

# UC Davis

## UC Davis Previously Published Works

### Title

Abundance of conserved CRISPR-Cas9 target sites within the highly polymorphic genomes of Anopheles and Aedes mosquitoes

### Permalink

<https://escholarship.org/uc/item/4qh9n1m0>

### Journal

Nature Communications, 11(1)

### ISSN

2041-1723

### Authors

Schmidt, Hanno

Collier, Travis C

Hanemaaijer, Mark J

et al.

### Publication Date

2020

### DOI




10.1038/s41467-020-15204-0

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

# Abundance of conserved CRISPR-Cas9 target sites within the highly polymorphic genomes of *Anopheles* and *Aedes* mosquitoes

Hanno Schmidt <sup>1</sup>, Travis C. Collier<sup>1</sup>, Mark J. Hanemaaijer<sup>1,2</sup>, Parker D. Houston <sup>1</sup>, Yoosook Lee<sup>1</sup> & Gregory C. Lanzaro <sup>1</sup>✉

A number of recent papers report that standing genetic variation in natural populations includes ubiquitous polymorphisms within target sites for Cas9-based gene drive (CGD) and that these “drive resistant alleles” (DRA) preclude the successful application of CGD for managing these populations. Here we report the results of a survey of 1280 genomes of the mosquitoes *Anopheles gambiae*, *An. coluzzii*, and *Aedes aegypti* in which we determine that ~90% of all protein-encoding CGD target genes in natural populations include at least one target site with no DRAs at a frequency of  $\geq 1.0\%$ . We conclude that the abundance of conserved target sites in mosquito genomes and the inherent flexibility in CGD design obviates the concern that DRAs present in the standing genetic variation of mosquito populations will be detrimental to the deployment of this technology for population modification strategies.

<sup>1</sup>Vector Genetics Laboratory, Department of Pathology, Microbiology, and Immunology, School of Veterinary Medicine, University of California, Davis, CA 95616, USA. <sup>2</sup>Present address: Winlove Probiotics, Hulstweg 11, 1032 LB Amesterdam, Netherlands. ✉email: [gclanzaro@ucdavis.edu](mailto:gclanzaro@ucdavis.edu)

The discovery of clustered regularly interspaced short palindromic repeats (CRISPR) in bacteria<sup>1,2</sup> and the Cas9 enzyme (CRISPR associated protein 9)<sup>3–5</sup>, has revolutionized our capacity to genetically engineer a wide range of organisms. The subsequent development of CRISPR-Cas9-based gene drives<sup>6</sup> has further increased the potential application of this technology. Gene drives promote the spread of introduced genetic elements (e.g., alternative alleles, exogenous genes) through populations by altering the way in which they are inherited, such that the desired genetic element is over-represented among progeny (“Super-Mendelian inheritance”)<sup>7</sup>. This leads to an increase in frequency of the introduced genetic element, potentially until fixation in the targeted population.

One application of CRISPR-Cas9 gene drive that has gained a great deal of attention is the possibility of controlling populations of disease vectors like mosquitoes. The focus of current efforts is *Anopheles gambiae* and *An. coluzzii* which transmit malaria, and *Aedes aegypti* which transmits dengue, chikungunya, yellow fever, and Zika. Collectively, these diseases cause hundreds of thousands of human deaths per year<sup>8</sup>. New strategies for controlling these vectors are sorely needed because currently available control methods are costly, increasingly ineffective due to insecticide resistance<sup>9</sup> and are generally difficult to deploy in rural endemic areas. Alternative genetic-based strategies for vector control are not new, however, the recent advances in genetic engineering and gene drive have sparked increased interest in this approach. There are two broad categories of strategies involving genetically engineered mosquitoes (GEM) with gene drive currently under development: population suppression aimed at greatly reducing or eliminating the mosquito population<sup>10</sup> and population modification, which renders mosquitoes incapable of transmitting a pathogen but otherwise leaves it unaltered<sup>11</sup>. Recently, CRISPR-Cas9-based gene drive systems have been designed for population modification in *Anopheles*<sup>12</sup> and *Aedes*<sup>13</sup> and for population suppression in *Anopheles*<sup>10,14</sup> mosquitoes.

Experiments demonstrating the capacity of gene-drive constructs to spread through wild-type populations in laboratory cages have yielded promising results<sup>10</sup>. A major limitation of these experiments is that they use populations of mosquitoes derived from long-standing laboratory colonies that do not replicate populations as they occur in nature<sup>15,16</sup>. Specifically, founder effects during establishment, repeated bottlenecks experienced during maintenance, and selection for adaptation to the laboratory environment in these colonies all result in the loss of genetic variability relative to their counterparts in nature<sup>17–19</sup>.

Recently, several population genomic studies have amassed a large volume of genomic data from natural populations of *An. gambiae*<sup>20,21</sup>, *An. coluzzii*<sup>21</sup>, and *Ae. aegypti*<sup>22</sup>. These surveys revealed exceptionally high levels of genetic variability leading some authors to warn that CRISPR-Cas9-based gene-drive systems (CGD) may be prone to failure due to drive resistance resulting from standing genetic variation. This includes unclonable alleles within the target sequence that are not recognized by the guide RNA<sup>21,23</sup>. A study of the impact of drive resistance alleles (DRAs) on the performance of CGD in natural populations of the flour beetle, *Tribolium castaneum*, concluded that population-specific rare alleles will probably reduce or eliminate drive efficacy<sup>24</sup>. General modeling approaches revealed that standing genetic variation could even exceed de novo mutations in contributing to CGD resistance<sup>25</sup>. Given the interest in the development of CGD, a systematic evaluation of the distribution of polymorphisms within the genomes of these critical mosquito species and its impact on potential target sites for CRISPR-Cas9 editing is warranted.

Here we present genome-wide screens of the three principal human disease vector species *An. gambiae*, *An. coluzzii*, and *Ae.*

*aegypti* for the presence of CRISPR-Cas9 target sites and an analysis of the degree of polymorphism therein. In detail, we search all transcribed regions of protein-coding genes in the species’ reference genomes for potential CRISPR-Cas9 target sites. We then subject each target site to a screen for nucleotide polymorphisms (single nucleotide polymorphisms, insertions, deletions) in the genomes of mosquitoes sampled directly from natural populations. Our analyses include 111 *An. gambiae*, 100 *An. coluzzii*, and 132 *Ae. aegypti* genomes from our lab plus publicly available polymorphism data from 937 additional *An. gambiae s.l.* samples. The special interest in *An. gambiae* as the principal vector of malaria in Africa results in a larger number of individual mosquito sequence data compared with any other mosquito species. Additional insights gained from including the larger number of sequences compared with *An. coluzzii* and *Ae. aegypti* outweigh the benefits of having equal numbers per species. We find that >30% of protein-coding genes have potential CRISPR-Cas9 targets with GC content between 30 and 70% and no off-target sequence. This drops to 8.4% if sites with DRAs at frequencies >1% in natural populations are excluded. Nonetheless ~90% of all protein-coding genes contain at least one target site that remain after this filtering. Based on these observations we conclude that DRAs within the standing variation that exists in natural populations of the mosquito species studied will not pose a problem to the successful deployment of CRISPR-Cas9-based gene drive for population modification strategies. Gene drive used as part of population suppression strategies are more likely to be unsustainable because of the presence of low-frequency DRAs and the fact that they impose much stronger selection favoring them.

## Results

**Identifying potential CRISPR-Cas9 target sites.** We began our analysis by identifying all potential CRISPR-Cas9 guide RNA (gRNA) target sites in each species’ genomes and subjecting each to an analysis to identify DRAs. We define potential target sites as 23 bp stretches with the nucleotides ‘NGG’ at one 3’-end (NGG = protospacer adjacent motif, PAM), located in a transcript of a protein-coding gene. To make the analysis more conservative, we restricted our search to target sites with a GC content between 30 and 70% and no close (<4 mismatches) sequence matching anywhere else in the genome that could produce off-target activity. The total number of potential target sites was estimated by screening the latest versions of the publicly available reference genomes of *An. gambiae* (AgamP4) and *Ae. aegypti* (AeagL4) using the program CHOPCHOP<sup>26–28</sup>. The AgamP4 genome is sequenced from an *An. gambiae*–*An. coluzzii*-hybrid laboratory strain and is suitable as a reference for both, *An. coluzzii* and *An. gambiae*<sup>29</sup>.

We identified 1,196,509 high-quality potential targets in the genome of *An. gambiae s.l.* and 828,454 for *Ae. aegypti* (Table 1). While 69.5% (*An. gambiae s.l.*) and 77.2% (*Ae. aegypti*) of the raw target sites were dismissed during quality filtering, the overwhelming majority of coding genes contain at least one potential CRISPR-Cas9 target site (97.2% in *An. gambiae s.l.* and 92.2% in *Ae. aegypti*, Fig. 1).

**DRA frequencies in natural mosquito populations.** We conducted screening of potential target sites for DRAs using five unfiltered datasets of nucleotide polymorphisms from natural mosquito populations (Table 2) from our lab (Vector Genetics Laboratory; VGL) and The *Anopheles gambiae* 1000 Genomes Consortium (Ag1000G). The nucleotide diversity ( $\pi$ ) of transcripts in protein-coding genes was between 0.94 and 1.02%

(Table 2, Supplementary Table 1). For comparison, a relatively high estimate of human nucleotide diversity is 0.6%<sup>30</sup>.

**DRA frequencies and abundance of good CRISPR-Cas9 targets.** We define a “good” CRISPR-Cas9 target site as a potential target which contains no DRAs above a predefined DRA

threshold frequency. Sample size limits our ability to detect DRAs below a certain frequency. If we set the threshold frequency at 0.00 the proportion of good targets is highly dependent on sample size, with an ordinary least-squares coefficient of determination ( $r^2$ ) of 0.99 (Fig. 2, orange bars). However, setting the DRA threshold at 0.01 essentially eliminates this sampling effect ( $r^2$  of 0.00) If this threshold (<1%) is applied to all five datasets we find ~90% of protein-coding genes contain at least one good target (Fig. 2, blue bars).

The relationship between the DRA threshold frequency and the percentage of genes containing at least one good target is illustrated in detail for the Ag1000G *An. gambiae* dataset ( $N = 654$ ) (Fig. 3). The chance that any specific target site will be free of DRAs is much lower than the chance that a given gene will contain at least one good target. Only 28.2% of specific targets are free from DRAs with  $\geq 1\%$  frequency, dropping to 6.3% for variants with a frequency of 0.15%. Less than 3% of potential targets are completely free of observed DRAs.

The fraction of protein-coding genes containing at least one good target is largely independent of the DRA frequency threshold down to a threshold of ~1%, at which point it drops steeply (Fig. 3, blue line). For example, 91.6% of protein-coding genes contain a good target, with no DRAs at frequencies  $\geq 1\%$ . However, only 58.5% of protein-coding genes contain at least one good target when the DRA threshold is set at 0.15%. The fraction of good targets among all potential targets is far more sensitive to the DRA frequency threshold, declining steadily as the DRA frequency threshold is decreased (Fig. 3, orange line).

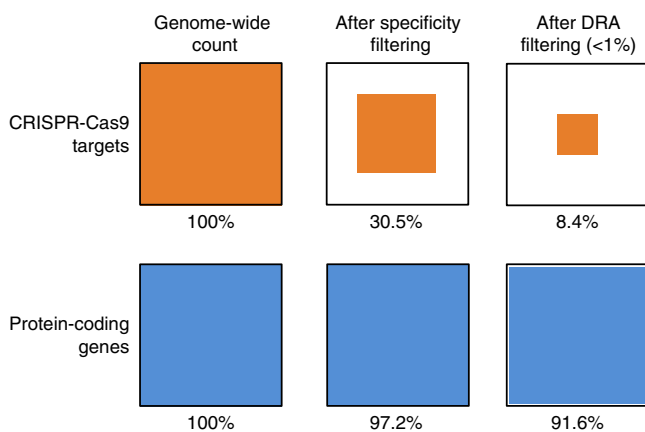
**Discussion**

In this study we confirm what has been widely reported that mosquitoes in the genera *Aedes* and *Anopheles* have genomes that are highly polymorphic<sup>21,22</sup>. The suggestion that standing genetic variation will render Cas9-based gene drives ineffective in natural mosquito populations seems to be plausible. Indeed, <3% of potential high-quality Cas9 targets are free from any observed variation and increasing the sample size analyzed will only reduce this value. However, protein-coding genes almost always contain many potential targets. The median number of potential targets per coding gene for *An. gambiae* and *An. coluzzii* is 72 and for *Ae. aegypti* it is 47. Even if there is only a 3% chance that any individual target will be a good target, with 72 options the chance that at least one will be good is 89% (Fig. 1). Consequently, our analyses show that ~90% of all protein-coding genes have conserved target sites for CRISPR-Cas9 editing in all three species examined. The broad similarity of results between the two *Anopheles* species and *Ae. aegypti*, which has quite different genomic characteristics, suggests the observed pattern represents a general principle. This could be tested in the future by examining population genomic data from additional species.

**Table 1 Target sites for CRISPR-Cas9 editing in mosquito genomes.**

	<i>Anopheles gambiae</i> s.l. (AgamP4.11)	<i>Aedes aegypti</i> (AaegL5.1)
Genome size (Mbp)	230.5	1195
Coding part of the genome (Mbp)	25.7 (11.2%)	59.1 (5%)
Raw targets	3,918,579	3,638,628
Potential targets	1,196,509 (30.5%)	828,454 (22.8%)
Protein-coding genes	12,562	13,601
Protein-coding genes with potential targets	12,213 (97.2%)	12,536 (92.2%)

Raw targets are all unique target sites suitable for CRISPR-Cas9 editing found in transcripts of protein-coding genes. Potential targets refer to sites that passed filtering for off-target effects and GC content. The same reference genome (AgamP4.11) was used for *A. gambiae* and *A. coluzzii*.  
Mbp mega base pairs.

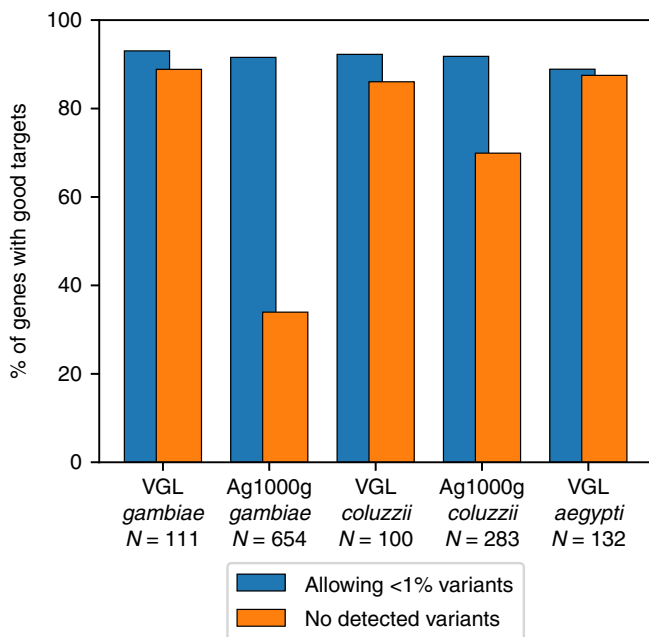


**Fig. 1 Sketch of the effect of quality filtering on the number of “good” targets/genes.** Genome-wide count of CRISPR-Cas9 targets (orange) and protein-coding genes (blue) is set to 100% each. During specificity filtering (GC content between 30 and 70% and no off-targets) and DRA filtering (DRA frequency <1%), the number of available targets drops well below 10% (Table 2). Nevertheless, ~90% of all protein-coding genes still contain at least one good target. Colored areas correspond to the values for the combined data of *An. gambiae* and *An. coluzzii*. The percentages are similar for *Ae. aegypti* (not shown). Source data are provided as a Source Data file.

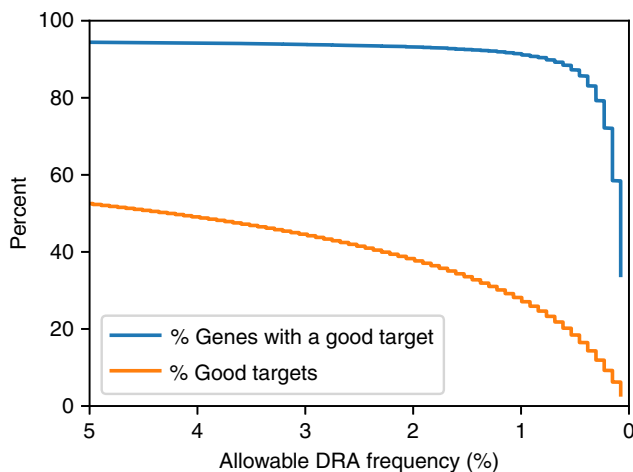
**Table 2 Effect of polymorphisms on potential targets.**

	<i>An. gambiae</i> VGL	<i>An. gambiae</i> Ag1000G	<i>An. coluzzii</i> VGL	<i>An. coluzzii</i> Ag1000G	<i>Ae. aegypti</i> VGL
Samples	111	654	100	283	132
Nucleotide diversity ( $\pi$ ) in transcribed regions of protein-coding genes	0.98%	1.02%	1%	0.95%	0.94%
Good targets (% of raw targets/% of potential targets)	216,793 (5.5%/18.1%)	34,995 (0.9%/2.9%)	174,117 (4.4%/14.6%)	93,142 (2.4%/7.8%)	273,627 (7.5%/33%)
Protein-coding genes with good targets, i.e., no variation at all (% of all protein-coding genes)	11,163 (88.9%)	4281 (34.1%)	10,832 (86.2%)	8796 (70%)	12,536 (87.5%)
Protein-coding genes with good targets, i.e., no variation at frequencies $\geq 1\%$ (% of all protein-coding genes)	11,721 (93.3%)	11,525 (91.3%)	11,415 (90.9%)	11,540 (91.9%)	12,096 (88.9%)

The variant data comprises unfiltered calls for the three mosquito species. Sequence data was taken from the UC Davis Vector Genetics Lab archive. We also included publicly available variant data for *An. gambiae* and *An. coluzzii* from the Ag1000G project. Good target sites are potential target sites that have been additionally filtered for variant data, i.e., target sites that are suitable for gene editing with a high probability of showing no resistance alleles in natural populations.



**Fig. 2 Frequency of genes with good targets.** The fraction of genes with good targets is dependent on the presence of low-frequency DRAs. When dismissing all targets with DRA frequencies > 0.0, the fraction of genes with good targets decreases with increasing sample size. Ignoring DRAs with frequencies below 1% in the dataset results in ~90% of genes having at least one good target in all datasets examined. VGL: Vector Genetics Laboratory, Ag1000G: The *Anopheles gambiae* 1000 Genomes Consortium. Source data are provided as a Source Data file.



**Fig. 3 Effects of polymorphisms on targets.** Percent of genes containing at least one good target (blue) and percent of good targets out of possible targets as a function of DRA frequency threshold (orange). This analysis is based on  $N = 654$  *An. gambiae* samples (Ag1000G data). DRA frequency threshold is the value beyond which alternative alleles are considered to be DRAs and are filtered out (i.e., a DRA frequency threshold of 0.01 means, alternative alleles with a frequency below 1% are ignored during filtering). Note the constant decline in the fraction of good targets (orange), which is not mirrored in the fraction of genes containing good targets (blue) until the DRA frequency threshold is set at <0.01. Source data are provided as a Source Data file.

These results have implications for evaluating the prospects of population modification versus population suppression strategies. GEMs for population suppression are designed to eliminate fertile females from a target population and so transgenic individuals

obviously have extremely low fitness<sup>31</sup>. Very strong selection favoring a wild-type genotype may be to some extent countered by the self-replicating gene drive. However, every individual with a genotype that includes a DRA will be subjected to strong positive selection. Therefore, even low-frequency DRAs (including private nucleotide polymorphisms) pose a high risk to population suppression strategies, since DRAs would rapidly increase in frequency<sup>32,33</sup>. This translates to establishing a DRA threshold near zero. As we demonstrate here, this will reduce the number of protein-coding genes that are useful candidates for genetic engineering (Fig. 3), especially in large natural populations as depicted by the sample size dependence in fraction of genes with targets having no polymorphisms at all (Fig. 2). A scenario where this might be useful is the targeting of small, defined populations, where spillover to neighboring populations is unwanted and can be avoided by the intended use of alleles that are fixed in the target population but absent in the neighbor population<sup>34,35</sup>. While this can also be applied with population modification strategies, these also can be specifically designed to have a negligible fitness cost relative to wild-type<sup>12</sup>. In this case, low frequency wild-type genotypes that include a DRA should not affect the gene drive behavior detrimentally, since they would likely remain at low frequencies or be eliminated by drift<sup>36</sup>. Even highly efficient gene-drive systems generate DRAs at frequencies of ~1%, thus choosing this value as a threshold for standing variation is justified from the point of view that such a level of DRAs would have only marginal effects of inherent gene-drive performance.

When seeking to design a GEM with CGD for release into a natural population, researchers will most likely consider target sites excluded by the stringent filtering applied in this study. For example, we did not apply quality filtering of polymorphic sites in order to be as conservative as possible. In the planning of a GEM design study, researchers would clearly make an effort to exclude false positive calls, thereby reducing the total number/density of DRAs. Also, a single polymorphic site distant from the PAM (e.g., >–10 bp from PAM) is not expected to have dramatic effects on cleavage efficiencies. Moreover, different nucleotide positions in CRISPR-Cas9 target sites have quite unequal effects on cleavability<sup>24,37</sup>. In summary, when evaluating a specific candidate gene for GEM design, a much larger number of target sites for CRISPR-Cas9 editing could be considered for empirical evaluation.

The extensive amount of data analyzed for three of the most important human disease vector species in large parts of their distributional area present an unprecedented view of the feasibility of CRISPR-Cas9-based gene drives in mosquitoes. The results demonstrate that good target sites lacking DRAs or with DRAs present at low frequency are abundant in the three species studied. The abundance of good target sites in mosquito genomes and the inherent flexibility in CRISPR-Cas9-based gene-drive design suggests that drive resistance arising from selection on standing genetic variation will not be a detriment to the deployment of this technology for eliminating mosquito-borne diseases.

## Methods

**Search for potential CRISPR-Cas9 target sites.** Potential CRISPR-Cas9 target sites were searched with the command line version of CHOPCHOP v6054ae8b29b9<sup>26–28</sup> with Python v.2.7.15<sup>38</sup>, applying default settings and ‘Xu\_2015’ efficiency scoring<sup>39</sup>. We modified CHOPCHOP slightly to fix a minor bug in the handling of chromosome names and to increase the maximum target (transcript) size. We restricted the search to the transcripts (coding sequences + untranslated regions) of protein-coding genes from the most recent annotation files downloaded from vectorbase.org, with 12,562 entries for protein-coding genes in *An. gambiae* (AgamP4.11) and 13,601 for *Ae. aegypti* (AaegL5.1). The output was filtered for targets that show no off-target sites with less than four mismatches to the original sequence and that have a GC content between 30 and 70%. We denote CRISPR-Cas9 target sites that passed this procedure as “potential target sites”.

**Anopheles gambiae s.l. data preparation.** We used individual whole genome sequencing data from  $N = 111$  *An. gambiae* s.s. samples from natural populations in Mali ( $N = 40$ ), Cameroon ( $N = 5$ ), Tanzania ( $N = 6$ ), Zambia ( $N = 6$ ), and the Comoro Islands ( $N = 54$ ) and  $N = 100$  *An. coluzzii* samples from natural populations in Mali ( $N = 66$ ), Benin ( $N = 11$ ), Equatorial Guinea ( $N = 3$ ), Cameroon ( $N = 1$ ), and São Tomé and Príncipe ( $N = 19$ ). Specimens for sequencing were taken from the Vector Genetics Laboratory's archive (Supplementary Data 1). Genomic DNA was sequenced on an Illumina HiSeq 4000 to a mean depth of 10.2X (Supplementary Table 2). Sequences were filtered for adapters with Trimmomatic v0.36<sup>40</sup> and then mapped against the most current reference genome assembly 'AgamP4'<sup>41</sup> using BWA MEM v0.7.17-r1188<sup>42</sup>. Polymorphic sites were called with FreeBayes v1.2.0<sup>43</sup> applying default parameters but 'theta=0.01' and 'max-complex-gap=3'.

The Ag1000G datasets were processed by MalariaGen ([www.malariagen.net](http://www.malariagen.net)) using BWA-MEM mapping and GATK UnifiedGenotyper<sup>44</sup> workflow<sup>45</sup>.

**Aedes aegypti data preparation.** We used individual whole genome data from  $N = 132$  *Ae. aegypti* samples from California ( $N = 122$ ), Florida ( $N = 4$ ), Mexico ( $N = 3$ ), and South Africa ( $N = 3$ ) from the Vector Genetics Laboratory's archive (Supplementary Data 1) sequenced to a mean depth of 10.5X (Supplementary Table 2). Sequences were generated and data were processed in the same way as described above for *Anopheles* samples, using the most current reference genome assembly 'AaegL5'<sup>46</sup>.

**Polymorphisms in potential CRISPR-Cas9 target sites.** We did not include any subsequent quality filtering of detected polymorphisms to ensure having the broadest possible set of potential polymorphic sites and hence being as conservative as possible (i.e., removing all doubtful/undecidable potential CRISPR-Cas9 target sites). The potential CRISPR-Cas9 target sites were then filtered for polymorphic sites using custom scripts available at [https://github.com/travc/Cas9\\_target\\_site\\_survey](https://github.com/travc/Cas9_target_site_survey).

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Sequence data sources are detailed in Supplemental Table S1. Data processing scripts and small datafiles are available in GitHub with the identifier: [<https://doi.org/10.5281/zenodo.3661448>]<sup>47</sup>. Any additional data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 18 September 2019; Accepted: 21 February 2020;

Published online: 18 March 2020

## References

- Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
- Jansen, R., Embden, J. D. v., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).
- Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
- Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
- Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
- Gantz, V. M. & Bier, E. The mutagenic chain reaction: a method for converting heterozygous to homozygous mutations. *Science* **348**, 442–444 (2015).
- Champer, J., Buchman, A. & Akbari, O. S. Cheating evolution: engineering gene drives to manipulate the fate of wild populations. *Nat. Rev. Genet.* **17**, 146–159 (2016).
- WHO. World Malaria Report 2017 (2017).
- Ranson, H. & Lissenden, N. Insecticide resistance in African *Anopheles* mosquitoes: a worsening situation that needs urgent action to maintain malaria control. *Trends Parasitol.* **32**, 187–196 (2016).
- Kyrou, K. et al. A CRISPR-Cas9 gene drive targeting doublesex causes complete population suppression in caged *Anopheles gambiae* mosquitoes. *Nat. Biotechnol.* **36**, 1062 (2018).
- Macias V. & James A. A. In *Genetic Control of Malaria and Dengue* (ed. Adelman, Z. N.) (Elsevier, 2016).
- Gantz, V. M. et al. Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi*. *Proc. Natl Acad. Sci. USA* **112**, E6736–E6743 (2015).
- Li, M. et al. Development of a confinable gene-drive system in the human disease vector, *Aedes aegypti*. *eLife* **9** (2020).
- Hammond, A. et al. A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*. *Nat. Biotechnol.* **34**, 78–83 (2016).
- Boëte, C. *Anopheles* mosquitoes: not just flying malaria vectors... especially in the field. *Trends Parasitol.* **25**, 53–55 (2009).
- Baeshen, R. et al. Differential effects of inbreeding and selection on male reproductive phenotype associated with the colonization and laboratory maintenance of *Anopheles gambiae*. *Malar. J.* **13**, 19 (2014).
- Ross, P. A., Endersby-Harshman, N. M. & Hoffmann, A. A. A comprehensive assessment of inbreeding and laboratory adaptation in *Aedes aegypti* mosquitoes. *Evol. Appl.* **12**, 572–586 (2019).
- Norris, D. E., Shurtleff, A. C., Touré, Y. T., Lanzaro, G. C. & Microsatellite, D. N. A. polymorphism and heterozygosity among field and laboratory populations of *Anopheles gambiae* s.s. (Diptera: Culicidae). *J. Med. Entomol.* **38**, 336–340 (2001).
- Lainhart, W. et al. Changes in genetic diversity from field to laboratory during colonization of *Anopheles darlingi* Root (Diptera: Culicidae). *Am. J. Trop. Med. Hyg.* **93**, 998–1001 (2015).
- Schmidt, H. et al. Transcontinental dispersal of *Anopheles gambiae* occurred from West African origin via serial founder events. *Commun. Biol.* **2**, 473 (2019).
- The *Anopheles gambiae* 1000 Genomes Consortium. Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature* **552**, 96 (2017).
- Lee, Y. et al. Genome-wide divergence among invasive populations of *Aedes aegypti* in California. *BMC Genomics* **20**, 204 (2019).
- Callaway, E. Gene drives meet the resistance. *Nature* **542**, 15 (2017).
- Drury, D. W., Dapper, A. L., Siniard, D. J., Zentner, G. E. & Wade, M. J. CRISPR-Cas9 gene drives in genetically variable and nonrandomly mating wild populations. *Sci. Adv.* **3**, e1601910 (2017).
- Unckless, R. L., Clark, A. G. & Messer, P. W. Evolution of resistance against CRISPR-Cas9 gene drive. *Genetics* **205**, 827–841 (2017).
- Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B. & Valen, E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res.* **44**, W272–W276 (2016).
- Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M. & Valen, E. CHOPCHOP: a CRISPR-Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* **42**, W401–W407 (2014).
- Labun, K., Montague, T. G., Krause, M. & Torres Cleuren, Y. N. Tjeldnes Hk, Valen E. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* **47**, W171–W174 (2019).
- Love, R. R. et al. Chromosomal inversions and ecotypic differentiation in *Anopheles gambiae*: the perspective from whole-genome sequencing. *Mol. Ecol.* **25**, 5889–5906 (2016).
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- Noble, C., Olejarz, J., Esvelt, K. M., Church, G. M. & Nowak, M. A. Evolutionary dynamics of CRISPR gene drives. *Sci. Adv.* **3**, e1601964 (2017).
- North, A. R., Burt, A. & Godfray, H. C. J. Modelling the potential of genetic control of malaria mosquitoes at national scale. *BMC Biol.* **17**, 26 (2019).
- KaramiNejadRanjbar, M. et al. Consequences of resistance evolution in a Cas9-based sex conversion-suppression gene drive for insect pest management. *Proc. Natl Acad. Sci. USA* **115**, 6189–6194 (2018).
- Sudweeks, J. et al. Locally Fixed Alleles: a method to localize gene drive to island populations. *bioRxiv* <https://doi.org/10.1101/509364> (2019).
- Greenbaum, G., Feldman, M. W., Rosenberg, N. A. & Kim, J. Designing gene drives to limit spillover to non-target populations. *bioRxiv* <https://doi.org/10.1101/680744> (2019).
- Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, 1983).
- Zheng, T. et al. Profiling single-guide RNA specificity reveals a mismatch sensitive core sequence. *Sci. Rep.* **7**, 40638 (2017).
- Van Rossum, G. & Drake, F. L. Jr. *Python Tutorial* (Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995).
- Xu, H. et al. Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25**, 1147–1157 (2015).
- Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- Sharakhova, M. V. et al. Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biol.* **8**, R5 (2007).
- Li, H. Aligning sequence reads, clone sequences and assembly contigs with Q5 BWA-MEM. arXiv:1303.3997 [q-bio.GN] (2013).
- Garrison E. & Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio.GN] (2012).
- McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

45. The *Anopheles gambiae* 1000 Genomes Consortium. Ag1000G phase 2 AR1 data release. (2017).
46. Matthews, B. J. et al. Improved reference genome of *Aedes aegypti* informs arbovirus vector control. *Nature* **563**, 501–507 (2018).
47. Collier, T. travc/Cas9\_target\_site\_survey: Publication release. Zenodo. <https://doi.org/10.5281/ZENODO.3661448> (2020).

### Acknowledgements

We thank Allison Weakley, Hans Gripkey, Kendra Person, Youki Yamasaki, and Catelyn Neiman for carrying out DNA extraction and genomic DNA library preparations. Funding for this work was provided by the University of California Irvine Malaria Initiative, Pacific Southwest Regional Center of Excellence for Vector-Borne Diseases funded by the U.S. Centers for Disease Control and Prevention (Cooperative Agreement 1U01CK000516) and from NIH R56 grant (R56AI130277). We thank Ethan Bier and Anthony James for useful comments on an earlier draft of this paper.

### Author contributions

G.C.L., Y.L., and H.S. conceived the study. T.C.C., H.S., P.H., and M.J.H. performed target site analysis. G.C.L. conducted field sample acquisition. Y.L. carried out whole genome sequencing of mosquito field isolates. T.C.C., H.S., and Y.L. performed polymorphism analysis. H.S., T.C.C., and G.C.L. drafted the first version of the paper. All authors contributed to the final version of the paper.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-15204-0>.

**Correspondence** and requests for materials should be addressed to G.C.L.

**Peer review information** *Nature Communications* thanks Antoinette Piaggio and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



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

© The Author(s) 2020