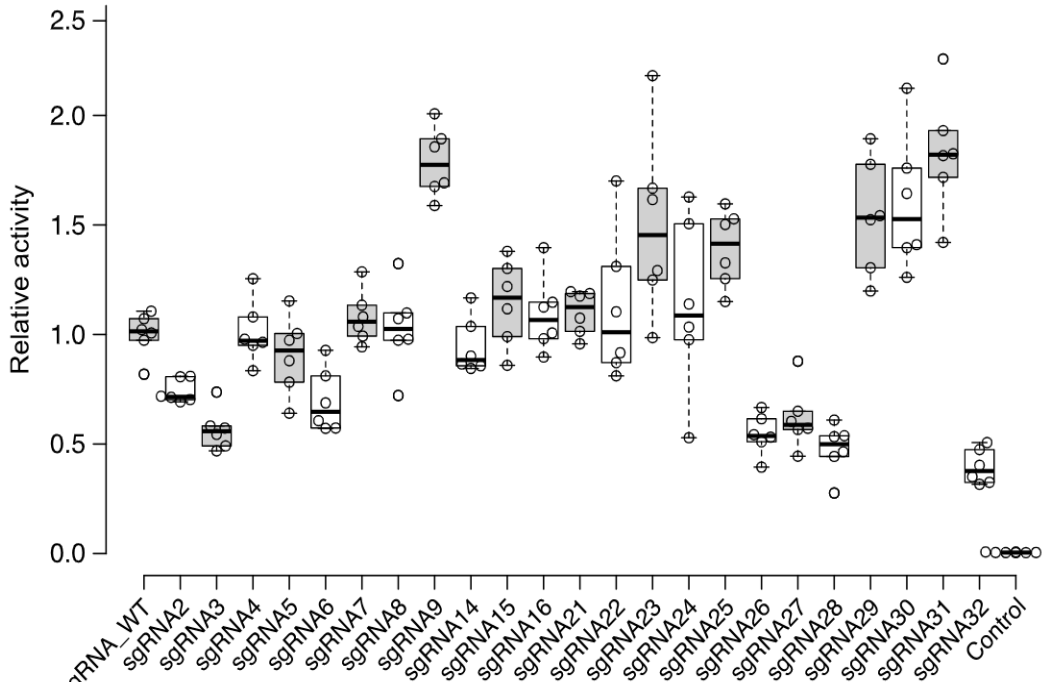


Figure 45: Graph of results showing relative activity of a panel of sgRNA variants. This image is reproduced from Noble et al. (2019), who show activity relative to the standard sgRNA sequence ("sgRNA_WT") on the y-axis. Each sgRNA variant is identified on the x-axis and data is reported as individual repeats (circles) with overlayed box and whisker plots. A negative control ("control") is shown at the far right of the x-axis.

The data shown in Figure 45 was generated "by using a Cas9 transcriptional activator screen using a (fluorescent) reporter in human cells" (Noble et al., 2019). It shows sgRNA_WT as a positive control at the far left of the x axis and a negative control group at the far right; results have been transformed to "sgRNA_WT" as 1. No statistical analysis is presented, but data is shown as individual repeats with box and whisker plots to indicate spread. There appears to be little background activity in the negative control group, if any, and many sgRNA variants show activity that overlaps with the spread of results for the positive control, or exceeds it.

Appendix Table 20: Primer sequences used to generate sgRNA variants

sgRNA F		sgRNA R		Target	Backbone Variant
LA986	GAAATTAATACGACTCACTATAGGACTTTTCTCTATCACTGATAGTTTTAGAGCTAGAAA	LA988	AAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC	TetO_2	sgRNA_WT
LA2180	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTCGAGAGGACG	LA2181	AAAAGCACCGACTCGGTGCCAGGTCTCCCTGTGATAACGGACTGGATTAAAATCACTGCCTTATTCGAACTTGGACTCTCGTCCTCTCGAAC	TetO_2	sgRNA_03
LA2162	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTCCAGAGTCGG	LA2163	AAAAGCACCGAATCGGTGCCTGCCTTCCGGCATGATAACGGACTGGTATATAATACACTGCCTTATTCCAACTTGTCGTTCCCGACTCTGGAAC	TetO_2	sgRNA_04
LA2178	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTGTAGAGCGTA	LA2179	AAAAGCACCTACTCGGTGCCAGCGTTTCCGCTTGATAACGGACTGGAATTTAATTCACTGCCTTATTGTAACTTGCGTATTTCTACGCTCTACAAC	TetO_2	sgRNA_05
LA2176	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTGCAGAGACAC	LA2177	AAAAGCACCGAATCGGTGCCGACGTTCCCACGTCTGATAACGGACTGGTTTAATAAACACTGCCTTATTGCAACTTGACACTCCCGTGTCTCTGCAAC	TetO_2	sgRNA_06
LA2164	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTGGAGAGGCAT	LA2165	AAAAGCACCGACTCGGTGCCCCTAGTCTCCTAGGTGTGTACGGACTAGCCTTATTGGAACTTGGCATTCTCATGCCTCTCCAAC	TetO_2	sgRNA_07

sgRNA F		sgRNA R		Target	Backbone Variant
LA2166	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTCTTAGAGTGTG	LA2167	AAAAGCACCGAATCGGTGCCTCAGGTTCCCCTGATGATAACGGACTAGCCTTATCTTAACTTGTGTGTTCCCACACTCTAAGAC	TetO_2	sgRNA_08
LA1417	AAAAGCACCGACTCGGTGCCACGCTTTTCAGCGTTGAATACGGACTAGCCTTATCCTAACTTGCCATTTTCATGGCTCTAGGAC	LA1420	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTCCTAGAGCCATGAA	TetO_2	sgRNA_09
LA2168	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTCGGAGAGAACA	LA2169	AAAAGCACCGAATCGGTGCCGTCGTTCGCACGACTGTGTACGGACTAGCCTTATCGGAACTTGAACAGTCCCCTGTTCTCTCCGAC	TetO_2	sgRNA_14
LA1549	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTGCTAGAGTACGTGGA	LA1543	AAAAGCACCGACTCGGTGCCCTGCATTCCTGCAGTGATAACGGACTAGCCTTATGTTAACTTGGGATATCTCTATCCCTCTAACAC	TetO_2	sgRNA_15
LA1550	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTGTTAGAGGGATAGAG	LA1544	AAAAGCACCGAATCGGTGCCTGTCGTTTCCGACATGAATACGGACTAGCCTTATGCTAACTTGTACGTTTCCACGTACTCTAGCAC	TetO_2	sgRNA_16
LA2170	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTGGGAGAGCCAA	LA2171	AAAAGCACCGACTCGGTGCCAGGTCTCCCGACCTTGTGTACGGACTAGCCTTATGGGAACTTGCCAAATTTCTTTGGCTCTCCCAC	TetO_2	sgRNA_21
LA1551	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTG	LA1545	AAAAGCACCGACTCGGTGCCCAGGTCTCCCTGTGATAACGGACTGGCCTTATTCGAACTT	TetO_2	sgRNA_22

sgRNA F		sgRNA R		Target	Backbone Variant
	ATAGTTCGAGAGGACGAGAG		GGACTCTCGTCCTCTCGAAC		
LA2172	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTCCAGAGTCGG	LA2173	AAAAGCACCGAATCGGTGCCTGCCTTCCGGCATGATAACGGACTTGCCTTATTCCAACTTGTCGTTCCCGACTCTGGAAC	TetO_2	sgRNA_23
LA1552	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTGTAGAGCGTAGAAA	LA1546	AAAAGCACCGACTCGGTGCCAGCGTTTCCGCTTGATAACGGACTCGCCTTATTGTAACTTGCGTATTTCTACGCTCTACAAC	TetO_2	sgRNA_24
LA1308	AAAAGCACCGAATCGGTGCCGACGTTCCCACGTCTGATAACGGACTGGCCTTATTGCAACTTGACACTCCCGTGTCTCTGCAAC	LA1418	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTTGCAGAGACACGGG	TetO_2	sgRNA_25
LA2174	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTCCAGAGGTTCG	LA2175	AAAAGCACCGACTCGGTGCCCGAACTCTCGTTCGTGATAACGGACTCGCCTTATCCCAACTTGGTTCTCTCGAACCTCTGGAC	TetO_2	sgRNA_27
LA1416	AAAAGCACCGACTCGGTGCCAGCTCTCCCGAGCTTGATAACGGACTTGCCTTATCGCAACTTGCATCTTTTCAGATGCTCTGCGAC	LA1419	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTCGCAGAGCATCTGA	TetO_2	sgRNA_29
LA1553	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTG	LA1547	AAAAGCACCGACTCGGTGCCACAGCTCCCGCTGTTGATAACGGACTCGCCTTATGCGAAC	TetO_2	sgRNA_30

sgRNA F		**sgRNA R**		**Target**	**Backbone Variant**
	ATAGTGCGAGAGCTTACGAA		TTGCTTACTTTCGTAAGCTCTCGCAC		
LA1554	GAAATTAATACGACTCACTATAGGTGCACTTTTCTCTATCACTGATAGTGCCAGAGAGTAGGGG	LA1548	AAAAGCACCGAATCGGTGCCGGTCATCTCTGACCTGATAACGGACTGGCCTTATGCCAACTTGAGTAGTCCCCTACTCTCTGGCAC	TetO_2	sgRNA_31

Appendix F: Publication

pubs.acs.org/synthbio Technical Note

Expanding the CRISPR Toolbox in Culicine Mosquitoes: *In Vitro* Validation of Pol III Promoters

Michelle A. E. Anderson,[‖] Jessica Purcell,[‖] Sebald A. N. Verkuijl, Victoria C. Norman, Philip T. Leftwich, Tim Harvey-Samuel, and Luke S. Alphey*

Cite This: *ACS Synth. Biol.* 2020, 9, 678–681 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: CRISPR–Cas9-based "gene drive" technologies have been proposed as a novel and effective means of controlling human diseases vectored by mosquitoes. However, more complex designs than those demonstrated to date—and an expanded molecular toolbox with which to build them—will be required to overcome the issues of resistance formation/evolution and drive spatial/temporal limitation. Foreseeing this need, we assessed the sgRNA transcriptional activities of 33 phylogenetically diverse insect Polymerase III promoters using three disease-relevant Culicine mosquito cell lines (*Aedes aegypti*, *Aedes albopictus*, and *Culex quinquefasciatus*). We show that U6 promoters work across species with a range of transcriptional activity levels and find 7SK promoters to be especially promising because of their broad phylogenetic activity. We further show that U6 promoters can be substantially truncated without affecting transcriptional levels. These results will be of great utility to researchers involved in developing the next generation of gene drives.

KEYWORDS: *Polymerase III, Cas9, U6 promoter, 7SK promoter, gene drive, mosquito*

A limited set of RNA Polymerase III (Pol III) promoters, mostly from U6 and H1 genes, have been used in eukaryotic synthetic biology systems to express short non-coding RNAs without the 5′ and 3′ mRNA modification associated with Polymerase II expression. The U6 small nuclear RNA (snRNA) has a highly conserved 106–108 nt sequence and an external 5′ promoter structure that is remarkably similar to that of an RNA Pol II promoter, namely, a TATA-like box and proximal sequence element (PSE). 7SK is another RNA Pol III-transcribed abundant snRNA, whose function in transcriptional regulation is conserved from invertebrates to humans. Like U6, 7SK has an external 5′ promoter structure, similar conserved domains (a TATA-like box and PSE), and in mammals distal elements consisting of an SPH domain and OCT motif.[1] 7SK RNAs have been identified in multiple arthropod species, including Dipterans.[2,3] In this work, we have explored their use for expression of sgRNAs.

In mosquitoes, Pol III promoters have been utilized for genetic control strategies that depend on CRISPR guide RNA (sgRNA) or RNAi expression.[4−8] The ability to express multiple noncoding RNAs while minimizing repetitive sequences is a significant advantage to these systems and may be necessary to create robust technologies.[9,10] More broadly, as the use of mosquitoes as insect synthetic biology chassis develops, it will be highly advantageous for researchers to have access to a diverse range of validated noncoding RNA promoters with varied expression levels; such a toolbox does not yet exist in the Culicines, and we address this need here. Alternative methods for multiplexing sgRNAs from a single transcript, *e.g.*, using tRNAs or ribozyme-based processing, have been demonstrated in other species[11−13] with varying efficiencies but have not yet been applied to mosquitoes.

We hypothesized that adapting existing promoters from related species would be a rapid and cost-effective way of expanding the available Pol III expression toolbox in Culicines, as cross-species activity of U6 promoters in mosquitoes has previously been demonstrated by Konet et al.[6] We systematically tested the activities of a range of previously reported insect U6 promoters in *Aedes aegypti*, *Aedes albopictus*, and *Culex quinquefasciatus* cell lines. This was supplemented by the identification and testing of additional new U6 promoters and the testing of U6[14] and 7SK[2] promoters that had previously been identified but not experimentally tested in cell lines. We used a standardized cell- CRISPR/dCas9–VPR binding assay to systematically quantify the promoter activity across cell lines. Our results represent a large advance in the available

Received: October 25, 2019
Published: March 4, 2020

https://dx.doi.org/10.1021/acssynbio.9b00436
ACS Synth. Biol. 2020, 9, 678–681

expression tools and provide a general guide for efficiently identifying additional expression modalities.

RESULTS AND DISCUSSION

The transcriptional activities of 33 phylogenetically diverse insect Pol III promoters were tested in three cell lines from disease-relevant Culicine mosquito species (*A. aegypti*, *A. albopictus* and *C. quinquefasciatus*). Potential U6 promoters were identified by BLAST using a previously published *A. aegypti* U6 RNA sequence, AAEL017774.[6] The presence of highly conserved sequence elements (a TATA-like box, PSE, and poly-T terminator) were verified for those sequences taken forward experimentally (sequence alignments are provided in the Supporting Information). Each promoter was used to express the same sgRNA, targeting a tetracycline response element (TRE) upstream of the coding sequence of the firefly luciferase gene. Expression of functional sgRNA by the putative promoters, in conjunction with dCas9−VPR, binds the TRE and activates expression of firefly luciferase. Firefly luciferase activity was normalized to the levels of *Renilla* luciferase expressed independently of the sgRNAs (Figure 1).

Figure 1. dCas9−VPR assay components. Our assay consists of four plasmids, each expressing a single component. HR5/IE1, a constitutive promoter in insect cells of baculoviral origin, is used to express a dCas9−VPR fusion protein. A second plasmid containing seven tetO repeats upstream of the *D. melanogaster* minimal hsp70 promoter expresses firefly luciferase upon activation. Test promoters all express the same sgRNA targeting the tetO repeat region. Finally, a plasmid expressing *Renilla* luciferase from the OpIE2 promoter was used as a control to normalize for transfection efficiency.

In Aag2 (*A. aegypti*) and Hsu (*C. quinquefasciatus*) cells, the levels of promoter activity were broadly in line with the species of origin of the promoter, decreased with phylogenetic distance, and accounted for most of our observed variance (R^2_m) (R^2_m = 0.73, R^2_c = 0.89 for Aag2, R^2_m = 0.46, R^2_c = 0.84 for Hsu; Tables S1 and S2 and Figure 2A,B), while random variance introduced by technical replicates (replicate wells transfected with the same mix on the same day) and experimental blocking were low.

In *A. albopictus*-derived U4.4 cells we found no significant effect of the species of origin of the promoter sequence on the relative luciferase expression and larger random variance than in our other experiments (R^2_m = 0.22, R^2_c = 0.89; Table S3 and Figure 2C). We speculate that there may be less overall activity from one or more of our promoter sequences in these cells, which with fewer replicate experiments (still at least three performed on different days) likely explains the lack of an observable pattern here.

Within those species where we tested U6 and 7SK promoter sequences, there was a trend toward 7SK promoter sequences having stronger activity levels than U6 promoter sequences regardless of their species of origin (Tables S1 and S2 and

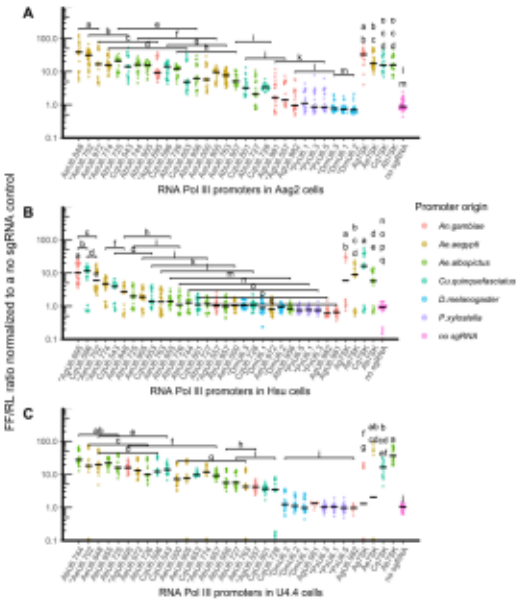


Figure 2. dCas9−VPR assay *in vitro*. Ratios of FF/RL luciferase normalized to a no-sgRNA control are shown. Promoters are organized by median relative activation within U6 and 7SK promoter categories, and the colors denote the promoter origin by species. Lowercase letter groupings denote significant differences at $P < 0.05$ following post hoc analysis. Each point represents one well of a 96-well plate, with at least eight replicate wells transfected in at least three replicate experiments.

Figure 2). None of the U6 promoters from *Drosophila melanogaster* or *Plutella xylostella* showed any activity above background in our mosquito cell lines (Figure 2). Promoters are denoted by the last three digits of their accession numbers.

Shorter versions of several U6 promoters were tested in Aag2 and Hsu cells (Figure 3) in order to determine the minimum possible promoter fragment without compromising the activity. For all seven promoters, the PSE and TATA-like box were present within 100 nt upstream of the transcriptional start and are likely the principal requirements for expression. We did not identify any distal sequence elements with a strong effect on the promoter activity, except in CuU6.801, where deletion from 200 bp to 100 bp essentially eliminated the activity. These results indicate that most of the U6 promoters identified can be used in a very compact form.

Furthering the work of Mount et al.[15] and Konet et al.,[6] we have demonstrated a pipeline for cell culture verification of Pol III promoter sequences in Culicine mosquitoes. In these experiments, we showed that Pol III promoter sequences from closely related species can be used to drive high levels of noncoding RNA expression in mosquito species of interest. Regulatory elements from more distantly related species may be applicable for complex applications where a range of expression levels is desirable. We anticipate that these findings will provide a valuable resource for those involved in the rapidly developing field of mosquito genome editing and synthetic biology.

https://dx.doi.org/10.1021/acssynbio.9b00436
ACS Synth. Biol. 2020, 9, 678−681

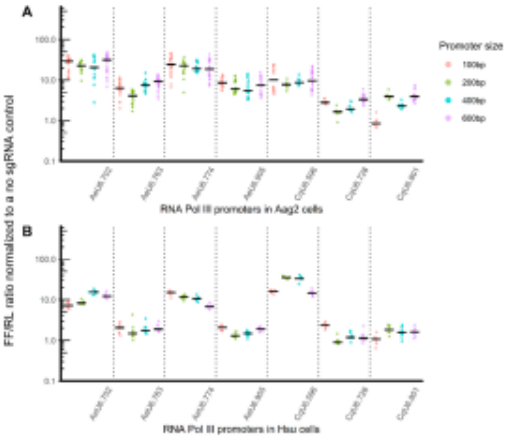


Figure 3. Mosquito U6 promoters maintain dCas9−VPR activity down to 100 bp. Seven promoters were deleted incrementally from a −600 bp fragment down to −100 bp upstream of the transcriptional start. Four promoter lengths were tested for each Pol III promoter, shown in each group left to right as 100 bp (orange), 200 bp (green), 400 bp (blue), and 600 bp (purple). Activity was assessed by the dCas9−VPR assay in Aag2 cells and Hsu cells. FF/RL luciferase ratios have been normalized to a no-sgRNA control. Each point represents one well of a 96-well plate, with at least eight replicate wells transfected at least three replicate experiments.

METHODS

Plasmids and Constructs. Cloning details and complete plasmid sequences are available in the Supporting Information.

Cells, Transfections, and Luciferase Assay. All of the cell lines were maintained at 28 °C without CO_2 or humidity control. Aag2 and U4.4 were maintained in L-15 (Thermo Fisher Scientific, Waltham, MA, U.S.) supplemented with 10% fetal bovine serum (FBS) (Labtech, Lewes, U.K.), 1% penicillin/streptomycin (Pen/Strep) (Thermo Fisher Scientific), and 10% tryptose phosphate broth (Thermo Fisher Scientific). Hsu cells were maintained in Schneider's Drosophila Medium (Lonza, Basel, Switzerland) supplemented with 10% FBS and 1% Pen/Strep. Cell lines were a kind gift of Rennos Fragkoudis.

Cells were seeded in 96-well plates 1 day prior to transfection with the TransIT-PRO transfection kit (Mirus Bio, Madison, WI, U.S.) according to the manufacturer's recommendations. Master mixes were prepared for 8.8 wells of a 96-well plate, and eight replicate wells per experimental construct were transfected in three to eight replicate experiments. In each well, 25 ng of dCas9−VPR plasmid, 25 ng of TRE-firefly reporter plasmid, 0.3 ng of Pol III-sgRNA expressing plasmid, and 50 ng (Aag2, U4.4) or 30 ng (Hsu) of pRL-OpIE2 were used.

Two days after transfection, the cells were washed twice with phosphate-buffered saline, lysed with 1× Passive Lysis Buffer, and then analyzed using the Dual Luciferase Assay Kit on a GloMax Multi+ plate reader (Promega, Southampton, U.K.).

Data Analysis. Luciferase readings were normalized for transfection by dividing the firefly activity by the *Renilla* activity and then normalized to the average of background readings (no-sgRNA control). Data were analyzed by generalized linear mixed models using a Γ distribution with a log link with the glmer function within lme4.[16] Models that encountered convergence errors were fitted with the boyqa optimizer. Each transformed data reading for a promoter was analyzed together with the species of origin and promoter type (U6 or 7SK), and experimental replicates and blocking were nested as a random effect within promoter identity. After each model was fitted, marginal and conditional R^2 values (R^2_m and R^2_c, respectively) were calculated to express the variance explained by the fixed and random factors using the package piecewiseSEM.[17] Pairwise comparisons of different promoter strengths were calculated using Tukey HSD multiple comparison tests using the lsmeans package.[18] All analyses were conducted in R ver. 3.5.3.[19] Scripts and raw data can be found at doi: 10.6084/m9.figshare.11407752.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.9b00436.

Additional information on all constructs, including complete sequences; sequence alignment of mosquito U6 and 7SK promoters analyzed; and summary output tables of full generalized linear mixed models (Tables S1−S3) (PDF)

AUTHOR INFORMATION

Corresponding Author

Luke S. Alphey − *Arthropod Genetics, The Pirbright Institute, Pirbright GU24 0NF, U.K.*; Email: luke.alphey@pirbright.ac.uk

Authors

Michelle A. E. Anderson − *Arthropod Genetics, The Pirbright Institute, Pirbright GU24 0NF, U.K.*; orcid.org/0000-0003-1510-2942

Jessica Purcell − *Arthropod Genetics, The Pirbright Institute, Pirbright GU24 0NF, U.K.*

Sebald A. N. Verkuijl − *Arthropod Genetics, The Pirbright Institute, Pirbright GU24 0NF, U.K.; Department of Zoology, University of Oxford, Oxford OX1 3SZ, U.K.*

Victoria C. Norman − *Arthropod Genetics, The Pirbright Institute, Pirbright GU24 0NF, U.K.*

Philip T. Leftwich − *Arthropod Genetics, The Pirbright Institute, Pirbright GU24 0NF, U.K.; School of Biological Sciences, University of East Anglia, Norwich, Norfolk NR4 7TJ, U.K.*; orcid.org/0000-0001-9500-6592

Tim Harvey-Samuel − *Arthropod Genetics, The Pirbright Institute, Pirbright GU24 0NF, U.K.*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.9b00436

Author Contributions

‖M.A.E.A. and J.P. contributed equally to this work. M.A.E.A., J.P., T.H.-S., and L.S.A. conceived and designed the experiments. M.A.E.A., J.P., S.A.N.V., T.H.-S., and V.C.N. designed and generated constructs or components. M.A.E.A. and J.P. performed the experiments. P.T.L. analyzed the data. M.A.E.A., S.A.N.V., P.T.L., and T.H.-S. wrote the manuscript, and all of the authors read and approved the final draft.

Notes

The authors declare no competing financial interest.

https://dx.doi.org/10.1021/acssynbio.9b00436
ACS Synth. Biol. 2020, 9, 678−681

■ ACKNOWLEDGMENTS

M.A.E.A., S.A.N.V., and L.S.A. were supported through an award from DARPA's Safe Genes Program to MIT (N66001-17-2-4054). The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the U.S. Government. J.P., S.A.N.V., and L.S.A. were supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) (Grant BB/M011224/1 to S.A.N.V. and Grants BBS/E/I/00007033, BBS/E/I/00007038, and BBS/E/I/00007039 to L.S.A.)]. J.P. was supported by a studentship from The Pirbright Institute. P.T.L. was funded through the Wellcome Trust (Investigator Award 110117/Z/15/Z). T.H.-S. and V.C.N. were supported by European Union H2020 Grant nEUROSTRESSPEP (634361).

■ REFERENCES

(1) Diribarne, G., and Bensaude, O. (2009) 7SK RNA, a non-coding RNA regulating P-TEFb, a general transcription factor. *RNA Biol. 6*, 122−128.
(2) Gruber, A. R., Kilgus, C., Mosig, A., Hofacker, I. L., Hennig, W., and Stadler, P. F. (2008) Arthropod 7SK RNA. *Mol. Biol. Evol. 25*, 1923−1930.
(3) Yazbeck, A. M., Tout, K. R., and Stadler, P. F. (2018) Detailed secondary structure models of invertebrate 7SK RNAs. *RNA Biol. 15*, 158−164.
(4) Amarzguioui, M., Rossi, J. J., and Kim, D. (2005) Approaches for chemically synthesized siRNA and vector-mediated RNAi. *FEBS Lett. 579*, 5974−5981.
(5) Bannister, S. C., Wise, T. G., Cahill, D. M., and Doran, T. J. (2007) Comparison of chicken 7SK and U6 RNA polymerase III promoters for short hairpin RNA expression. *BMC Biotechnol. 7*, 79−79.
(6) Konet, D. S., Anderson, J., Piper, J., Akkina, R., Suchman, E., and Carlson, J. (2007) Short-hairpin RNA expressed from polymerase III promoters mediates RNA interference in mosquito cells. *Insect Mol. Biol. 16*, 199−206.
(7) Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., and Zhang, F. (2013) Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science 339*, 819−823.
(8) McManus, M. T., and Sharp, P. A. (2002) Gene silencing in mammals by small interfering RNAs. *Nat. Rev. Genet. 3*, 737.
(9) Hammond, A. M., Kyrou, K., Bruttini, M., North, A., Galizi, R., Karlsson, X., Kranjc, N., Carpi, F. M., D'Aurizio, R., Crisanti, A., and Nolan, T. (2017) The creation and selection of mutations resistant to a gene drive over multiple generations in the malaria mosquito. *PLoS Genet. 13*, No. e1007039.
(10) Oberhofer, G., Ivy, T., and Hay, B. A. (2018) Behavior of homing endonuclease gene drives targeting genes required for viability or female fertility with multiplexed guide RNAs. *Proc. Natl. Acad. Sci. U. S. A. 115*, E9343−E9352.
(11) Xie, K., Minkenberg, B., and Yang, Y. (2015) Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proc. Natl. Acad. Sci. U. S. A. 112*, 3570−3575.
(12) Port, F., and Bullock, S. L. (2016) Augmenting CRISPR applications in Drosophila with tRNA-flanked sgRNAs. *Nat. Methods 13*, 852−854.
(13) Nissim, L., Perli, S. D., Fridkin, A., Perez-Pinera, P., and Lu, T. K. (2014) Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. *Mol. Cell 54*, 698−710.
(14) Hernandez, G., Jr., Valafar, F., and Stumph, W. E. (2007) Insect small nuclear RNA gene promoters evolve rapidly yet retain conserved features involved in determining promoter activity and RNA polymerase specificity. *Nucleic Acids Res. 35*, 21−34.
(15) Mount, S. M., Gotea, V., Lin, C. F., Hernandez, K., and Makalowski, W. (2006) Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA 13*, 5−14.
(16) Bates, D., Machler, M., Bolker, B. M., and Walker, S. C. (2015) Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Software 67* (1), 1−48.
(17) Lefcheck, J. S. (2016) PIECEWISESEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods Ecol. Evol. 7*, 573−579.
(18) Lenth, R. V. (2016) Least-Squares Means: The R Package lsmeans. *J. Stat. Software 69* (1), 1−33.
(19) R Core Team (2019) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria; http://www.R-project.org/.

https://dx.doi.org/10.1021/acssynbio.9b00436
ACS Synth. Biol. 2020, 9, 678−681

References

AKBARI, O. S., PAPATHANOS, P. A., SANDLER, J. E., KENNEDY, K. & HAY, B. A. 2014. Identification of germline transcriptional regulatory elements in Aedes aegypti. *Sci Rep,* 4**,** 3954.

ALPHEY, L. 2002. Re-engineering the sterile insect technique. *Insect Biochem Mol Biol,* 32**,** 1243-7.

ALPHEY, L. 2014. Genetic control of mosquitoes. *Annu Rev Entomol,* 59**,** 205-24.

ALPHEY, L., BENEDICT, M., BELLINI, R., CLARK, G. G., DAME, D. A., SERVICE, M. W. & DOBSON, S. L. 2010. Sterile-insect methods for control of mosquito-borne diseases: an analysis. *Vector Borne Zoonotic Dis,* 10**,** 295-311.

ALPHEY, L. S., CRISANTI, A., RANDAZZO, F. F. & AKBARI, O. S. 2020. Opinion: Standardizing the definition of gene drive. *Proc Natl Acad Sci U S A,* 117**,** 30864-30867.

AMRAOUI, F., VAZEILLE, M. & FAILLOUX, A. B. 2016. French Aedes albopictus are able to transmit yellow fever virus. *Euro Surveill,* 21.

ANDERSON, M. A., GROSS, T. L., MYLES, K. M. & ADELMAN, Z. N. 2010. Validation of novel promoter sequences derived from two endogenous ubiquitin genes in transgenic Aedes aegypti. *Insect Mol Biol,* 19**,** 441-9.

ANDERSON, M. A. E., PURCELL, J., VERKUIJL, S. A. N., NORMAN, V. C., LEFTWICH, P. T., HARVEY-SAMUEL, T. & ALPHEY, L. S. 2020. Expanding the CRISPR Toolbox in Culicine Mosquitoes: In Vitro Validation of Pol III Promoters. *ACS Synth Biol,* 9**,** 678-681.

ANNAS, G. J., BEISEL, C. L., CLEMENT, K., CRISANTI, A., FRANCIS, S., GALARDINI, M., GALIZI, R., GRUNEWALD, J., IMMOBILE, G., KHALIL, A. S., MULLER, R., PATTANAYAK, V., PETRI, K., PAUL, L., PINELLO, L., SIMONI, A., TAXIARCHI, C. & JOUNG, J. K. 2021. A Code of Ethics for Gene Drive Research. *CRISPR J,* 4**,** 19-24.

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. & TRACKING RESISTANCE TO ARTEMISININ, C. 2014. Spread of artemisinin resistance in Plasmodium falciparum malaria. *N Engl J Med,* 371**,** 411-23.

ASHLEY, E. A. & PHYO, A. P. 2018. Drugs in Development for Malaria. *Drugs,* 78**,** 861-879.

ATCC. 2010. *Cell Line Authentication Test Recommendations* [Online]. Available: https://www.lgcstandards-atcc.org/~/media/PDFs/Technical%20Bulletins/tb08.ashx [Accessed 02/2021 2021].

BENEDICT, M. Q., BURT, A., CAPURRO, M. L., DE BARRO, P., HANDLER, A. M., HAYES, K. R., MARSHALL, J. M., TABACHNICK, W. J. & ADELMAN, Z. N. 2018. Recommendations for

Laboratory Containment and Management of Gene Drive Systems in Arthropods. *Vector Borne Zoonotic Dis,* 18**,** 2-13.

BERGHAMMER, A. J., KLINGLER, M. & WIMMER, E. A. 1999. A universal marker for transgenic insects. *Nature,* 402**,** 370-1.

BHATT, S., GETHING, P. W., BRADY, O. J., MESSINA, J. P., FARLOW, A. W., MOYES, C. L., DRAKE, J. M., BROWNSTEIN, J. S., HOEN, A. G., SANKOH, O., MYERS, M. F., GEORGE, D. B., JAENISCH, T., WINT, G. R., SIMMONS, C. P., SCOTT, T. W., FARRAR, J. J. & HAY, S. I. 2013. The global distribution and burden of dengue. *Nature,* 496**,** 504-7.

BHATT, S., WEISS, D. J., CAMERON, E., BISANZIO, D., MAPPIN, B., DALRYMPLE, U., BATTLE, K., MOYES, C. L., HENRY, A., ECKHOFF, P. A., WENGER, E. A., BRIET, 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.

BHAYA, D., DAVISON, M. & BARRANGOU, R. 2011. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu Rev Genet,* 45**,** 273-97.

BIAN, G., JOSHI, D., DONG, Y., LU, P., ZHOU, G., PAN, X., XU, Y., DIMOPOULOS, G. & XI, Z. 2013. Wolbachia invades Anopheles stephensi populations and induces refractoriness to Plasmodium infection. *Science,* 340**,** 748-51.

BIAN, G., XU, Y., LU, P., XIE, Y. & XI, Z. 2010. The endosymbiotic bacterium Wolbachia induces resistance to dengue virus in Aedes aegypti. *PLoS Pathog,* 6**,** e1000833.

BIBIKOVA, M., BEUMER, K., TRAUTMAN, J. K. & CARROLL, D. 2003. Enhancing gene targeting with designed zinc finger nucleases. *Science,* 300**,** 764.

BIKARD, D., JIANG, W., SAMAI, P., HOCHSCHILD, A., ZHANG, F. & MARRAFFINI, L. A. 2013. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res,* 41**,** 7429-37.

BOLDSYSTEMS. 2007. *Animal Identification [COI] - Species Level Barcode Records* [Online]. Available: http://www.boldsystems.org/index.php/IDS_OpenIdEngine [Accessed].

BOLLAG, R. J., WALDMAN, A. S. & LISKAY, R. M. 1989. Homologous recombination in mammalian cells. *Annu Rev Genet,* 23**,** 199-225.

BRAND, A. H. & PERRIMON, N. 1993. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development,* 118**,** 401-15.

BREWER, N., MCKENZIE, M. S., MELKONJAN, N., ZAKY, M., VIK, R., STOFFOLANO, J. G. & WEBLEY, W. C. 2021. Persistence and Significance of Chlamydia trachomatis in the Housefly, Musca domestica L. *Vector Borne Zoonotic Dis,* 21**,** 854-863.

BRINER, A. E., DONOHOUE, P. D., GOMAA, A. A., SELLE, K., SLORACH, E. M., NYE, C. H., HAURWITZ, R. E., BEISEL, C. L., MAY, A. P. & BARRANGOU, R. 2014. Guide RNA functional modules direct Cas9 activity and orthogonality. *Mol Cell,* 56**,** 333-339.

BRINSTER, R. L., ALLEN, J. M., BEHRINGER, R. R., GELINAS, R. E. & PALMITER, R. D. 1988. Introns increase transcriptional efficiency in transgenic mice. *Proc Natl Acad Sci U S A,* 85**,** 836-40.

BROCKEN, D. J. W., TARK-DAME, M. & DAME, R. T. 2018. dCas9: A Versatile Tool for Epigenome Editing. *Curr Issues Mol Biol,* 26**,** 15-32.

BUCHMAN, A. R. & BERG, P. 1988. Comparison of intron-dependent and intron-independent gene expression. *Mol Cell Biol,* 8**,** 4395-405.

BURT, A. 2003. Site-specific selfish genes as tools for the control and genetic engineering of natural populations. *Proc Biol Sci,* 270**,** 921-8.

CASAS-MOLLANO, J. A., ZINSELMEIER, M. H., ERICKSON, S. E. & SMANSKI, M. J. 2020. CRISPR-Cas Activators for Engineering Gene Expression in Higher Eukaryotes. *CRISPR J,* 3**,** 350-364.

CATTERUCCIA, F., NOLAN, T., BLASS, C., MULLER, H. M., CRISANTI, A., KAFATOS, F. C. & LOUKERIS, T. G. 2000. Toward Anopheles transformation: Minos element activity in anopheline cells and embryos. *Proc Natl Acad Sci U S A,* 97**,** 2157-62.

CAVENER, D. R. 1987. Comparison of the Consensus Sequence Flanking Translational Start Sites in Drosophila and Vertebrates. *Nucleic Acids Research,* 15**,** 1353-1361.

CAVENER, D. R. & RAY, S. C. 1991. Eukaryotic Start and Stop Translation Sites. *Nucleic Acids Research,* 19**,** 3185-3192.

CHANDRAMOHAN, D., ZONGO, I., SAGARA, I., CAIRNS, M., YERBANGA, R. S., DIARRA, M., NIKIEMA, F., TAPILY, A., SOMPOUGDOU, F., ISSIAKA, D., ZOUNGRANA, C., SANOGO, K., HARO, A., KAYA, M., SIENOU, A. A., TRAORE, S., MAHAMAR, A., THERA, I., DIARRA, K., DOLO, A., KUEPFER, I., SNELL, P., MILLIGAN, P., OCKENHOUSE, C., OFORI-ANYINAM, O., TINTO, H., DJIMDE, A., OUEDRAOGO, J. B., DICKO, A. & GREENWOOD, B. 2021. Seasonal Malaria Vaccination with or without Seasonal Malaria Chemoprevention. *N Engl J Med,* 385**,** 1005-1017.

CHANG, M. J., KUZIO, J. & BLISSARD, G. W. 1999. Modulation of translational efficiency by contextual nucleotides flanking a Baculovirus initiator AUG codon. *Virology,* 259**,** 369-383.

CHARREL, R. N., LEPARC-GOFFART, I., GALLIAN, P. & DE LAMBALLERIE, X. 2014. Globalization of Chikungunya: 10 years to invade the world. *Clin Microbiol Infect,* 20**,** 662-3.

CHATTERJEE, R. 2007. Cell biology. Cases of mistaken identity. *Science,* 315**,** 928-31.

CHAVEZ, A., SCHEIMAN, J., VORA, S., PRUITT, B. W., TUTTLE, M., E, P. R. I., LIN, S., KIANI, S., GUZMAN, C. D., WIEGAND, D. J., TER-OVANESYAN, D., BRAFF, J. L., DAVIDSOHN, N., HOUSDEN, B. E., PERRIMON, N., WEISS, R., AACH, J., COLLINS, J. J. & CHURCH, G. M. 2015. Highly efficient Cas9-mediated transcriptional programming. *Nat Methods,* 12**,** 326-8.

CHAVEZ, A., TUTTLE, M., PRUITT, B. W., EWEN-CAMPEN, B., CHARI, R., TER-OVANESYAN, D., HAQUE, S. J., CECCHI, R. J., KOWAL, E. J. K., BUCHTHAL, J., HOUSDEN, B. E., PERRIMON, N., COLLINS, J. J. & CHURCH, G. 2016. Comparison of Cas9 activators in multiple species. *Nat Methods,* 13**,** 563-567.

CHENG, A. W., WANG, H., YANG, H., SHI, L., KATZ, Y., THEUNISSEN, T. W., RANGARAJAN, S., SHIVALILA, C. S., DADON, D. B. & JAENISCH, R. 2013. Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Res,* 23**,** 1163-71.

CHRISTIAN, K. & ANDREAS, V. 2013. Production of Recombinant Proteins in Insect Cells. *American Journal of Biochemistry and Biotechnology,* 9.

CHYLINSKI, K., LE RHUN, A. & CHARPENTIER, E. 2013. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol,* 10**,** 726-37.

CLARK, K., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & SAYERS, E. W. 2016. GenBank. *Nucleic Acids Res,* 44**,** D67-72.

DANG, Y., JIA, G., CHOI, J., MA, H., ANAYA, E., YE, C., SHANKAR, P. & WU, H. 2015. Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. *Genome Biol,* 16**,** 280.

DIAGNE, C. T., DIALLO, D., FAYE, O., BA, Y., FAYE, O., GAYE, A., DIA, I., FAYE, O., WEAVER, S. C., SALL, A. A. & DIALLO, M. 2015. Potential of selected Senegalese Aedes spp. mosquitoes (Diptera: Culicidae) to transmit Zika virus. *BMC Infect Dis,* 15**,** 492.

DOBSON, S. L., MARSLAND, E. J. & RATTANADECHAKUL, W. 2001. Wolbachia-induced cytoplasmic incompatibility in single- and superinfected Aedes albopictus (Diptera: Culicidae). *J Med Entomol,* 38**,** 382-7.

DONG, S., LIN, J., HELD, N. L., CLEM, R. J., PASSARELLI, A. L. & FRANZ, A. W. 2015. Heritable CRISPR/Cas9-mediated genome editing in the yellow fever mosquito, Aedes aegypti. *PLoS One,* 10**,** e0122353.

DORIGATTI, I., MCCORMACK, C., NEDJATI-GILANI, G. & FERGUSON, N. M. 2018. Using Wolbachia for Dengue Control: Insights from Modelling. *Trends Parasitol,* 34**,** 102-113.

DOURIS, V., SWEVERS, L., LABROPOULOU, V., ANDRONOPOULOU, E., GEORGOUSSI, Z. & IATROU, K. 2006. Stably transformed insect cell lines: tools for expression of secreted and membrane-anchored proteins and high-throughput screening platforms for drug and insecticide discovery. *Adv Virus Res,* 68**,** 113-56.

DUNCKER, B. P., DAVIES, P. L. & WALKER, V. K. 1997. Introns boost transgene expression in Drosophila melanogaster. *Mol Gen Genet,* 254**,** 291-6.

DUNN, D. W. & FOLLETT, P. A. 2017. The Sterile Insect Technique (SIT) – an introduction. *Entomologia Experimentalis et Applicata,* 164**,** 151-154.

ECDC., E. C. F. D. P. A. C. 2017. Vector control with a focus on *Aedes aegypti* and *Aedes albopictus* mosquitoes. *Technical Report.* Stockholm: ECDC.

ECDC., E. C. F. D. P. A. C. 2020. *Culex pipiens - Factsheet for experts* [Online]. Available: https://www.ecdc.europa.eu/en/all-topics-z/disease-vectors/facts/mosquito-factsheets/culex-pipiens-factsheet-experts [Accessed Nov 2021 2021].

EMCA., E. M. C. A., WHO 2013. Guidelines for the Control of Mosquitoes of Public Health Importance in Europe. 2013 ed.

ESU, E., LENHART, A., SMITH, L. & HORSTICK, O. 2010. Effectiveness of peridomestic space spraying with insecticide on dengue transmission; systematic review. *Trop Med Int Health,* 15**,** 619-31.

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.

FARAJOLLAHI, A., FONSECA, D. M., KRAMER, L. D. & MARM KILPATRICK, A. 2011. "Bird biting" mosquitoes and human disease: a review of the role of Culex pipiens complex mosquitoes in epidemiology. *Infect Genet Evol,* 11**,** 1577-85.

FDA FDA Vaccines Licensed for Use in the United States.

FENG, X., LOPEZ DEL AMO, V., MAMELI, E., LEE, M., BISHOP, A. L., PERRIMON, N. & GANTZ, V. M. 2021. Optimized CRISPR tools and site-directed transgenesis towards gene drive development in Culex quinquefasciatus mosquitoes. *Nat Commun,* 12**,** 2960.

FERGUSON, N. M., KIEN, D. T., CLAPHAM, H., AGUAS, R., TRUNG, V. T., CHAU, T. N., POPOVICI, J., RYAN, P. A., O'NEILL, S. L., MCGRAW, E. A., LONG, V. T., DUI LE, T., NGUYEN, H. L., CHAU, N. V., WILLS, B. & SIMMONS, C. P. 2015. Modeling the impact on virus transmission of Wolbachia-mediated blocking of dengue virus infection of Aedes aegypti. *Sci Transl Med,* 7**,** 279ra37.

FOIL, L. & GORHAM, J. 2000. Mechanical Transmission of Disease Agents by Arthropods. *Medical Entomology***,** 461-514.

FOLMER, O., BLACK, M., HOEH, W., LUTZ, R. & VRIJENHOEK, R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol,* 3**,** 294-9.

FU, G., LEES, R. S., NIMMO, D., AW, D., JIN, L., GRAY, P., BERENDONK, T. U., WHITE-COOPER, H., SCAIFE, S., KIM PHUC, H., MARINOTTI, O., JASINSKIENE, N., JAMES, A. A. & ALPHEY, L. 2010. Female-specific flightless phenotype for mosquito control. *Proc Natl Acad Sci U S A,* 107**,** 4550-4.

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. *Proc Natl Acad Sci U S A,* 112**,** E6736-43.

GARNEAU, J. E., DUPUIS, M. E., VILLION, M., ROMERO, D. A., BARRANGOU, R., BOYAVAL, P., FREMAUX, C., HORVATH, P., MAGADAN, A. H. & MOINEAU, S. 2010. The CRISPR/Cas

bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature,* 468**,** 67-71.

GARSKE, T., VAN KERKHOVE, M. D., YACTAYO, S., RONVEAUX, O., LEWIS, R. F., STAPLES, J. E., PEREA, W., FERGUSON, N. M. & YELLOW FEVER EXPERT, C. 2014. Yellow Fever in Africa: estimating the burden of disease and impact of mass vaccination from outbreak and serological data. *PLoS Med,* 11**,** e1001638.

GEURTS, A. M., YANG, Y., CLARK, K. J., LIU, G., CUI, Z., DUPUY, A. J., BELL, J. B., LARGAESPADA, D. A. & HACKETT, P. B. 2003. Gene transfer into genomes of human cells by the sleeping beauty transposon system. *Mol Ther,* 8**,** 108-17.

GILBERT, L. A., LARSON, M. H., MORSUT, L., LIU, Z., BRAR, G. A., TORRES, S. E., STERN-GINOSSAR, N., BRANDMAN, O., WHITEHEAD, E. H., DOUDNA, J. A., LIM, W. A., WEISSMAN, J. S. & QI, L. S. 2013. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell,* 154**,** 442-51.

GIRALDO-CALDERON, G. I., EMRICH, S. J., MACCALLUM, R. M., MASLEN, G., DIALYNAS, E., TOPALIS, P., HO, N., GESING, S., VECTORBASE, C., 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 Res,* 43**,** D707-13.

GIRARD, M., NELSON, C. B., PICOT, V. & GUBLER, D. J. 2020. Arboviruses: A global public health threat. *Vaccine,* 38**,** 3989-3994.

GLASER, R. L. & MEOLA, M. A. 2010. The native Wolbachia endosymbionts of Drosophila melanogaster and Culex quinquefasciatus increase host resistance to West Nile virus infection. *PLoS One,* 5**,** e11977.

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. *Nat Biotechnol,* 23**,** 453-6.

GOOD, P. D., KRIKOS, A. J., LI, S. X., BERTRAND, E., LEE, N. S., GIVER, L., ELLINGTON, A., ZAIA, J. A., ROSSI, J. J. & ENGELKE, D. R. 1997. Expression of small, therapeutic RNAs in human cell nuclei. *Gene Ther,* 4**,** 45-54.

GOTUZZO, E., YACTAYO, S. & CORDOVA, E. 2013. Efficacy and duration of immunity after yellow fever vaccination: systematic review on the need for a booster every 10 years. *Am J Trop Med Hyg,* 89**,** 434-44.

GRATZ, N. G. 1999. Emerging and resurging vector-borne diseases. *Annu Rev Entomol,* 44**,** 51-75.

GRATZ, S. J., CUMMINGS, A. M., NGUYEN, J. N., HAMM, D. C., DONOHUE, L. K., HARRISON, M. M., WILDONGER, J. & O'CONNOR-GILES, K. M. 2013. Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. *Genetics,* 194**,** 1029-35.

GROSSMAN, G. L., RAFFERTY, C. S., CLAYTON, J. R., STEVENS, T. K., MUKABAYIRE, O. & BENEDICT, M. Q. 2001. Germline transformation of the malaria vector, Anopheles gambiae, with the piggyBac transposable element. *Insect Mol Biol,* 10**,** 597-604.

GRUBER, A. R., KILGUS, C., MOSIG, A., HOFACKER, I. L., HENNIG, W. & STADLER, P. F. 2008a. Arthropod 7SK RNA. *Mol Biol Evol,* 25**,** 1923-30.

GRUBER, A. R., KOPER-EMDE, D., MARZ, M., TAFER, H., BERNHART, S., OBERNOSTERER, G., MOSIG, A., HOFACKER, I. L., STADLER, P. F. & BENECKE, B. J. 2008b. Invertebrate 7SK snRNAs. *J Mol Evol,* 66**,** 107-15.

GUARINO, L. A., GONZALEZ, M. A. & SUMMERS, M. D. 1986. Complete Sequence and Enhancer Function of the Homologous DNA Regions of Autographa californica Nuclear Polyhedrosis Virus. *J Virol,* 60**,** 224-9.

GUBLER, D. J. 2002. The global emergence/resurgence of arboviral diseases as public health problems. *Arch Med Res,* 33**,** 330-42.

HAGHIGHAT-KHAH, R. E., HARVEY-SAMUEL, T., BASU, S., STJOHN, O., SCAIFE, S., VERKUIJL, S., LOVETT, E. & ALPHEY, L. 2019. Engineered action at a distance: Blood-meal-inducible paralysis in Aedes aegypti. *PLoS Negl Trop Dis,* 13**,** e0007579.

HAGHIGHAT-KHAH, R. E., SCAIFE, S., MARTINS, S., ST JOHN, O., MATZEN, K. J., MORRISON, N. & ALPHEY, L. 2015. Site-specific cassette exchange systems in the Aedes aegypti mosquito and the Plutella xylostella moth. *PLoS One,* 10**,** e0121097.

HAMER, G. L., KITRON, U. D., GOLDBERG, T. L., BRAWN, J. D., LOSS, S. R., RUIZ, M. O., HAYES, D. B. & WALKER, E. D. 2009. Host selection by Culex pipiens mosquitoes and West Nile virus amplification. *Am J Trop Med Hyg,* 80**,** 268-78.

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 Biotechnol,* 34**,** 78-83.

HAMMOND, A., KARLSSON, X., MORIANOU, I., KYROU, K., BEAGHTON, A., GRIBBLE, M., KRANJC, N., GALIZI, R., BURT, A., CRISANTI, A. & NOLAN, T. 2021. Regulating the expression of gene drives is key to increasing their invasive potential and the mitigation of resistance. *PLoS Genet,* 17**,** e1009321.

HAMMOND, A. M., 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 Genet,* 13**,** e1007039.

HARVEY-SAMUEL, T. D., XU, X., LOVETT, E., DAFA'ALLA, T., WALKER, A., NORMAN, V. C., CARTER, R., TEAL, J., AKILAN, L., LEFTWICH, P. T. & ALPHEY, L. 2020. Engineered expression of the invertebrate-specific scorpion toxin AaHIT reduces adult longevity and female fecundity in the diamondback moth Plutella xylostella. *bioRxiv*.

HEGEDUS, D. D., PFEIFER, T. A., HENDRY, J., THEILMANN, D. A. & GRIGLIATTI, T. A. 1998. A series of broad host range shuttle vectors for constitutive and inducible expression of heterologous proteins in insect cell lines. *Gene,* 207**,** 241-9.

HELINSKI, M. E., PARKER, A. G. & KNOLS, B. G. 2009. Radiation biology of mosquitoes. *Malar J,* 8 Suppl 2**,** S6.

HERNANDEZ, G., JR., VALAFAR, F. & STUMPH, W. E. 2007. Insect small nuclear RNA gene promoters evolve rapidly yet retain conserved features involved in determining promoter activity and RNA polymerase specificity. *Nucleic Acids Res,* 35**,** 21-34.

HILGENBOECKER, K., HAMMERSTEIN, P., SCHLATTMANN, P., TELSCHOW, A. & WERREN, J. H. 2008. How many species are infected with Wolbachia?--A statistical analysis of current data. *FEMS Microbiol Lett,* 281**,** 215-20.

HILL, C. A., KAFATOS, F. C., STANSFIELD, S. K. & COLLINS, F. H. 2005. Arthropod-borne diseases: vector control in the genomics era. *Nat Rev Microbiol,* 3**,** 262-8.

HOFFMANN, A. A., ITURBE-ORMAETXE, I., CALLAHAN, A. G., PHILLIPS, B. L., BILLINGTON, K., AXFORD, J. K., MONTGOMERY, B., TURLEY, A. P. & O'NEILL, S. L. 2014. Stability of the wMel Wolbachia Infection following invasion into Aedes aegypti populations. *PLoS Negl Trop Dis,* 8**,** e3115.

HORSTICK, E. J., JORDAN, D. C., BERGERON, S. A., TABOR, K. M., SERPE, M., FELDMAN, B. & BURGESS, H. A. 2015. Increased functional protein expression using nucleotide sequence features enriched in highly expressed genes in zebrafish. *Nucleic Acids Res,* 43**,** e48.

HORVATH, P. & BARRANGOU, R. 2010. CRISPR/Cas, the immune system of bacteria and archaea. *Science,* 327**,** 167-70.

HSU, S. H., MAO, W. H. & CROSS, J. H. 1970. Establishment of a line of cells derived from ovarian tissue of Clex quinquefasciatus Say. *J Med Entomol,* 7**,** 703-7.

HUANG, M. T. & GORMAN, C. M. 1990. Intervening sequences increase efficiency of RNA 3' processing and accumulation of cytoplasmic RNA. *Nucleic Acids Res,* 18**,** 937-47.

HUANG, Y., WANG, Y., ZENG, B., LIU, Z., XU, X., MENG, Q., HUANG, Y., YANG, G., VASSEUR, L., GURR, G. M. & YOU, M. 2017. Functional characterization of Pol III U6 promoters for gene knockdown and knockout in Plutella xylostella. *Insect Biochem Mol Biol,* 89**,** 71-78.

HUYNH, C. Q. & ZIELER, H. 1999. Construction of modular and versatile plasmid vectors for the high-level expression of single or multiple genes in insects and insect cell lines. *J Mol Biol,* 288**,** 13-20.

IGARASHI, A. 1978. Isolation of a Singh's Aedes albopictus cell clone sensitive to Dengue and Chikungunya viruses. *J Gen Virol,* 40**,** 531-44.

ITURBE-ORMAETXE, I., WALKER, T. & SL, O. N. 2011. Wolbachia and the biological control of mosquito-borne disease. *EMBO Rep,* 12**,** 508-18.

JACKSON, R. J., HELLEN, C. U. & PESTOVA, T. V. 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol,* 11**,** 113-27.

JARVIS, D. L., WEINKAUF, C. & GUARINO, L. A. 1996. Immediate-early baculovirus vectors for foreign gene expression in transformed or infected insect cells. *Protein Expr Purif,* 8**,** 191-203.

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**,** 816-21.

KABADI, A. M., OUSTEROUT, D. G., HILTON, I. B. & GERSBACH, C. A. 2014. Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. *Nucleic Acids Res,* 42**,** e147.

KAMBHAMPATI, S., RAI, K. S. & BURGUN, S. J. 1993. Unidirectional Cytoplasmic Incompatibility in the Mosquito, Aedes Albopictus. *Evolution,* 47**,** 673-677.

KATAHIRA, J. 2015. Nuclear export of messenger RNA. *Genes (Basel),* 6**,** 163-84.

KOCH, R., LEDERMANN, R., URWYLER, O., HELLER, M. & SUTER, B. 2009. Systematic functional analysis of Bicaudal-D serine phosphorylation and intragenic suppression of a female sterile allele of BicD. *PLoS One,* 4**,** e4552.

KONET, D. S., ANDERSON, J., PIPER, J., AKKINA, R., SUCHMAN, E. & CARLSON, J. 2007. Short-hairpin RNA expressed from polymerase III promoters mediates RNA interference in mosquito cells. *Insect Mol Biol,* 16**,** 199-206.

KOZAK, M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell,* 44**,** 283-92.

KOZAK, M. 1987a. An Analysis of 5'-Noncoding Sequences from 699 Vertebrate Messenger-Rnas. *Nucleic Acids Research,* 15**,** 8125-8148.

KOZAK, M. 1987b. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J Mol Biol,* 196**,** 947-50.

KRAEMER, M. U., SINKA, M. E., DUDA, K. A., MYLNE, A. Q., SHEARER, F. M., BARKER, C. M., MOORE, C. G., CARVALHO, R. G., COELHO, G. E., VAN BORTEL, W., HENDRICKX, G., SCHAFFNER, F., ELYAZAR, I. R., TENG, H. J., BRADY, O. J., MESSINA, J. P., PIGOTT, D. M., SCOTT, T. W., SMITH, D. L., WINT, G. R., GOLDING, N. & HAY, S. I. 2015. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *Elife,* 4**,** e08347.

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. *Nat Biotechnol,* 36**,** 1062-1066.

LABBE, G. M., NIMMO, D. D. & ALPHEY, L. 2010. piggybac- and PhiC31-mediated genetic transformation of the Asian tiger mosquito, Aedes albopictus (Skuse). *PLoS Negl Trop Dis,* 4**,** e788.

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**,** W272-6.

LEDOGAR, R. J., AROSTEGUI, J., HERNANDEZ-ALVAREZ, C., MORALES-PEREZ, A., NAVA-AGUILERA, E., LEGORRETA-SOBERANIS, J., SUAZO-LAGUNA, H., BELLI, A., LAUCIRICA, J., COLOMA, J., HARRIS, E. & ANDERSSON, N. 2017. Mobilising communities for Aedes aegypti control: the SEPA approach. *BMC Public Health,* 17**,** 403.

LENTH, R. V. 2020. emmeans: Estimated Marginal Means, aka Least-Squares Means. *R package version 1.5.3.* https://CRAN.R-project.org/package=emmeans.

LI, M., BUI, M., YANG, T., BOWMAN, C. S., WHITE, B. J. & AKBARI, O. S. 2017. Germline Cas9 expression yields highly efficient genome engineering in a major worldwide disease vector, Aedes aegypti. *Proc Natl Acad Sci U S A,* 114**,** E10540-E10549.

LI, M., YANG, T., KANDUL, N. P., BUI, M., GAMEZ, S., RABAN, R., BENNETT, J., SANCHEZ, C. H., LANZARO, G. C., SCHMIDT, H., LEE, Y., MARSHALL, J. M. & AKBARI, O. S. 2020. Development of a confinable gene drive system in the human disease vector Aedes aegypti. *Elife,* 9.

LIEBER, M. R. 2010. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu Rev Biochem,* 79**,** 181-211.

LIU, Y., ZHANG, Y., YAO, L., HAO, H., FU, X., YANG, Z. & DU, E. 2015. Enhanced production of porcine circovirus type 2 (PCV2) virus-like particles in Sf9 cells by translational enhancers. *Biotechnol Lett,* 37**,** 1765-71.

LOGAN, J. G., SEAL, N. J., COOK, J. I., STANCZYK, N. M., BIRKETT, M. A., CLARK, S. J., GEZAN, S. A., WADHAMS, L. J., PICKETT, J. A. & MORDUE, A. J. 2009. Identification of human-derived volatile chemicals that interfere with attraction of the Scottish biting midge and their potential use as repellents. *J Med Entomol,* 46**,** 208-19.

LURA, T., CUMMINGS, R., VELTEN, R., DE COLLIBUS, K., MORGAN, T., NGUYEN, K. & GERRY, A. 2012. Host (avian) biting preference of southern California Culex mosquitoes (Diptera: Culicidae). *J Med Entomol,* 49**,** 687-96.

MA, S., CHANG, J., WANG, X., LIU, Y., ZHANG, J., LU, W., GAO, J., SHI, R., ZHAO, P. & XIA, Q. 2014. CRISPR/Cas9 mediated multiplex genome editing and heritable mutagenesis of BmKu70 in Bombyx mori. *Sci Rep,* 4**,** 4489.

MABASHI-ASAZUMA, H. & JARVIS, D. L. 2017. CRISPR-Cas9 vectors for genome editing and host engineering in the baculovirus-insect cell system. *Proc Natl Acad Sci U S A,* 114**,** 9068-9073.

MARSHALL, J. M., BUCHMAN, A., SANCHEZ, C. H. & AKBARI, O. S. 2017. Overcoming evolved resistance to population-suppressing homing-based gene drives. *Sci Rep,* 7**,** 3776.

MARTINS, S., NAISH, N., WALKER, A. S., MORRISON, N. I., SCAIFE, S., FU, G., DAFA'ALLA, T. & ALPHEY, L. 2012. Germline transformation of the diamondback moth, Plutella xylostella L., using the piggyBac transposable element. *Insect Mol Biol,* 21**,** 414-21.

MATERA, A. G., TERNS, R. M. & TERNS, M. P. 2007. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol,* 8**,** 209-20.

MAYR, C. 2019. What Are 3' UTRs Doing? *Cold Spring Harb Perspect Biol,* 11.

MCLEAN, K. J. & JACOBS-LORENA, M. 2016. Genetic Control Of Malaria Mosquitoes. *Trends Parasitol,* 32**,** 174-176.

MEREDITH, J. M., UNDERHILL, A., MCARTHUR, C. C. & EGGLESTON, P. 2013. Next-generation site-directed transgenesis in the malaria vector mosquito Anopheles gambiae: self-docking strains expressing germline-specific phiC31 integrase. *PLoS One,* 8**,** e59264.

MONTAGUE, T. G., CRUZ, J. M., GAGNON, J. A., CHURCH, G. M. & VALEN, E. 2014. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res,* 42**,** W401-7.

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.

MOYES, C. L., VONTAS, J., MARTINS, A. J., NG, L. C., KOOU, S. Y., DUSFOUR, I., RAGHAVENDRA, K., PINTO, J., CORBEL, V., DAVID, J. P. & WEETMAN, D. 2017. Contemporary status of insecticide resistance in the major Aedes vectors of arboviruses infecting humans. *PLoS Negl Trop Dis,* 11**,** e0005625.

MULLER, H. M., DIMOPOULOS, G., BLASS, C. & KAFATOS, F. C. 1999. A hemocyte-like cell line established from the malaria vector Anopheles gambiae expresses six prophenoloxidase genes. *J Biol Chem,* 274**,** 11727-35.

NOBLE, C., MIN, J., OLEJARZ, J., BUCHTHAL, J., CHAVEZ, A., SMIDLER, A. L., DEBENEDICTIS, E. A., CHURCH, G. M., NOWAK, M. A. & ESVELT, K. M. 2019. Daisy-chain gene drives for the alteration of local populations. *Proc Natl Acad Sci U S A,* 116**,** 8275-8282.

NOBLE, C., OLEJARZ, J., ESVELT, K. M., CHURCH, G. M. & NOWAK, M. A. 2017. Evolutionary dynamics of CRISPR gene drives. *Sci Adv,* 3**,** e1601964.

NOSTEN, F. & WHITE, N. J. 2007. Artemisinin-based combination treatment of falciparum malaria. *Am J Trop Med Hyg,* 77**,** 181-92.

O'MEARA, G. F., EVANS, L. F., JR., GETTMAN, A. D. & CUDA, J. P. 1995. Spread of Aedes albopictus and decline of Ae. aegypti (Diptera: Culicidae) in Florida. *J Med Entomol,* 32**,** 554-62.

O'NEILL, S. L., RYAN, P. A., TURLEY, A. P., WILSON, G., RETZKI, K., ITURBE-ORMAETXE, I., DONG, Y., KENNY, N., PATON, C. J., RITCHIE, S. A., BROWN-KENYON, J., STANFORD, D., WITTMEIER, N., JEWELL, N. P., TANAMAS, S. K., ANDERS, K. L. & SIMMONS, C. P. 2018. Scaled deployment of Wolbachia to protect the community from dengue and other Aedes transmitted arboviruses. *Gates Open Res,* 2**,** 36.

OGUNLADE, S. T., MEEHAN, M. T., ADEKUNLE, A. I., ROJAS, D. P., ADEGBOYE, O. A. & MCBRYDE, E. S. 2021. A Review: Aedes-Borne Arboviral Infections, Controls and Wolbachia-Based Strategies. *Vaccines (Basel),* 9.

PEDERSEN, T. L. 2020. patchwork: The Composer of Plots. *R package version 1.1.1.* https://CRAN.R-project.org/package=patchwork.

PELEG, J. 1968a. Growth of arboviruses in monolayers from subcultured mosquito embryo cells. *Virology,* 35**,** 617-9.

PELEG, J. 1968b. Growth of arboviruses in primary tissue culture of Aedes aegypti embryos. *Am J Trop Med Hyg,* 17**,** 219-23.

PESTOVA, T. V. & HELLEN, C. U. 2001. Functions of eukaryotic factors in initiation of translation. *Cold Spring Harb Symp Quant Biol,* 66**,** 389-96.

PFEIFER, T. A., HEGEDUS, D. D., GRIGLIATTI, T. A. & THEILMANN, D. A. 1997. Baculovirus immediate-early promoter-mediated expression of the Zeocin resistance gene for use as a dominant selectable marker in dipteran and lepidopteran insect cell lines. *Gene,* 188**,** 183-90.

PFEIFFER, B. D., NGO, T. T., HIBBARD, K. L., MURPHY, C., JENETT, A., TRUMAN, J. W. & RUBIN, G. M. 2010. Refinement of tools for targeted gene expression in Drosophila. *Genetics,* 186**,** 735-55.

PFEIFFER, B. D., TRUMAN, J. W. & RUBIN, G. M. 2012. Using translational enhancers to increase transgene expression in Drosophila. *Proc Natl Acad Sci U S A,* 109**,** 6626-31.

PINKERTON, A. C., MICHEL, K., O'BROCHTA, D. A. & ATKINSON, P. W. 2000. Green fluorescent protein as a genetic marker in transgenic Aedes aegypti. *Insect Mol Biol,* 9**,** 1-10.

PORT, F. & BULLOCK, S. L. 2016. Augmenting CRISPR applications in Drosophila with tRNA-flanked sgRNAs. *Nat Methods,* 13**,** 852-4.

PORT, F., CHEN, H. M., LEE, T. & BULLOCK, S. L. 2014. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. *Proc Natl Acad Sci U S A,* 111**,** E2967-76.

PROMEGA 2015. Dual-Luciferase Reporter Assay System. *Instructions for use of Products E1910 and E1960.*

PROST, E., DERYCKERE, F., ROOS, C., HAENLIN, M., PANTESCO, V. & MOHIER, E. 1988. Role of the Oocyte Nucleus in Determination of the Dorsoventral Polarity of Drosophila as Revealed by Molecular Analysis of the K10-Gene. *Genes & Development,* 2**,** 891-900.

PRYCE, J., MEDLEY, N. & CHOI, L. 2022. Indoor residual spraying for preventing malaria in communities using insecticide-treated nets. *Cochrane Database Syst Rev,* 1**,** CD012688.

PRYCE, J., RICHARDSON, M. & LENGELER, C. 2018. Insecticide-treated nets for preventing malaria. *Cochrane Database Syst Rev,* 11**,** CD000363.

PULLEN, S. S. & FRIESEN, P. D. 1995. Early transcription of the ie-1 transregulator gene of Autographa californica nuclear polyhedrosis virus is regulated by DNA sequences within its 5' noncoding leader region. *J Virol,* 69**,** 156-65.

QI, L. S., LARSON, M. H., GILBERT, L. A., DOUDNA, J. A., WEISSMAN, J. S., ARKIN, A. P. & LIM, W. A. 2013. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell,* 152**,** 1173-83.

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.

RANSON, H. & LISSENDEN, N. 2016. Insecticide Resistance in African Anopheles Mosquitoes: A Worsening Situation that Needs Urgent Action to Maintain Malaria Control. *Trends Parasitol,* 32**,** 187-196.

RATNASINGHAM, S. & HEBERT, P. D. 2007. bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol Ecol Notes,* 7**,** 355-364.

REITER, P. 2001. Climate change and mosquito-borne disease. *Environ Health Perspect,* 109 Suppl 1**,** 141-61.

REN, L., PANG, D., ZHANG, M., LI, L., WANG, T., OUYANG, H. & LI, A. 2011. Comparative analysis of the activity of two promoters in insect cells. *African Journal of Biotechnology,* 10**,** 8930-8941.

RORTH, P. 1998. Gal4 in the Drosophila female germline. *Mech Dev,* 78**,** 113-8.

ROUET, P., SMIH, F. & JASIN, M. 1994. Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proc Natl Acad Sci U S A,* 91**,** 6064-8.

ROZENDAAL, J. A. 1997. *Vector Control. Methods for use by individuals and communities,* Geneva, World Health Organisation.

RTS, S. C. T. P. 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. *Lancet,* 386**,** 31-45.

SAITO, M., MANSOOR, R., KENNON, K., ANVIKAR, A. R., ASHLEY, E. A., CHANDRAMOHAN, D., COHEE, L. M., D'ALESSANDRO, U., GENTON, B., GILDER, M. E., JUMA, E., KALILANI-PHIRI, L., KUEPFER, I., LAUFER, M. K., LWIN, K. M., MESHNICK, S. R., MOSHA, D., MWAPASA, V., MWEBAZA, N., NAMBOZI, M., NDIAYE, J. A., NOSTEN, F., NYUNT, M., OGUTU, B., PARIKH, S., PAW, M. K., PHYO, A. P., PIMANPANARAK, M., PIOLA, P., RIJKEN, M. J., SRIPRAWAT, K., TAGBOR, H. K., TARNING, J., TINTO, H., VALEA, I., VALECHA, N., WHITE, N. J., WILADPHAINGERN, J., STEPNIEWSKA, K., MCGREADY, R.

& GUERIN, P. J. 2020. Efficacy and tolerability of artemisinin-based and quinine-based treatments for uncomplicated falciparum malaria in pregnancy: a systematic review and individual patient data meta-analysis. *Lancet Infect Dis,* 20**,** 943-952.

SANO, K. I., MAEDA, K., OKI, M. & MAEDA, Y. 2002. Enhancement of protein expression in insect cells by a lobster tropomyosin cDNA leader sequence. *Febs Letters,* 532**,** 143-146.

SANSBURY, B. M., HEWES, A. M. & KMIEC, E. B. 2019. Understanding the diversity of genetic outcomes from CRISPR-Cas generated homology-directed repair. *Commun Biol,* 2**,** 458.

SCHNEIDER, I. 1972. Cell lines derived from late embryonic stages of Drosophila melanogaster. *J Embryol Exp Morphol,* 27**,** 353-65.

SCHOCH, C. L., CIUFO, S., DOMRACHEV, M., HOTTON, C. L., KANNAN, S., KHOVANSKAYA, R., LEIPE, D., MCVEIGH, R., O'NEILL, K., ROBBERTSE, B., SHARMA, S., SOUSSOV, V., SULLIVAN, J. P., SUN, L., TURNER, S. & KARSCH-MIZRACHI, I. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford),* 2020.

SCHRAMM, L. & HERNANDEZ, N. 2002. Recruitment of RNA polymerase III to its target promoters. *Genes Dev,* 16**,** 2593-620.

SCHULLER, A. P. & GREEN, R. 2018. Roadblocks and resolutions in eukaryotic translation. *Nat Rev Mol Cell Biol,* 19**,** 526-541.

SCIENCES, N. A. O. Global Health Impacts of Vector-Borne Diseases. Forum on Microbial Threats; Board on Global Health, 2016 Washington (DC). National Academies of Sciences.

SERANO, T. L., CHEUNG, H. K., FRANK, L. H. & COHEN, R. S. 1994. P element transformation vectors for studying Drosophila melanogaster oogenesis and early embryogenesis. *Gene,* 138**,** 181-6.

SIMMONS, C. P., FARRAR, J. J., NGUYEN V, V. & WILLS, B. 2012. Dengue. *N Engl J Med,* 366**,** 1423-32.

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 Res,* 42**,** 7461-72.

SINGH, K. R. & PAVRI, K. M. 1967. Experimental studies with chikungunya virus in Aedes aegypti and Aedes albopictus. *Acta Virol,* 11**,** 517-26.

SINKINS, S. P. 2004. Wolbachia and cytoplasmic incompatibility in mosquitoes. *Insect Biochem Mol Biol,* 34**,** 723-9.

SMITH, G. E., SUMMERS, M. D. & FRASER, M. J. 1983. Production of human beta interferon in insect cells infected with a baculovirus expression vector. *Mol Cell Biol,* 3**,** 2156-65.

SMITH, R. C., VEGA-RODRIGUEZ, J. & JACOBS-LORENA, M. 2014. The Plasmodium bottleneck: malaria parasite losses in the mosquito vector. *Mem Inst Oswaldo Cruz,* 109**,** 644-61.

SNOW, R. W., GUERRA, C. A., NOOR, A. M., MYINT, H. Y. & HAY, S. I. 2005. The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature,* 434**,** 214-7.

SPICKLER, A. R. 2016. *Screwworm Myiasis* [Online]. The Center for Food Security & Public Health Available: cfsph.iastate.edu/diseaseinfo/factsheets/ [Accessed November 2021 2021].

STAPLES, J. E., GERSHMAN, M., FISCHER, M., CENTERS FOR DISEASE, C. & PREVENTION 2010. Yellow fever vaccine: recommendations of the Advisory Committee on Immunization Practices (ACIP). *MMWR Recomm Rep,* 59**,** 1-27.

STEWART, M. 2010. Nuclear export of mRNA. *Trends Biochem Sci,* 35**,** 609-17.

SUZUKI, T., ITO, M., EZURE, T., KOBAYASHI, S., SHIKATA, M., TANIMIZU, K. & NISHIMURA, O. 2006. Performance of expression vector, pTD1, in insect cell-free translation system. *J Biosci Bioeng,* 102**,** 69-71.

TAMURA, T., THIBERT, C., ROYER, C., KANDA, T., ABRAHAM, E., KAMBA, M., KOMOTO, N., THOMAS, J. L., MAUCHAMP, B., CHAVANCY, G., SHIRK, P., FRASER, M., PRUDHOMME, J. C. & COUBLE, P. 2000. Germline transformation of the silkworm Bombyx mori L. using a piggyBac transposon-derived vector. *Nat Biotechnol,* 18**,** 81-4.

TATEMATSU, K., UCHINO, K., SEZUTSU, H. & TAMURA, T. 2014. Effect of ATG initiation codon context motifs on the efficiency of translation of mRNA derived from exogenous genes in the transgenic silkworm, Bombyx mori. *Springerplus,* 3**,** 136.

THEILMANN, D. A. & STEWART, S. 1992. Molecular analysis of the trans-activating IE-2 gene of Orgyia pseudotsugata multicapsid nuclear polyhedrosis virus. *Virology,* 187**,** 84-96.

THOMAS, S. J. & YOON, I. K. 2019. A review of Dengvaxia(R): development to deployment. *Hum Vaccin Immunother,* 15**,** 2295-2314.

TNG, P. Y. L., CARABAJAL PALADINO, L., VERKUIJL, S. A. N., PURCELL, J., MERITS, A., LEFTWICH, P. T., FRAGKOUDIS, R., NOAD, R. & ALPHEY, L. 2020. Cas13b-dependent and Cas13b-independent RNA knockdown of viral sequences in mosquito cells following guide RNA expression. *Commun Biol,* 3**,** 413.

TOLLE, M. A. 2009. Mosquito-borne diseases. *Curr Probl Pediatr Adolesc Health Care,* 39**,** 97-140.

TRAVERSA, D. 2013. Fleas infesting pets in the era of emerging extra-intestinal nematodes. *Parasit Vectors,* 6**,** 59.

TURELLI, M. 2010. Cytoplasmic incompatibility in populations with overlapping generations. *Evolution,* 64**,** 232-41.

UNCKLESS, R. L., CLARK, A. G. & MESSER, P. W. 2017. Evolution of Resistance Against CRISPR/Cas9 Gene Drive. *Genetics,* 205**,** 827-841.

VAN DEN BERG, H., DA SILVA BEZERRA, H. S., AL-ERYANI, S., CHANDA, E., NAGPAL, B. N., KNOX, T. B., VELAYUDHAN, R. & YADAV, R. S. 2021. Recent trends in global insecticide use for disease vector control and potential implications for resistance management. *Sci Rep,* 11**,** 23867.

VAN OERS, M. M., PIJLMAN, G. P. & VLAK, J. M. 2015. Thirty years of baculovirus-insect cell protein expression: from dark horse to mainstream technology. *Journal of General Virology,* 96**,** 6-23.

VAN OERS, M. M., VLAK, J. M., VOORMA, H. O. & THOMAS, A. A. M. 1999. Role of the 3' untranslated region of baculovirus p10 mRNA in high-level expression of foreign genes. *J Gen Virol,* 80 (Pt 8)**,** 2253-2262.

VANNICE, K. S., HILLS, S. L., SCHWARTZ, L. M., BARRETT, A. D., HEFFELFINGER, J., HOMBACH, J., LETSON, G. W., SOLOMON, T., MARFIN, A. A. & JAPANESE ENCEPHALITIS VACCINATION EXPERTS, P. 2021. The future of Japanese encephalitis vaccination: expert recommendations for achieving and maintaining optimal JE control. *NPJ Vaccines,* 6**,** 82.

VAUGHN, J. L., GOODWIN, R. H., TOMPKINS, G. J. & MCCAWLEY, P. 1977. The establishment of two cell lines from the insect Spodoptera frugiperda (Lepidoptera; Noctuidae). *In Vitro,* 13**,** 213-7.

VEGA-RUA, A., ZOUACHE, K., GIROD, R., FAILLOUX, A. B. & LOURENCO-DE-OLIVEIRA, R. 2014. High level of vector competence of Aedes aegypti and Aedes albopictus from ten American countries as a crucial factor in the spread of Chikungunya virus. *J Virol,* 88**,** 6294-306.

VITOR, A. C., HUERTAS, P., LEGUBE, G. & DE ALMEIDA, S. F. 2020. Studying DNA Double-Strand Break Repair: An Ever-Growing Toolbox. *Front Mol Biosci,* 7**,** 24.

VLAK, J. M., SCHOUTEN, A., USMANY, M., BELSHAM, G. J., KLINGE-ROODE, E. C., MAULE, A. J., VAN LENT, J. W. & ZUIDEMA, D. 1990. Expression of cauliflower mosaic virus gene I using a baculovirus vector based upon the p10 gene and a novel selection method. *Virology,* 179**,** 312-20.

VOLOHONSKY, G., TERENZI, O., SOICHOT, J., NAUJOKS, D. A., NOLAN, T., WINDBICHLER, N., KAPPS, D., SMIDLER, A. L., VITTU, A., COSTA, G., STEINERT, S., LEVASHINA, E. A., BLANDIN, S. A. & MAROIS, E. 2015. Tools for Anopheles gambiae Transgenesis. *G3 (Bethesda),* 5**,** 1151-63.

WAKIYAMA, M., MATSUMOTO, T. & YOKOYAMA, S. 2005. Drosophila U6 promoter-driven short hairpin RNAs effectively induce RNA interference in Schneider 2 cells. *Biochem Biophys Res Commun,* 331**,** 1163-70.

WALKER, T., JOHNSON, P. H., MOREIRA, L. A., ITURBE-ORMAETXE, I., FRENTIU, F. D., MCMENIMAN, C. J., LEONG, Y. S., DONG, Y., AXFORD, J., KRIESNER, P., LLOYD, A. L., RITCHIE, S. A., O'NEILL, S. L. & HOFFMANN, A. A. 2011. The wMel Wolbachia strain blocks dengue and invades caged Aedes aegypti populations. *Nature,* 476**,** 450-3.

WANG, J., CHEN, R., ZHANG, R., DING, S., ZHANG, T., YUAN, Q., GUAN, G., CHEN, X., ZHANG, T., ZHUANG, H., NUNES, F., BLOCK, T., LIU, S., DUAN, Z., XIA, N., XU, Z. & LU, F. 2017. The gRNA-miRNA-gRNA Ternary Cassette Combining CRISPR/Cas9 with RNAi Approach Strongly Inhibits Hepatitis B Virus Replication. *Theranostics,* 7**,** 3090-3105.

WEBSTER, S. H. & SCOTT, M. J. 2021. The Aedes aegypti (Diptera: Culicidae) hsp83 Gene Promoter Drives Strong Ubiquitous DsRed and ZsGreen Marker Expression in Transgenic Mosquitoes. *J Med Entomol,* 58**,** 2533-2537.

WEST, S. C., BLANCO, M. G., CHAN, Y. W., MATOS, J., SARBAJNA, S. & WYATT, H. D. 2015. Resolution of Recombination Intermediates: Mechanisms and Regulation. *Cold Spring Harb Symp Quant Biol,* 80**,** 103-9.

WHO. 2017. *Global Vector Control Response 2017 - 2030: A strategic approach to tackle vector-borne diseases* [Online]. Available: https://www.who.int/publications/i/item/WHO-HTM-GVCR-2017.01 [Accessed Nov 2021 2021].

WHO 2019a. Dengue vaccine: WHO position paper, September 2018 - Recommendations. *Vaccine,* 37**,** 4848-4849.

WHO 2019b. Japanese encephalitis. *Fact sheets detail.* https://www.who.int/news-room/fact-sheets/detail/japanese-encephalitis: World Health Organisation.

WHO 2020. Vector-borne diseases. *WHO Fact sheets*.

WHO 2021a. Guidance framework for testing genetically modified mosquitoes, second edition.

WHO 2021b. WHO Guidelines for malaria. *In:* ORGANISATION, W. H. (ed.). Geneva: WHO.

WHO 2021c. World Malaria Report *In:* WHO (ed.).

WHO. 2022a. *Consolidated Guidelines for malaria* [Online]. Available: https://www.who.int/teams/global-malaria-programme/guidelines-for-malaria [Accessed 30/04/2022].

WHO 2022b. Dengue and severe dengue. *Fact sheets detail.* https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue: World Health Organisation.

WICKHAM, H. 2016. ggplot2 : Elegant Graphics for Data Analysis. *Use R!,*. 2nd ed. Cham: Springer International Publishing : Imprint: Springer,.

WICKHAM, H., AVERICK, M., BRYAN, J., CHANG, W., MCGOWAN, L., FRANÇOIS, R., GROLEMUND, G., HAYES, A., HENRY, L., HESTER, J., KUHN, M., PEDERSEN, T., MILLER, E., BACHE, S., MÜLLER, K., OOMS, J., ROBINSON, D., SEIDEL, D., SPINU, V. & YUTANI, H. 2019. Welcome to the Tidyverse. *Journal of Open Source Software,* 4**,** 1686.

WILKE, A. B., SCAIFE, S., ALPHEY, L. & MARRELLI, M. T. 2013. DsRed2 transient expression in Culex quinquefasciatus mosquitoes. *Mem Inst Oswaldo Cruz,* 108**,** 529-31.

WORLD_BANK. 2014. *Poverty and Health* [Online]. Available: https://www.worldbank.org/en/topic/health/brief/poverty-health [Accessed Nov 2021 2021].

XIE, C., CHEN, Y. L., WANG, D. F., WANG, Y. L., ZHANG, T. P., LI, H., LIANG, F., ZHAO, Y. & ZHANG, G. Y. 2017. SgRNA Expression of CRIPSR-Cas9 System Based on MiRNA Polycistrons as a Versatile Tool to Manipulate Multiple and Tissue-Specific Genome Editing. *Sci Rep,* 7**,** 5795.

XIE, K., MINKENBERG, B. & YANG, Y. 2015. Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proc Natl Acad Sci U S A,* 112**,** 3570-5.

YAMADA, H., MAIGA, H., JUAREZ, J., DE OLIVEIRA CARVALHO, D., MAMAI, W., ALI, A., BIMBILE-SOMDA, N. S., PARKER, A. G., ZHANG, D. & BOUYER, J. 2019. Identification of critical factors that significantly affect the dose-response in mosquitoes irradiated as pupae. *Parasit Vectors,* 12**,** 435.

YAN, Q., XU, K., XING, J., ZHANG, T., WANG, X., WEI, Z., REN, C., LIU, Z., SHAO, S. & ZHANG, Z. 2016. Multiplex CRISPR/Cas9-based genome engineering enhanced by Drosha-mediated sgRNA-shRNA structure. *Sci Rep,* 6**,** 38970.

YAZBECK, A. M., TOUT, K. R. & STADLER, P. F. 2018. Detailed secondary structure models of invertebrate 7SK RNAs. *RNA Biol,* 15**,** 158-164.

ZALUCKI, M. P., SHABBIR, A., SILVA, R., ADAMSON, D., SHU-SHENG, L. & FURLONG, M. J. 2012. Estimating the economic cost of one of the world's major insect pests, Plutella xylostella (Lepidoptera: Plutellidae): just how long is a piece of string? *J Econ Entomol,* 105**,** 1115-29.

ZHOU, Z., DANG, Y., ZHOU, M., LI, L., YU, C. H., FU, J., CHEN, S. & LIU, Y. 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc Natl Acad Sci U S A,* 113**,** E6117-E6125.

ZIELER, H. & HUYNH, C. Q. 2002. Intron-dependent stimulation of marker gene expression in cultured insect cells. *Insect Mol Biol,* 11**,** 87-95.

ZUG, R. & HAMMERSTEIN, P. 2012. Still a host of hosts for Wolbachia: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. *PLoS One,* 7**,** e38544.

Blank Page