RESEARCH ARTICLE

Open Access

Engineered sex ratio distortion by X-shredding in the global agricultural pest *Ceratitis capitata*

Angela Meccariello[1], Flavia Krsticevic[2], Rita Colonna[1], Giuseppe Del Corsano[1], Barbara Fasulo[1], Philippos Aris Papathanos[2*] and Nikolai Windbichler[1*]

Abstract

Background: Genetic sex ratio distorters are systems aimed at effecting a bias in the reproductive sex ratio of a population and could be applied for the area-wide control of sexually reproducing insects that vector disease or disrupt agricultural production. One example of such a system leading to male bias is X-shredding, an approach that interferes with the transmission of the X-chromosome by inducing multiple DNA double-strand breaks during male meiosis. Endonucleases targeting the X-chromosome and whose activity is restricted to male gametogenesis have recently been pioneered as a means to engineer such traits.

Results: Here, we enabled endogenous CRISPR/Cas9 and CRISPR/Cas12a activity during spermatogenesis of the Mediterranean fruit fly *Ceratitis capitata*, a worldwide agricultural pest of extensive economic significance. In the absence of a chromosome-level assembly, we analysed long- and short-read genome sequencing data from males and females to identify two clusters of abundant and X-chromosome-specific sequence repeats. When targeted by gRNAs in conjunction with Cas9, cleavage of these repeats yielded a significant and consistent distortion of the sex ratio towards males in independent transgenic strains, while the combination of distinct distorters induced a strong bias (~ 80%).

Conclusion: We provide a first demonstration of CRISPR-based sex distortion towards male bias in a non-model organism, the global pest insect *Ceratitis capitata*. Although the sex ratio bias reached in our study would require improvement, possibly through the generation and combination of additional transgenic lines, to result in a system with realistic applicability in the field, our results suggest that strains with characteristics suitable for field application can now be developed for a range of medically or agriculturally relevant insect species.

Background

Effecting a substantial bias in the reproductive sex ratio of a population towards males has long been recognized as a potentially powerful means for genetic control [1]. While naturally occurring sex distortion traits have been of longstanding interest to evolutionary biologists [2], the advent of powerful genome-editing tools and synthetic biology now allows for the generation of artificial sex distortion traits designed to realize this untapped potential. Sexually reproducing insect species that vector disease or disrupt agricultural production are considered the prime targets for such a genetic control approach that, in its various designs, is predicted to be more efficacious than the mass release of sterile males [3], the current gold standard in genetic pest control [4]. Several molecular mechanisms could be engineered to bring about a genetic male bias, which is preferred as females

---

* Correspondence: p.papathanos@mail.huji.ac.il;
nikolai.windbichler@imperial.ac.uk
[2]Department of Entomology, Robert H. Smith Faculty of Agriculture, Food and Environment, Hebrew University of Jerusalem, Rehovot, Israel
[1]Department of Life Sciences, Imperial College London, Sir Alexander Fleming Building, South Kensington Campus, London, UK

are generally the sex causing damage or transmitting disease. X-shredding has been pioneered in the mosquito *Anopheles gambiae* due to the fortuitous discovery of an X-linked target gene cluster and a suitable endonuclease [5, 6]. X-shredding interferes with the transmission of the X-chromosome by inducing multiple DNA double-strand breaks (DSBs) during male meiosis. This leads to an overrepresentation of Y-bearing gametes amongst those that successfully fertilized embryos. The mechanism which links DNA cleavage to a transmission advantage of the Y-chromosome is not fully understood. The competitiveness of mature sperm has been suggested as a contributing factor in the mosquito X-shredding system [7]. Recently, the advent of CRISPR/Cas9 enabled the design of RNA-guided X-shredders that have now successfully been demonstrated in *Anopheles gambiae* [8] and *Drosophila melanogaster* [9]. As an alternative approach, the male-determining genes or M-factors that have recently been identified in several insect species could also be used to engineer effective sex distortion traits [10–14].

Artificial sex distorters could be applied in a variety of ways, some of which have now been demonstrated in the laboratory. Autosomal distorter traits have been shown to eliminate caged mosquito populations [6] and could, possibly in the form of repressible or inducible transgenes, be developed as a next-generation inundative genetic control strategy for insects that are currently controlled by the sterile insect technique (SIT). Both X-shredders and M-factors could be mobilized via gene drive for a powerful inoculative type of control as has been recently demonstrated in *Anopheles gambiae* [15]. Similarly, when linked to the Y-chromosome, X-shredders could constitute a driving Y-chromosome that would favour its transmission over the X-chromosome and the resulting decline in female numbers is predicted to lead to the eventual collapse of the population [16]. This strategy was one of the first proposed but may be challenging to engineer due to transcriptional silencing of the Y-chromosome during male meiosis [17]. Finally, post-zygotic genetic sex distorters that act in and are transmitted via the male germline could also be linked to the Y-chromosome and would obtain a level of persistence in the population. Such post-zygotic sex ratio distorters have recently been demonstrated in *Drosophila melanogaster* via the targeting of X-linked haplolethal genes during male meiosis [9].

Ceratitis capitata, the Mediterranean fruit fly or medfly, is one of the most destructive global agricultural pests due to its wide host range of close to 300 different crop plants [18]. The control of medfly populations relies primarily on mass releases of sterilized males as the main component of an area-wide integrated pest management strategy [19]. Nevertheless, the global economic costs due to crop damage, quarantine restrictions affecting exports, and the control as well as prevention of medfly infestation amount to well over US\$ 1 billion annually [20] making the medfly an excellent candidate for further development of novel genetic control strategies.

We have recently identified the Y-linked M-factor as one possible avenue to engineer sex distortion traits in the medfly [13], and female-to-male sex conversion systems based on targeting key genes in the medfly sex determination cascade are also being developed [21]. Here, we sought to explore the possibility of engineering CRISPR-based sex distortion traits following the X-shredding paradigm, which operates during male meiosis to increase the transmission of Y-bearing gametes and does not entail the manipulation of the endogenous sex determination pathway. No chromosome-level assembly is currently available for *C. capitata*, and although CRISPR/Cas9 ribonucleoprotein complexes have been successfully used to induce heritable genetic changes in the medfly [21, 22], there are currently no available tools for the endogenous expression of CRISPR. The establishment and demonstration of the X-shredding mechanism in the medfly would thus not only add further evidence that this mechanism is broadly applicable but also that such traits can be established in other organisms of great medical or economic importance, for which a full suite of genomic resources and a powerful set of genome editing tools are currently lacking.

Results

Identification of targetable sequence repeats on the medfly X-chromosome

We have previously described the redkmer pipeline which identifies putative X-linked repeat sequences that are absent from other chromosomes using whole-genome sequencing reads as its primary input [23]. The redkmer pipeline was designed to identify short and highly repeated sequences that are specific to the X-chromosome by combining differential representation in female versus male whole-genome sequence data (also known as chromosome quotient (CQ)) [24] with estimates of sequence abundance in the genome, selecting those that are predicted to be most abundant on the X-chromosome. Here, we used available medfly WGS data [13] consisting of ~ 1.9 M PacBio reads from male medflies of the Fam18 strain and approximately 4.9×10^9 Illumina reads from each sex to identify X-linked sequences which could be targeted by CRISPR/Cas9 [25] or CRISPR/Cas12a (Cpf1) [26] in the male germline. The two endonuclease platforms were used for their differing and complementary properties. They recognize different protospacer adjacent motifs (PAMs) enabling the targeting of a more diverse set of sites. Cas9 and Cas12a cut distal and proximal from their target sites

and produce blunt and sticky ends, respectively. Whereas Cas9 is considered a more active enzyme, Cas12a is smaller and, due to its intrinsic RNase activity, has the ability to process its own gRNAs thus facilitating multiplex genome editing [27]. The Fam18 strain was previously sequenced in efforts to identify the primary signal of the medfly sex determination cascade, *Moy* [13], as it contains a shorter Y-chromosome [28]. Starting from a set of 5.3×10^9 kmers of 25 nucleotides, red-kmer selected 5×10^4 kmers representing the top 0.05% most abundant and X-chromosome-specific kmers (Fig. 1a). We applied further rounds of selection to this initial set. Targetability by either Cas9 or Cas12a was predicted by FlashFry [29] after which we applied criteria that ensure that selected kmers are not only highly abundant in both Illumina and PacBio libraries, but also represented on a maximum number of independent PacBio reads (Fig. 1b–d, see the "Methods" section for additional details). This was done to select against spurious repeats arising from sequencing artefacts and for those where individual repeat units are sufficiently spaced for independent targeting by Cas9 or Cas12a. From the resulting top 25 kmers, we manually selected two Cas9 and two Cas12a kmers for the experimental validation steps (Additional file 1: Table S1, Fig. 1b).

Activity of Cas9 and Cas12a activity during medfly spermatogenesis

To evaluate CRISPR function in the male germline of the medfly, we generated piggyBac transformation constructs for random integration in which Cas9 or Cas12a coding sequences were placed under the transcription control of the previously characterized *C. capitata* β2-tubulin promoter [30]. In both *An. gambiae* and *D. melanogaster*, this promoter drives the expression of genes during the primary spermatocyte stage at the onset of meiosis I, and endonuclease activity at this stage has been demonstrated to cause shredding of the X-chromosome and to interfere with its transmission to progeny [8, 9]. For the expression of the gRNA, we identified a putative U6 gene using BLAST of the *Drosophila* U6snRNA sequence (GenBank accession no. NR002083) against the medfly genome (NW_019376346.1, Ccap_2.1). A candidate medfly promoter element was then obtained by gene synthesis. In addition to two constructs each harbouring a gRNA targeting one of the two top Cas9 candidate repeat sequences (Cas9.1 and Cas9.2), we also generated constructs with a previously described gRNA [22] targeting the autosomal *white eye* (Cas9.w) gene for evaluating CRISPR activity (Fig. 2a). Two Cas12a constructs were generated, a construct targeting only the *white eye* gene (Cas12a.w) and a multi-gRNA construct (Cas12a.m) containing both gRNAs for the top X-shredding targets in addition to the *white eye* gRNA

(Fig. 2a). We generated multiple independent transgenic medfly strains using piggyBac-mediated random integration. We performed inverse PCR to characterize insertion sites (Additional file 2: Figure S1, Additional file 3: Table S2). A total of 7 and 4 independent strains were established for the constructs targeting Cas9 kmer1 and Cas9 kmer2, respectively, and single strains for all other constructs. To assess Cas9 and Cas12a activity against the *white eye* marker gene, we crossed transgenic sons or daughters of Cas9.w, Cas12a.w or Cas12a.m females to the w2Δ strain which carries a deletion spanning exon 2 of the *white eye* gene [31]. Mutagenesis of the *C. capitata white eye* by transient CRISPR/Cas9 targeting of a conserved domain within exon 3 has been previously shown to result in a high rate of individuals lacking eye pigment [22]. We compared the activity of the two endonuclease platforms by counting the fraction of offspring with white eyes compared to the female control in which we did not expect endonuclease expression (Fig. 2). In the progeny of transgenic males, we observed on average 98% white-eyed individuals for Cas9.w strain and < 25% for strain Cas12a.w. (Fig. 2b, c). We scored individuals as mutant only when pigment loss was complete (Fig. 2c), but not when we observed mosaic or orange eyes which were infrequently found in the progeny of Cas12a.w males. We detected no white-eyed individuals in the cross using the multiplexed strain Cas12a.m and also in a control where we crossed wild-type individuals to the w2Δ strain (Fig. 2). We found evidence for limited Cas9 activity in the Cas9.w female control cross with up to 0.34% mutant progeny detected in one family. This suggests that some level of leaky somatic expression of the β2-tubulin driven transgene occurs in females, for example, by a nearby enhancer that overrides its male specificity. Genotyping and sequencing of mutant *white eye* alleles from strain Cas9.w showed a spectrum of indels at the expected target site within exon 3 (Fig. 2d). We concluded that, using the β2-tubulin promoter in combination with Cas9, we were able to achieve high rates of activity during medfly spermatogenesis, although some level of ectopic expression in females could not be excluded.

Characterization of candidate X-shredding strains

Transgenic individuals carrying gRNAs targeting Cas9 kmer1 (Cas9.1strains a to g), those targeting Cas9 kmer2 (Cas9.2 strains a to d) and the Cas12a.m strain were crossed to the wild types to determine the sex ratio of their offspring (Fig. 3a). We expected to observe an effect in males due to the strong expression of the endonuclease during spermatogenesis; the reciprocal cross, due to the absence or comparatively low levels of β2-tubulin activity in females, was used as the control. Figure 3a summarizes the results of these experiments
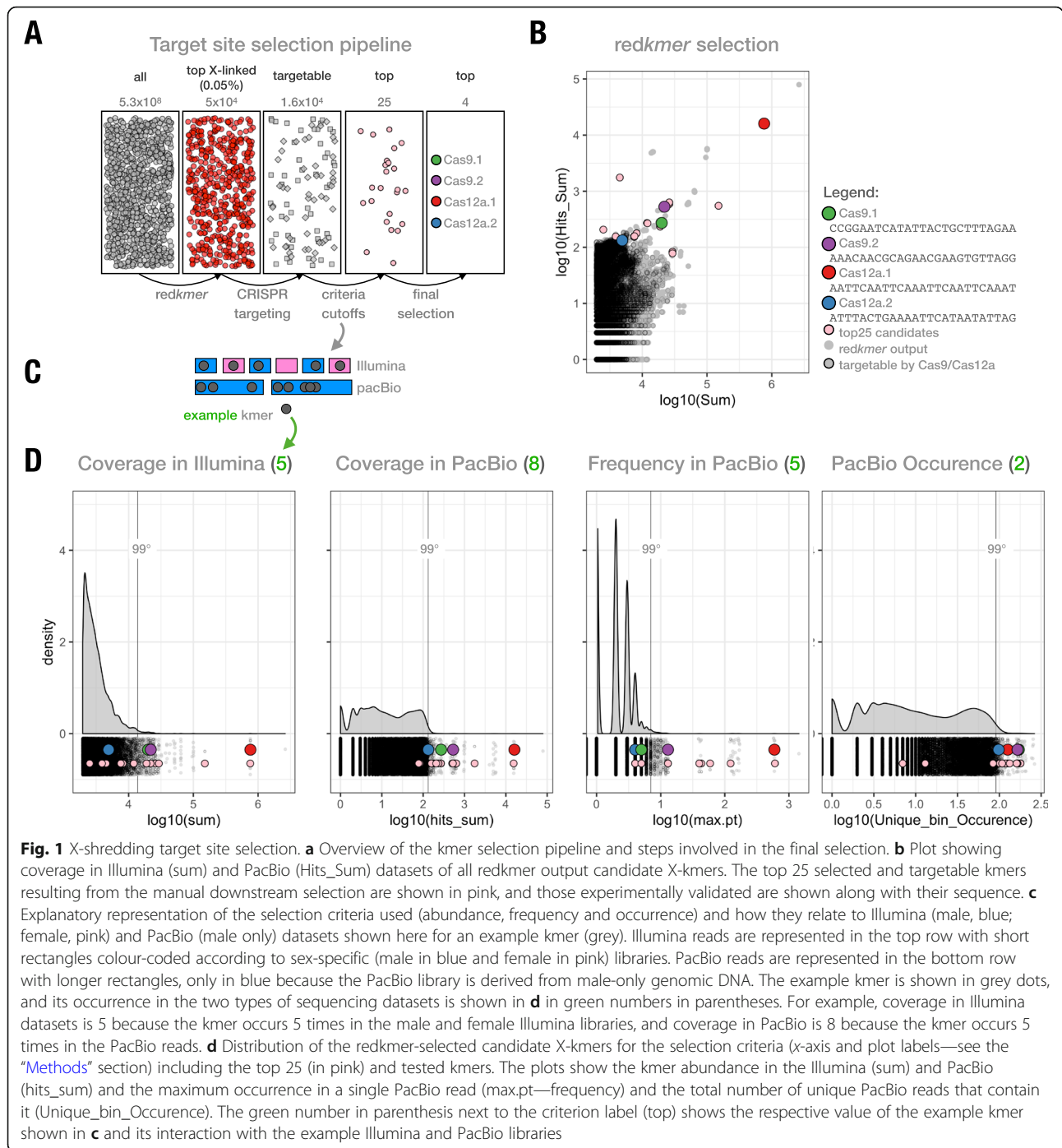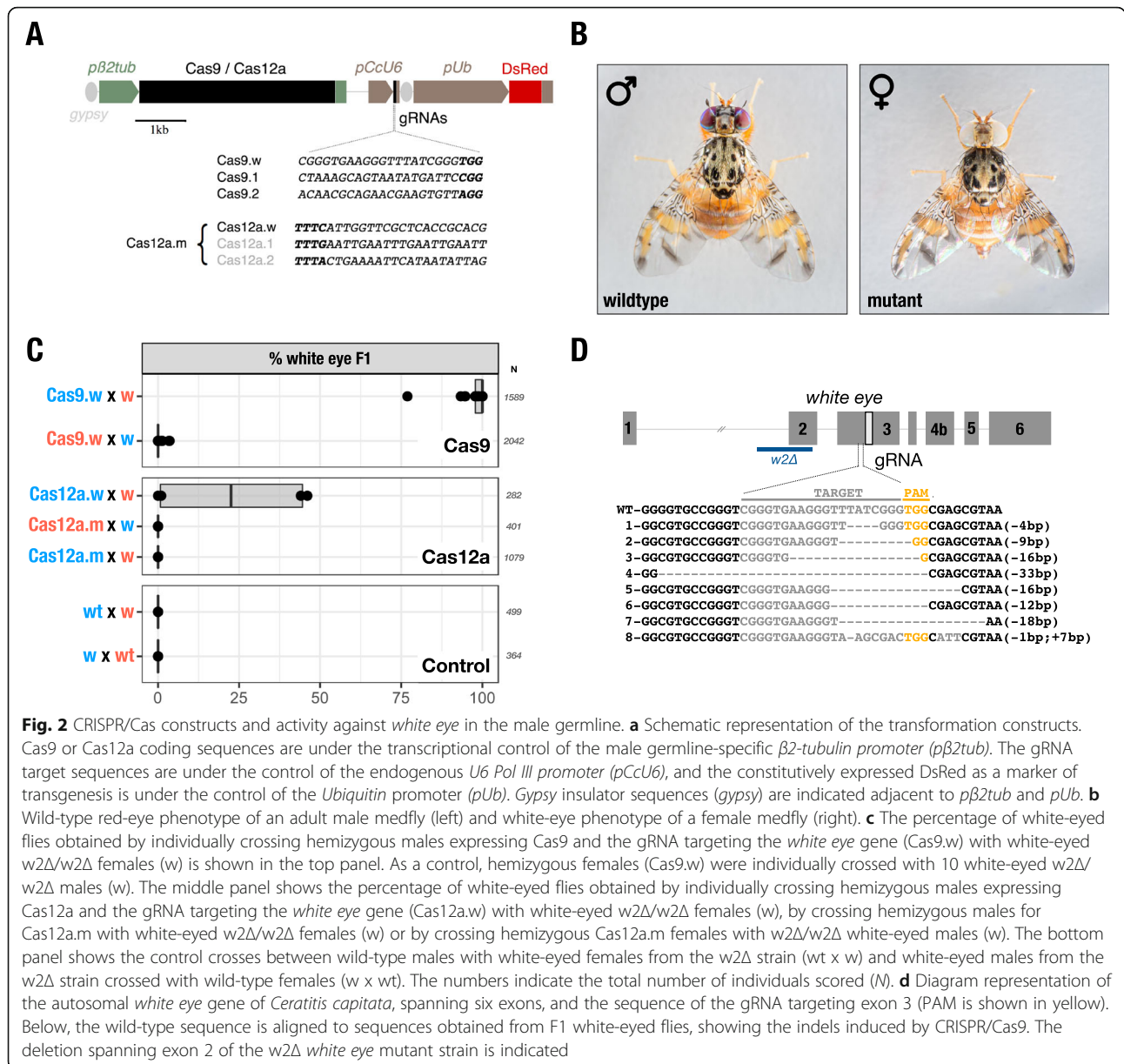
Fig. 1 X-shredding target site selection. **a** Overview of the kmer selection pipeline and steps involved in the final selection. **b** Plot showing coverage in Illumina (sum) and PacBio (Hits_Sum) datasets of all redkmer output candidate X-kmers. The top 25 selected and targetable kmers resulting from the manual downstream selection are shown in pink, and those experimentally validated are shown along with their sequence. **c** Explanatory representation of the selection criteria used (abundance, frequency and occurrence) and how they relate to Illumina (male, blue; female, pink) and PacBio (male only) datasets shown here for an example kmer (grey). Illumina reads are represented in the top row with short rectangles colour-coded according to sex-specific (male in blue and female in pink) libraries. PacBio reads are represented in the bottom row with longer rectangles, only in blue because the PacBio library is derived from male-only genomic DNA. The example kmer is shown in grey dots, and its occurrence in the two types of sequencing datasets is shown in **d** in green numbers in parentheses. For example, coverage in Illumina datasets is 5 because the kmer occurs 5 times in the male and female Illumina libraries, and coverage in PacBio is 8 because the kmer occurs 5 times in the PacBio reads. **d** Distribution of the redkmer-selected candidate X-kmers for the selection criteria (*x*-axis and plot labels—see the "Methods" section) including the top 25 (in pink) and tested kmers. The plots show the kmer abundance in the Illumina (sum) and PacBio (hits_sum) and the maximum occurrence in a single PacBio read (max.pt—frequency) and the total number of unique PacBio reads that contain it (Unique_bin_Occurence). The green number in parenthesis next to the criterion label (top) shows the respective value of the example kmer shown in **c** and its interaction with the example Illumina and PacBio libraries

performed over four consecutive generations of pooled crosses. We found that all Cas9 strains displayed a significant bias towards males in the progeny of transgenic fathers, but not in the progeny of transgenic mothers. An exception was strain Cas9.2a in which a significant deviation from the expected 50% sex ratio was also observed in the progeny of female transgenics, albeit less pronounced (55% males) compared to the male transgenic cross (66% males). Since this was not seen in the

other Cas9.2 strains carrying the same construct, we attributed this to a position/integration effect which may or may not be causally linked to the action of Cas9. No detectable effect on the sex ratio was observed for strain Cas12a.m. Cas9.1 strains showed a male-biassed sex ratio of 54% on average whereas all Cas9.2 strains scored a highly significant and more substantial sex bias of on average 66% male progeny with the best performing strain Cas9.2b reaching a bias of 68% towards males.

Fig. 2 CRISPR/Cas constructs and activity against *white eye* in the male germline. **a** Schematic representation of the transformation constructs. Cas9 or Cas12a coding sequences are under the transcriptional control of the male germline-specific *β2-tubulin promoter (pβ2tub)*. The gRNA target sequences are under the control of the endogenous *U6 Pol III promoter (pCcU6)*, and the constitutively expressed DsRed as a marker of transgenesis is under the control of the *Ubiquitin promoter (pUb)*. *Gypsy* insulator sequences (*gypsy*) are indicated adjacent to *pβ2tub* and *pUb*. **b** Wild-type red-eye phenotype of an adult male medfly (left) and white-eye phenotype of a female medfly (right). **c** The percentage of white-eyed flies obtained by individually crossing hemizygous males expressing Cas9 and the gRNA targeting the *white eye* gene (Cas9.w) with white-eyed w2Δ/w2Δ females (w) is shown in the top panel. As a control, hemizygous females (Cas9.w) were individually crossed with 10 white-eyed w2Δ/w2Δ males (w). The middle panel shows the percentage of white-eyed flies obtained by individually crossing hemizygous males expressing Cas12a and the gRNA targeting the *white eye* gene (Cas12a.w) with white-eyed w2Δ/w2Δ females (w), by crossing hemizygous males for Cas12a.m with white-eyed w2Δ/w2Δ females (w) or by crossing hemizygous Cas12a.m females with w2Δ/w2Δ white-eyed males (w). The bottom panel shows the control crosses between wild-type males with white-eyed females from the w2Δ strain (wt x w) and white-eyed males from the w2Δ strain crossed with wild-type females (w x wt). The numbers indicate the total number of individuals scored (*N*). **d** Diagram representation of the autosomal *white eye* gene of *Ceratitis capitata*, spanning six exons, and the sequence of the gRNA targeting exon 3 (PAM is shown in yellow). Below, the wild-type sequence is aligned to sequences obtained from F1 white-eyed flies, showing the indels induced by CRISPR/Cas9. The deletion spanning exon 2 of the w2Δ *white eye* mutant strain is indicated
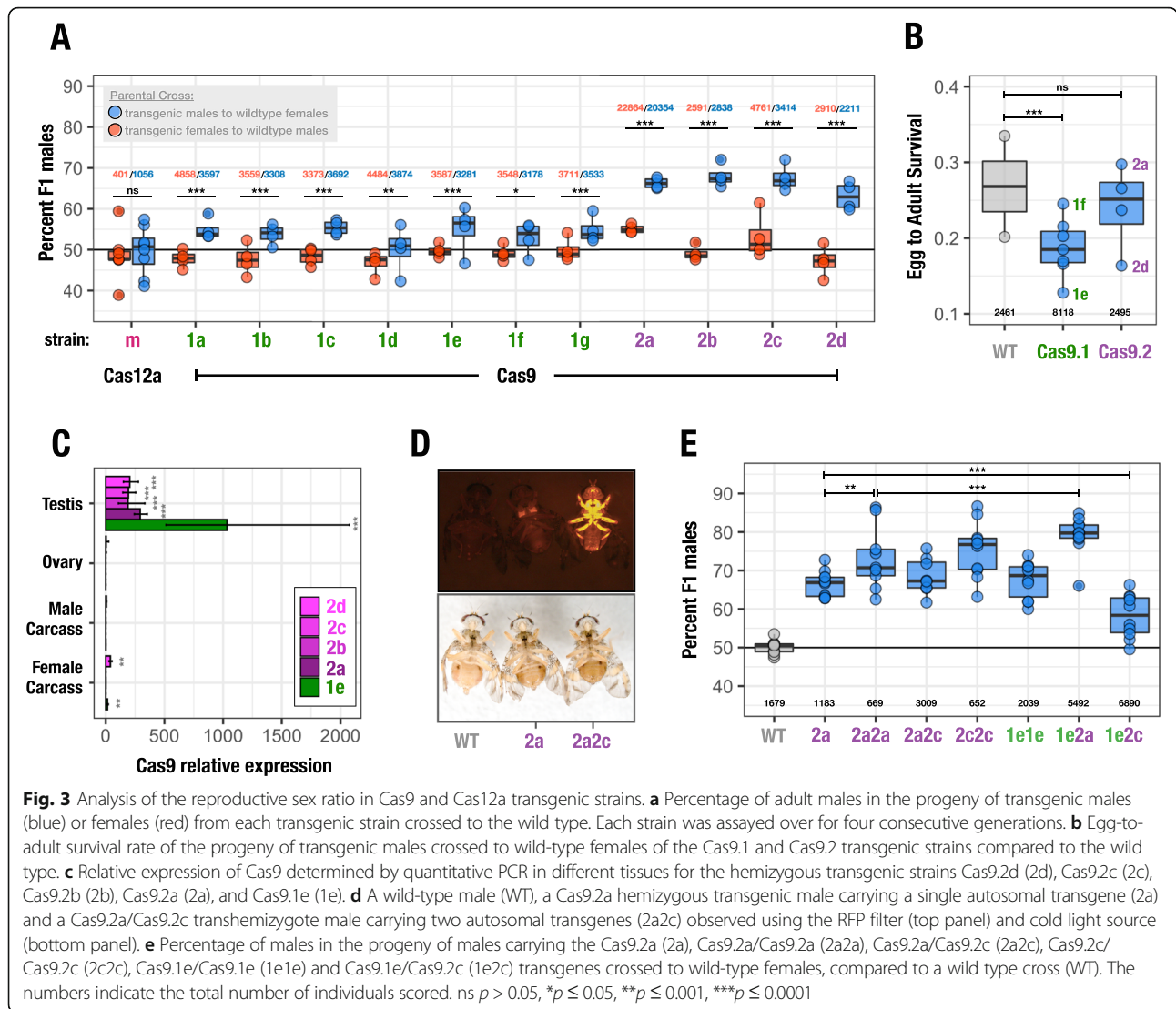
We measured egg-to-adult survival rates for all Cas9 strains to determine whether targeting these two clusters had a significant postzygotic effect on survival. We found highly variable rates of survival across experimental groups including also the two wild-type controls. To detect effects of X-shredding rather than strain-specific or insertion-related effects, we combined data from all Cas9.1 and from all Cas9.2 strains. Overall, a significant decrease in egg-to-adult survival was observed for Cas9.1 but not for Cas9.2 although the latter set of strains shows a consistently more biased sex ratio towards males (Fig. 3b). We next performed a quantitative PCR analysis to determine the level of Cas9 expression in relevant tissues (testes, ovaries, male and female carcasses) of these strains

(Fig. 3c). As expected, the most substantial expression was observed in the testes for all strains analysed (Cas9.1e, Cas9.2a, Cas9.2b, Cas9.2c, Cas9.2d). The average expression in strain Cas9.1e was 3.5× fold (Cas9.2a, $p = 0.060$, t test) to 5.4× fold (Cas9.2b, $p = 0.103$, t test) higher compared to the Cas9.2 strains, but this difference was not statistically significant. A low but significant expression of Cas9 was detected also in the female carcass of strains Cas9.1e ($p \leq 0.001$) and Cas9.2d ($p \leq 0.001$) but not in any other tissue.

Analysis of transgene combinations

We next tested the effects of homozygous and transhemizygous combinations of transgenes that each

Fig. 3 Analysis of the reproductive sex ratio in Cas9 and Cas12a transgenic strains. **a** Percentage of adult males in the progeny of transgenic males (blue) or females (red) from each transgenic strain crossed to the wild type. Each strain was assayed over for four consecutive generations. **b** Egg-to-adult survival rate of the progeny of transgenic males crossed to wild-type females of the Cas9.1 and Cas9.2 transgenic strains compared to the wild type. **c** Relative expression of Cas9 determined by quantitative PCR in different tissues for the hemizygous transgenic strains Cas9.2d (2d), Cas9.2c (2c), Cas9.2b (2b), Cas9.2a (2a), and Cas9.1e (1e). **d** A wild-type male (WT), a Cas9.2a hemizygous transgenic male carrying a single autosomal transgene (2a) and a Cas9.2a/Cas9.2c transhemizygote male carrying two autosomal transgenes (2a2c) observed using the RFP filter (top panel) and cold light source (bottom panel). **e** Percentage of males in the progeny of males carrying the Cas9.2a (2a), Cas9.2a/Cas9.2a (2a2a), Cas9.2a/Cas9.2c (2a2c), Cas9.2c/Cas9.2c (2c2c), Cas9.1e/Cas9.1e (1e1e) and Cas9.1e/Cas9.2c (1e2c) transgenes crossed to wild-type females, compared to a wild type cross (WT). The numbers indicate the total number of individuals scored. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.001$, ***$p \leq 0.0001$

individually was able to bias the sex ratio by either targeting the same or both X-linked repeat clusters. We crossed individuals of transgenic strains Cas9.1e, Cas9.2a and Cas9.2c and identified homozygous or transhemizygous individuals in the progeny using the intensity of polyubiquitin DsRed expression (Fig. 3d) to distinguish individuals carrying single or double transgene insertions. We confirmed by PCR genotyping on a subset of individuals that fluorescence intensity is a suitable indicator. We subsequently crossed individual homozygous or transhemizygous males to 10 wild-type females and determined the sex ratio of their progeny in 10 replicate crosses for each combination (Fig. 3e). As an internal control, we used wild-type males and hemizygous Cas9.2a males which showed a similar level of distortion in individual crosses (Fig. 3e) compared to pooled crosses (Fig. 3a). We observed that doubling the transgene dose can boost distortion. The Cas9.2a/Cas9.2a

combination for example performed significantly better than the Cas9.2a hemizygous cross. The Cas9.1e/Cas9.1e combination resulted in 67.5% males whereas all Cas9.1 hemizygous crosses in panel A averaged well below 60%. The progeny of males carrying the Cas9.1e/Cas9.2a combination consisted of ∼ 80% males on average, the highest rate of distortion observed in this study, suggesting that combinations of gRNAs with different targets can boost the sex bias substantially. In contrast, we observed a substantially reduced distortion on Cas9.1e/Cas9.2c males, significantly lower than the hemizygous Cas9.2a control (Fig. 3e).

Analysis of the target repeat structures

We used a recent re-assembly of the medfly genome built with PacBio and Hi-C data (EgII_Ccap3.2.1, GenBank assembly accession: GCA_905071925.1) [32] to characterize the possible origin and structure of the

sequences selected by redkmer and in particular of the experimentally targeted kmers. While the assembly remains somewhat fragmented (N50 ~ 77.38 Mb), we reasoned that it could possibly help to understand the nature and function of the sequences targeted.

We first evaluated the chromosomal origin for each of the scaffolds of the EgII_Ccap3.2.1 assembly by calculating the chromosome quotient (CQ) for all 18,520 genes in the annotated EgII_Ccap3.2.1 geneset and then assigned CQ to each of the 2712 contigs based on the average CQ of genes it contains (Fig. 4a, Additional file 4: Table S3). This analysis revealed that the X-chromosome constitutes the bulk of scaffold 3 but that this scaffold also contains Y-chromosome-derived sequences (Fig. 4a). We then used BLASTN to locate the redkmer candidate X-kmers, including the top twenty-five kmers and the set of experimentally validated kmers, within the genome assembly. In accordance with expectation, the candidate X-kmers mapped predominantly to scaffold 3 (Fig. 4a and Additional file 5: Figure S2),

demonstrating the specificity of X-chromosome targeting by redkmer and our downstream selection. Both experimentally validated Cas9 kmers mapped exclusively to scaffold 3 with 66 and 50 hits for Cas9.1 and Cas9.2, respectively. Cas12a.2 was also specific to scaffold 3, but with only 8 target site hits, while Cas12a.1 also had one hit on scaffold 2 and 15 hits on scaffold 3 and was not orders of magnitude more abundant than Cas12a.2 in the assembly, as predicted by the redkmer output. This discrepancy is likely the outcome of assembly-based loss or reduction of repetitive sequences, which does not affect redkmer which is based purely on raw coverage data (Additional file 5: Figure S2). Interestingly, the best match in the assembly for kmer Cas9.1 showed a single-base A to G mismatch in 56 of the 66 target sites; the remaining had additional mismatches (Fig. 4b), which may reflect strain differences.

To understand the possible origin and function of the sequences we targeted, we next used RepeatModeler [33] to annotate and catalogue the resident repeats in
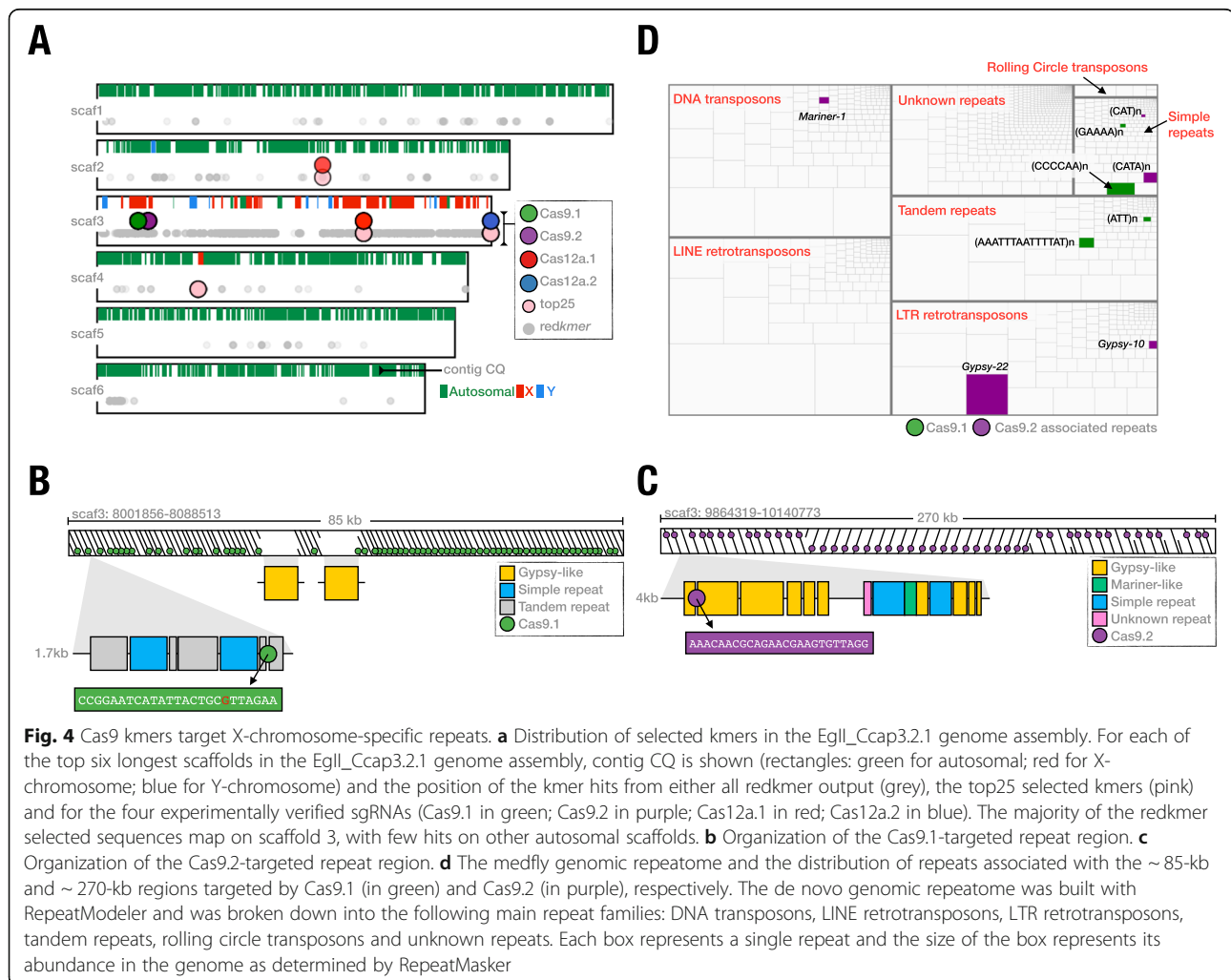
Fig. 4 Cas9 kmers target X-chromosome-specific repeats. **a** Distribution of selected kmers in the EgII_Ccap3.2.1 genome assembly. For each of the top six longest scaffolds in the EgII_Ccap3.2.1 genome assembly, contig CQ is shown (rectangles: green for autosomal; red for X-chromosome; blue for Y-chromosome) and the position of the kmer hits from either all redkmer output (grey), the top25 selected kmers (pink) and for the four experimentally verified sgRNAs (Cas9.1 in green; Cas9.2 in purple; Cas12a.1 in red; Cas12a.2 in blue). The majority of the redkmer selected sequences map on scaffold 3, with few hits on other autosomal scaffolds. **b** Organization of the Cas9.1-targeted repeat region. **c** Organization of the Cas9.2-targeted repeat region. **d** The medfly genomic repeatome and the distribution of repeats associated with the ~ 85-kb and ~ 270-kb regions targeted by Cas9.1 (in green) and Cas9.2 (in purple), respectively. The de novo genomic repeatome was built with RepeatModeler and was broken down into the following main repeat families: DNA transposons, LINE retrotransposons, LTR retrotransposons, tandem repeats, rolling circle transposons and unknown repeats. Each box represents a single repeat and the size of the box represents its abundance in the genome as determined by RepeatMasker

the EgII_Ccap3.2.1 assembly, since the kmer hits on scaffold 3 did not overlap with any known genome annotations. A library of 2037 consensus repeats including transposable elements, tandem and low-complexity repeats was mapped on the assembly using RepeatMasker, and regions surrounding the Cas9 kmer targets on scaffold 3 were manually annotated (Additional file 6). Cas9.1 kmer hits were distributed in a region spanning ~ 85 kb of scaffold 3 (8,001,979-8,087,013), and Cas9.2 kmer hits spanned a region of ~ 270 kb (9,864,418-10, 124,412). The Cas9.1 target sequence was embedded in an approximately 1.7-kb-long repeat unit, which itself was composed of distinct simple and tandem repeats that were repeated and arranged head-to-tail throughout the array, with two interruptions from a Gypsy-like element (Fig. 4b). The Cas9.2 target sequence instead was derived from a larger 4-kb repeat unit that is repeated 50 times in a head to tail fashion with a large repeat inversion in the middle (Fig. 4c). We found that the 4-kb repeat unit is mostly composed of fragmented retrotransposons and simple repeat elements, and the Cas9.2 target sequence was located in an incomplete Gypsy-like element. To understand the representation of the targeted repeats within the entire landscape of medfly repeatome, we calculated the relative abundance of each of the consensus sequences of our custom repeat library using RepeatMasker and then highlighted those derived from the 85-kb and 270-kb regions of the Cas9.1 and Cas9.2 target kmer, respectively (Fig. 4d). We found no evidence of active transcription occurring at these two repeat clusters (data not shown). We concluded from these results that X-chromosome shredding sex ratio distorters can be engineered by targeting resident repeats of the X-chromosome, and these repeats do not have to be highly abundant, conserved or functional.

Discussion

Our primary goals were to enable CRISPR/Cas9 and Cas12a genome editing in the medfly germline, to identify CRISPR-targetable sequence repeats located on the medfly X-chromosome using existing male and female genome sequencing data and, finally, to target these repeats with specific gRNAs in the hope of biasing the sex ratio of the progeny of transgenic males towards sons. We were able to achieve nearly optimal rates of Cas9 activity in the male germline against the *white eye* gene using the *C. capitata* β2-tubulin promoter and the CcU6.1 promoter we have isolated. The establishment of an efficient CRISPR/Cas9 toolset for transgenesis opens a number of possibilities from the generation of strains for achieving genetic sterility to the development of suppressive gene drives. Activity against *white eye* was lower for strain Cas12a.w, but Cas12a appears to be a viable genome editing system in the medfly offering a

complementary set of targetable sites. No activity against *white eye* could be detected for the multiplexed strain Cas12a.m although it shared the same gRNA targeting *white eye* with strain Cas12a.w. This either suggests a strong position effect affecting Cas12a or gRNA expression levels between these two strains or, as previously demonstrated, a loss in efficiency of expression and cleavage when gRNA arrays are utilized [34]. It also does not allow us to conclude on the suitability of the Cas12a.1 and the Cas12a.2 target repeats for X-shredding. The kmers for Cas12a.1, predicted by redkmer to be an order of magnitude more abundant than both selected Cas9 kmers (Fig. 1b), remain an attractive target to be evaluated going forward. We chose two Cas9 targets predicted by redkmer for experimental validation in the medfly and both Cas9 gRNAs gave rise to a significant sex bias, which improves upon previous results in *Drosophila* where only one out of eight gRNAs designed for X-shredding were found to be effective [9]. The consistency of the relative strength of the observed bias within the set of Cas9.1 and Cas9.2 strains suggested that the nature of the target sequence was more important in determining the effect on the sex ratio than possible expression differences between different insertion sites of the transgene. We did not find significant Cas9 expression differences in the testes of the strains analysed by qPCR. This result is in stark contrast with results in *Anopheles* where different integrations generated widely varying levels of expression and distortion [6, 8] and is possibly explained by our use of insulator sequences flanking Cas9. It also suggests that constructs lacking such insulators could be used to select for more active X-shredders.

The target sequences of the two previously studied successful X-shredding systems fell within the coding sequence of the X-linked *Muc14a* gene and within the X-linked rDNA gene cluster, in *Drosophila melanogaster* and *Anopheles gambiae*, respectively. For the present set of experiments in the medfly, our bioinformatic pipeline, making no assumption about the nature or function of the most abundant and X-specific repeats, identified and selected targets that are associated with clusters of X-linked repeats. These repeats, featuring an average distance of approximately 1.7 and 4 kb between the Cas9.1 and Cas9.2 targets, respectively, were found to consist of both simple repeat sequences and fragments of retrotransposons. They are unlikely to be functional and presumably constitute heterochromatic regions of the X-chromosome. This suggests that sex distorters targeting repeats could be engineered with relative ease in species with an abundance of X-specific heterochromatin, likely a common occurrence in many insects. The fact that the mere presence of a sex chromosome-specific repeat sequence constitutes a sufficient substrate for the

Meccariello *et al. BMC Biology*   (2021) 19:78

Page 9 of 14

operation of an efficient sex distortion trait also supports the notion that the evolution of both X- and Y-specific sequence heterochromatin could to a large extent drive and in turn be driven by the evolutionary dynamics of sexually antagonistic selfish genes [35]. A downside of selecting such "junk" DNA targets may be the low level of functional conservation which would facilitate the selection for resistance [36], although the actual mechanism by which X-shredding eliminates X-bearing gametes remains unknown and hence also the pathways via which resistance is likely to arise. Several caveats apply to the post hoc analysis we performed based on the genome assembly and the conclusions we reached above. As mentioned, the actual structure of these repeat clusters may differ in size and composition from what is predicted by the assembly and possibly also between strains and even individuals. Furthermore, the EgII_Ccap3.2.1 [32] assembly was generated from individuals of a medfly laboratory strain (EgyptII) which differs from our experimental strain (Benakeion) and the strain used for redkmer target site selection (Fam18). Our abundance estimates and predictions on chromosome specificity and/or size of the target locus may in fact differ between these strains. This is exemplified by Cas9.1 kmer where we observed strain-specific differences between the kmer targets predicted by redkmer and the genome assembly. While the single nucleotide mismatch 16 nucleotides away from the PAM was not predicted to abolish Cas9 activity, the weaker sex distortion effect we observed for Cas9.1 may be partially explained by this discrepancy. The analysis also underlined the importance of a stringent downselection of putative X-specific targets. Both repeat clusters we analysed provided a large set of alternative targetable sites many of which are more abundant than the Cas9 kmers we had selected. A subset of these possible targets also appears to be X-specific, i.e. abundant on Scaffold 3 yet absent from the remaining Scaffolds in the assembly (Additional file 7: Figure S3A,B). Redkmer however excluded such targets due to their lack of X-specificity determined on the basis of long-read and short-read quotients. It further proves the utility of dedicated tools such as redkmer for the bioinformatic identification of chromosome-specific sequence families, e.g. CRISPR target sequences to induce X-shredding for species of medical or agricultural importance for which high-quality genomes are not available.

We found evidence for positive synergy when a combination of gRNAs targeting different repeats was used, resulting in a level of male bias that is closer to what would be required for application. These experiments were of particular interest as they have not previously been possible in *An. gambiae* or *D. melanogaster*, where for each organism only a single functional X-shredding target is currently available. Combining two different gRNAs (Cas9.1e and Cas9.2a) resulted in the highest levels of distortion obtained in this study, above the level of any single gRNA or a double dose of a well-performing gRNA and Cas9 in homozygous males (Fig. 3e). However, a different combination of transgenes (Cas9.1e and Cas9.2c) resulted in a significantly reduced level of distortion indicating that further research on elucidating the consequences of X-cleavage in different positions is needed. In particular, an analysis of the structure of X-chromosomes following cycles of cleavage and repair in one or both target regions could help to better understand these observations. Individual genetic differences possibly linked to leaky expression of Cas9 in some of these strains could also contribute to this outcome.

Conclusions

Our findings suggest that an autosomal and self-limiting medfly sex distorter strain with performance and fitness characteristics suitable for inundative releases could now be developed. Such a strain could conceivably combine the existing genetic sexing system and mass rearing technology of the medfly with an inducible or repressible sex distortion system. However, the sex distortion traits we have described would need to be improved further to achieve field applicability. One possibility is to remobilize the existing constructs or generate and test further insertions and their various combinations to obtain more efficient sex distorters. Such an approach was recently taken to successfully generate an autocidal lethality strain in the medfly where a large set of combinations of independent driver lines and effector strains were screened to obtain the desired lethal trait [37]. Further improvements in the CRISPR toolkit for the efficient expression of multiple gRNAs could allow the generation of X-shredders simultaneously targeting a class of less-repetitive X-linked sequences. The gRNAs we describe could be improved by combining them with a more active source of Cas9 or with other gRNAs targeting the large set of additional X-specific sequences predicted by redkmer that still remains available for experimental evaluation (Fig. 1a). In *Drosophila*, the combination of X-shredding and X-poisoning systems has also been demonstrated and could help to strengthen the robustness of such traits against resistance or escape [9]. X-shredders could also be deployed as part of a sex-distorting autosomal gene drive, or if they are placed and expressed from the Y-chromosome could constitute suppressive sex-chromosome drive. X-shredding gRNAs could be combined into a highly active and male-specific Cas9 gene drive to boost its suppressive effect. The genetic architecture controlling medfly sex determination in particular, where both fertile XY females and XX males have been generated [13], suggests flexibility to engineer

a range of novel, sex-distorting gene drives in this species. The generation of a driving Y-chromosome for example would be a fruitful endeavour in this species but would require further advances in enabling the engineering of the medfly Y-chromosome and a better understanding to what degree meiotic sex chromosome inactivation is operational in this species. It would not only provide, for the first time, a well-defined system to study Y drive, which is currently lacking altogether, but also a powerful inoculative method for control of this agricultural pest.

Methods

X-shredding target site selection

To select our Cas9 and Cas12a X-chromosome target sequences, we ran redkmer [23] using medfly male and female Illumina whole-genome sequencing (WGS) data and canu-corrected [38] PacBio whole-genome sequencing reads from males with the following parameters: TRIMM5 = 5, TRIMM3 = 5, pac_length = 2000, pac_length_max = 100,000, LSum = 100, XSI = 0.995, kmer-noise = 5. The mitochondrial genome was downloaded from NCBI (NC_000857.1) and used to exclude mitochondrial-derived reads. The initial output of redkmer was a set of kmers that were analysed for Cas9 and Cas12a targeting using Flashfry [29] using as a background genome the PacBio reads deriving from the redkmer autosomal and Y-derived PacBio bins. Candidate X-kmers were then filtered based on their abundance in the Illumina (sum) and PacBio (hits_sum) WGS libraries (Fig. 1b). For each candidate X-kmer, we also calculated its maximum occurrence in a single PacBio read (max.pt – frequency) and the total number of unique PacBio reads that contain it (Unique_bin_Occurence, Fig. 1d). For each of these four parameters (sum, hits_sum, max.pt and Unique_bin_Occurence), kmers occurring in the top 0.05% percentile amongst the redkmer candidate output kmers were identified, and these criteria were then used to select the top 25 kmers with Cas9 or Cas12a targeting potential and excluding those with high off-targets. Scripts used for the filtering are available on https://github.com/genome-traffic/medflyXpaper.

Isolation of the *CcU6* promoter

The *D. melanogaster* U6snRNA sequence (GenBank accession no. NR002083) was used to perform a BLAST search of the *Ceratitis capitata* genome (NW_019376346.1 Reference Ccap_2.1). A sequence 450 bp upstream of the region coding for a predicted U6 spliceosomal RNA in *C. capitata* (LOC111591613, ncRNA, NCBI Reference Sequence: XR_002750719.1) was then selected as the *C. capitata* U6 promoter (*CcU6*).

Cas9 and Cas12a constructs

The p1260 plasmid (kindly provided by Francesca Scolari, University of Pavia, Italy) with piggyBac recombination sequences and *Drosophila melanogaster* pUb-DsRed as described in Scolari et al. [30] was used as the backbone of PiggyBac_Cas9 and PiggyBac_Cas12a constructs. For this study, we used the following synthetic plasmids and primers by Eurofins genomics: CcU6-gRNA-Cas9, CcU6-gRNA-Cas12a.w and CcU6-gRNA-Cas12a.m. To generate the Cas9 constructs, three DNA fragments were amplified: (1) a 1250-bp fragment containing the attP site and the Ccβ2-tubulin 5′ regulatory region was amplified from p1260 with primers *β2-promoterF* and *β2-promoterR*, (2) the human codon-optimized Cas9 coding sequence including two nuclear localization signals (SV40 NLS at the 5′ and nucleoplasmin NLS at the 3′) was amplified from hCas9 (Addgene #41815; http://n2t.net/addgene:41815) using primers *F_Cas9* and *R_Cas9* and (3) the 641 bp 3′-UTR of the β2-tubulin gene was isolated from genomic DNA (extracted from a pool of five adult *C. capitata* males using the Holmes-Bonner protocol [39]) using primers *F_3UTR-β2* and *R_3UTR-β2*. The PCR fragment was amplified after purification with primers *F1_3UTR-β2* and *R1_3UTR-β2-SacII*, the latter containing the *SacII* restriction site. These three PCR products were assembled into an *AscI*-linearized p1260 plasmid to create the PiggyBac_Cas9 construct. The CcU6-gRNA-Cas9 plasmid containing the pair of *AarI* restriction sites used for golden gate cloning of the specific gRNA served as the PCR template for the cloning of the individual gRNA expression plasmids using primers containing *SacII* site, *SacII-CcU6-F* and *SacII-CcU6R*, and the obtained amplicon was cut with *SacII* and then ligated to *SacII*-linearized plasmid PiggyBac_Cas9. All gRNAs were cloned in *AarI*-linearized plasmid PiggyBac_Cas9 creating the three vectors PiggyBac_Cas9.w, PiggyBac_Cas9.1 and PiggyBac_Cas9.2. The PiggyBac_Cas12a plasmid was generated through three steps. In the first step, the *HindIII*-linearized pUK21 plasmid (Addgene #49788; https://www.addgene.org/49788) was used as an intermediate backbone to clone: (1) the Ccβ2-tubulin-promoter (from p1260) amplified with the forward primer designed with restriction site *MluI*, *1a-Bp-MluI-F*, and the reverse primer *1a-Bp-R* and (2) 3′UTR β2-tubulin terminator with the forward primer containing the protospacer-*BsaI* site, *BsaI-Bt-F*, and the reverse primer containing the *SacII* site, *Bt-SacII-R*. The two PCR fragments were re-amplified with the following primers with all characteristics necessary for assembly: *Gb1-pUk21-B2p-F*, *Gb1-pUk21-B2p-R* and *Gb2-BpSpacerBT-F*, *Gb2-BpSpacerBT-R*, respectively. The final products were assembled in pUK-Ccβ2. In the second step, Cas12a (Addgene #69988; http://www.addgene.org/69988/) was amplified

with the primers *Cas12a-BsaI-F* and *Cas12a-BsaI-R* and was assembled through golden gate cloning in the *BsaI*-linerarized plasmid pUK-Ccβ2, generating the plasmid pUKCas12a. In the last step, the pUKCas12a plasmid was cut with the enzyme *MluI* and the digested fragment ligated to the *AscI*-linearized p1260. The CcU6-gRNA-Cas12a.w; CcU6g-RNA-Cas12a.m plasmids, provided with the specific gRNA sequence, were amplified with primers *CcU6-Cas12a-F* and *CcU6-Cas12a-R*. The PCR products were assembled in the final PiggyBac_Cas12a.w and PiggyBac_Cas12.a.m. constructs, respectively. The annotated GenBank files for all plasmid vectors are provided in Additional file 8, for primers see Additional file 9: Table S4.

Ceratitis capitata germline transformation

Germline transformation was performed by microinjection of piggyBac constructs (500 ng/μL) together with the ${}^{i}hyPBase$ [40] transposase helper plasmid (300 ng/μL) into the wild-type embryos as described in Meccariello et al. [13]. Hatched larvae were transferred to Petri dishes containing larval food [22]. Surviving G_0 individuals were crossed to wild-type flies, and positive transformants were identified under fluorescence microscopy for the expression of the PUb-DsRed amongst the G_1 progeny. Transgenic lines originated from a single integration event were selected using inverse PCR following the protocol as described in Scolari et al. [30] using the primers from Galizi et al. [6] (Additional file 3: Table S2). In each generation, transgenic females were backcrossed to Benakeion wild-type males. Flies were screened using the fluorescence stereomicroscope MVX-ZB10 Olympus with the filter for RFP (exciter filter 530–560 nm, dichroic beam splitter 570 nm, barrier filter 585–670 nm).

Medfly rearing

The *C. capitata* transgenic target line, the wild-type strain Benakeion, was kindly provided by Giuseppe Saccone's Lab, University of Naples "Federico II", and Benakeion *white eye* mutant strain (w2Δ) was provided by the FAO/IAEA Agriculture and Biotechnology Laboratory, Seibersdorf, Austria. The strains were maintained in standard laboratory conditions at 26 °C, 65% relative humidity and a 12:12-h light-dark regimen. The adult flies were fed yeast/sucrose powder (1:2).

Cas9.w and Cas12a activity assays against *white eye*

A transgenic male *we*⁺ homozygote was individually crossed to five homozygous w2Δ mutant females, and as a control, a transgenic female *we*⁺ homozygote was individually crossed to five homozygous w2Δ mutant males; a wild-type male was individually crossed to five homozygous w2Δ mutant females; a wild-type female was individually crossed to five homozygous w2Δ mutant males and the number of white eye phenotype adults in the progeny was counted. The crosses of the Cas12a transgenic lines were set up at 34 °C, 65% relative humidity and a 12:12-h light-dark regimen. For the molecular analysis of CRISPR/Cas9-induced mutations in *white eye* gene genomic DNA was extracted from individual flies in 200 μL Holmes Bonner buffer, with minor modifications, according to the protocol of Holmes and Bonner [39]. The resulting DNA was used as a template to amplify the region encompassing the target sites, using the following primers: Cas9_W-F and Cas9_W-R. The PCR products were purified with Monarch® PCR & DNA Cleanup Kit (New England Biolabs) and subcloned using StrataClone PCR cloning Kit (Agilent Technologies) and sequenced.

Determination of the sex ratios and egg-to-adult survival assay

To assay the adult sex ratio, 20 transgenic males were crossed to 30 wild-type females and, as a control, 30 transgenic females were crossed to 20 wild-type males. Each transgenic line was assayed for at least four consecutive generations. Eggs were collected three times, with an interval of 2 to 3 days and reared to adulthood. The total number of adult male and female flies from each cross were counted. The survival test was performed, 20 transgenic males were crossed to 30 wild-type female and, as a control, 20 wild-type males were crossed to 30 wild type females. The number of eggs laid and the number of adults hatching were counted. Each transgenic line was assayed in triplicate, i.e. in three cages of 20 males and 30 females.

Three crosses were performed to generate homozygous individuals for the analysis of the effect of copy number of each transgene: (1) 10 Cas9.1e males and 20 Cas9.1a females, (2) 10 Cas9.2c males and 20 Cas9.2c females and (3) 10 Cas9.2a males and 20 Cas9.2a females. Also, transhemizygous individuals were generated with the following three crosses: (1) 10 Cas9.1e males and 20 Cas9.2a females, (2) 10 Cas9.2c males and 20 Cas9.1e females and (3) 10 Cas9.2a males and 20 Cas9.2c females.

From each cross, 10 homozygous males and 10 transhemizygous males were selected based on the intensity of DsRed fluorescence and were then individually crossed to 10 wild-type females. We validated that DsRed fluorescence intensity could be used as an indicator for the number of transgenes using PCR on genomic DNA extracted from 9 putative transhemizygous males. For this, we used primer 3F1 in combination with either primers Rev-Cas9.1e-DNA, Rev-Cas9.2a-DNA or Rev-Cas9.2c-DNA (Additional file 9: Table S4) and amplified regions ranging from the transgene to insertion-specific genomic flanking sequences. Egg collection and the

determination of adult sex ratio in the progeny were performed as described above. To analyse the sex ratios observed, we used a generalized linear model in which the sex ratios were modelled as a function of the sex of the transgenic parent and the experimental generation using a binomial error distribution (Additional file 10: Table S5). (R scripts and input data are available at https://github.com/genome-traffic/medflyXpaper).

Real-time quantitative PCR

RNA was isolated from dissected testes, male carcasses, dissected ovaries and female carcasses from the Cas9.1e, Cas9.2a, Cas9.2b, Cas9.2c and Cas9.d transgenic lines in biological triplicates using TRIzol Reagent (Thermo Fisher Scientific) following the manufacturer's instructions. The total RNA obtained from each tissue was reverse transcribed using the Thermo Scientific Maxima H Minus First Strand cDNA Synthesis Kit with dsDNase according to the manufacturer's instructions. The real-time quantitative PCR (qPCR) assays were performed using the 7500 Fast Real-Time PCR system. Amplifications were carried out in a solution containing 10 μL 2X Fast SYBR™ Green Master Mix (Thermo Scientific), 2 μL first-stranded cDNA (diluted 1:10) and 800 nM of each primer (for primers see Additional file 9: Table S4), to a final volume of 20 μL. To check reproducibility, each assay was performed with technical triplicates for each of the three biological samples. The Cas9 expression levels were normalized to the housekeeping gene Rpl19 [41] and expression to wild-type samples, and data analysis was performed using the PCR package for R [42] (R scripts and input data are available at https://github.com/genome-traffic/medflyXpaper).

Reagents

All amplification steps were performed using Phusion High-Fidelity DNA Polymerase (New England Biolabs). The cloning was performed with the NEBuilder Hi Fi DNA Assembly kit (New England Biolabs). The enzymes mentioned were purchased from New England Biolabs: *AscI* (#R0558), *MluI* (#R3198), *BsaI* (#R3733) and *SacII* (#R0157). The *AarI* (ER1581) restriction enzyme type II was purchased from Thermo Fisher Scientific. All inserts were verified by sequencing (Genewiz).

Origin of the targeted X-chromosome sequences in the EgII_Ccap3.2.1 medfly genome assembly

To identify sequences of the medfly X-chromosome within the EgII_Ccap3.2.1 assembly, we calculated the chromosome quotient for each of the 18,520 augustus-predicted genes of the assembly using the EgyptII male and female WGS Illumina data [32]. WGS data from males and females were mapped using bowtie [43] following the standard CQ parameters (reporting all

alignments (-a) and allowing no mismatches (-v 0) and CQ was calculated as the ratio of female over male (library-size normalized) reads). Contig CQ was then calculated as the median CQ of genes contained in the 2712 contigs, excluding all contigs containing less than 10 predicted genes. Doing so resulted in median CQ values for the 465 largest contigs representing 15,809 genes (Additional file 4: Table S3). To generate a custom repeat library for the medfly, we used the assembly of the medfly genome and ran RepeatModeler (v2.0.1-0) [32], using RECON (v1.08) [44], RepeatScout (v1.0.6) [45], and TRF (v4.09) [46], with default parameters. The consensus classes were compiled in a set of 2037 sequences belonging to families of transposable elements (TEs), tandem and other low-complexity repeats. Out of the medfly 2037 reference REs library, 1312 consensus classes were annotated and classified hierarchically as TEs, tandem repeats, and simple repeats sequences (https://www.dfam.org/classification/tree and Additional file 11: Table S6). kmers were mapped to the Ccap3.2 assembly using BLASTN (-word_size 5, -max_target_seqs 10,000 and -max_hsps 10,000). To dissect the structure of the Cas9 targeted repeats in the ~ 85-kb and ~ 270-kb arrays of scaffold 3, we extracted the full sequences 100 bp upstream of the first kmer until the next kmer using the Seqtk toolkit (https://github.com/lh3/seqtk) and then mapped it back to the array using Yass [47]. To annotate the higher-order repeat landscape of the ~ 85-kb and ~ 270-kb arrays of scaffold 3, we used RepeatMasker [48] with default parameters and our previously generated repeats database with the -lib option to generate a gff file. This data was used to map the repeats within the targeted arrays. Additionally, we ran BLASTN using the two Cas9 kmers directly against the RepeatModeler library (-word_size 5, 90% identity, and 50% of coverage) and found a unique hit from the Cas9.1 kmer with a tandem repeat sequence and for the Cas9.2 kmer a significant alignment with an incomplete gypsy-like element. We also ran a BLASTN search (cut-off at 80% identity and e-value of 1e−10) using each of the ~ 85-kb and ~ 270-kb arrays against a merged repeats library that included both the medfly RepeatModeler library and the *Drosophila melanogaster* transposon database (https://github.com/bergmanlab/transposons/tree/master/current). Both approaches, RepeatModeler and BLASTN gave similar results. To generate the genome-wide medfly repeat landscape plot shown in Fig. 4d, we used RepeatMasker output to calculate the relative percentage of the total genome assembly masked by each repeat in the library. All repeats were assigned Dfam hierarchical classification nomenclature (full name, sub name, section, type, subtype, section, subsection and name) to cluster the information into boundary boxes. The percentage of each element associated with

the kmers was plotted in R using the treemap package and ggplot (supplemental R scripts and input data are available at https://github.com/genome-traffic/medflyXpaper). To evaluate alternative kmers originating from the targeted regions, the 85- and 270-kb regions of scaffold 3 for Cas9.1 and Cas9.2, respectively, we generated 25 bp kmers from the genome assembly of these regions using jellyfish count (-C, -c 3 and -s 1000000000 parameters), "dump" (-c) and "histo" [49]. The resulting kmers were then mapped by BLASTN either back to the entire assembly, scaffold 3 alone or just the region originally used for generating the kmers by jellyfish. The number of hits against each library was extracted from the output, and the specificity to either scaffold 3 was calculated.

Supplementary Information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12915-021-01010-7.

Additional file 1: Table S1. Kmers and gRNAs sequences shown in 5′ to 3′ orientation.

Additional file 2: Figure S1. Predicted transgene integration sites within the EgII_Ccap3.2.1 assembly.

Additional file 3: Table S2. Sequences flanking piggyBac insertions.

Additional file 4: Table S3. Median CQ values for the 465 largest contigs representing 15,809 genes.

Additional file 5: Figure S2. BLASTN hits of selected kmers to the Ccap3.2 assembly.

Additional file 6:. Library of the consensus repeats including transposable elements.

Additional file 7: Figure S3. Specificity and abundance of kmers originating in the target regions.

Additional file 8:. Genbank plasmids.

Additional file 9: Table S4. Primer sequences shown in 5′ to 3′ orientation.

Additional file 10: Table S5. Statistical analysis.

Additional file 11: Table S6. Annotation of the resident repeats in the medfly genome EgII_Ccap3.2.1 using the Dfam.

Authors' contributions
N.W. and P.A.P. conceived the project. A.M. designed and performed the experimental work. P.A.P., N.W. and F.K. performed the bioinformatics analyses. A.M., R.C., G.D.C. and B.F. maintained the strains and performed the crosses. N.W., P.A.P. and A.M. analysed the data. N.W., P.A.P. and A.M. wrote the paper. All authors read and approved the final manuscript.

Availability of data and materials
All data generated or analysed during this study are included in this published article and its supplementary information files.

Declarations

Competing interests
The authors declare that they have no competing interests.

References
1. Papathanos PA, Windbichler N, Akbari OS. Sex ratio manipulation for insect population control. Transgenic Insects: Techniques andApplications. 2014: 83-100.
2. Jaenike J. Sex chromosome meiotic drive. Annu Rev Ecol Syst. 2001;32(1): 25–49.
3. Schliekelman P, Ellner S, Gould F. Pest control by genetic manipulation of sex ratio. J Econ Entomol. 2005;98(1):18–34. https://doi.org/10.1093/jee/98.1.18.
4. A.S., R. Genetic basis of the sterile insect technique. In: Dyck VA, Hendrichs J, Robinson A, editors. Sterile insect technique. Dordrecht: Springer; 2005. https://doi.org/10.1007/1-4020-4051-2_4.
5. Windbichler N, Papathanos PA, Catteruccia F, Ranson H, Burt A, Crisanti A. Homing endonuclease mediated gene targeting in Anopheles gambiae cells and embryos. Nucleic Acids Res. 2007;35(17):5922–33. https://doi.org/10.1093/nar/gkm632.
6. Galizi R, Doyle LA, Menichelli M, Bernardini F, Deredec A, Burt A, Stoddard BL, Windbichler N, Crisanti A. A synthetic sex ratio distortion system for the control of the human malaria mosquito. Nat Commun. 2014;5(1):3977. https://doi.org/10.1038/ncomms4977.
7. Haghighat-Khah RE, Sharma A, Wunderlich MR, Morselli G, Marston LA, Bamikole C, Hall A, Kranjc N, Taxiarchi C, Sharakhov I, Galizi R. Cellular mechanisms regulating synthetic sex ratio distortion in the Anopheles gambiae germline. Pathog Glob Health. 2020;114(7):370–8. https://doi.org/10.1080/20477724.2020.1803628.
8. Galizi R, Hammond A, Kyrou K, Taxiarchi C, Bernardini F, O'Loughlin SM, Papathanos PA, Nolan T, Windbichler N, Crisanti A. A CRISPR-Cas9 sex-ratio distortion system for genetic control. Sci Rep. 2016;6(1):31139. https://doi.org/10.1038/srep31139.
9. Fasulo B, Meccariello A, Morgan M, Borufka C, Papathanos PA, Windbichler N. A fly model establishes distinct mechanisms for synthetic CRISPR/Cas9 sex distorters. PLoS Genet. 2020;16(3):e1008647. https://doi.org/10.1371/journal.pgen.1008647.
10. Hall AB, Basu S, Jiang X, Qi Y, Timoshevskiy VA, Biedler JK, Sharakhova MV, Elahi R, Anderson MAE, Chen XG, Sharakhov IV, Adelman ZN, Tu Z. Sex determination. A male-determining factor in the mosquito Aedes aegypti. Science. 2015;348(6240):1268–70. https://doi.org/10.1126/science.aaa2850.
11. Krzywinska E, Dennison NJ, Lycett GJ, Krzywinski J. A maleness gene in the malaria mosquito Anopheles gambiae. Science. 2016;353(6294):67–9. https://doi.org/10.1126/science.aaf5605.
12. Sharma A, Heinze SD, Wu Y, Kohlbrenner T, Morilla I, Brunner C, Wimmer EA, van de Zande L, Robinson MD, Beukeboom LW, Bopp D. Male sex in houseflies is determined by Mdmd, a paralog of the generic splice factor gene CWC22. Science. 2017;356(6338):642–5. https://doi.org/10.1126/science.aam5498.
13. Meccariello A, Salvemini M, Primo P, Hall B, Koskinioti P, Dalíková M, Gravina A, Gucciardino MA, Forlenza F, Gregoriou ME, Ippolito D, Monti SM, Petrella V, Perrotta MM, Schmeing S, Ruggiero A, Scolari F, Giordano E, Tsoumani KT, Marec F, Windbichler N, Arunkumar KP, Bourtzis K, Mathiopoulos KD, Ragoussis J, Vitagliano L, Tu Z, Papathanos PA, Robinson MD, Saccone G. Maleness-on-the-Y (MoY) orchestrates male sex determination in major agricultural fruit fly pests. Science. 2019;365(6460):1457–60. https://doi.org/10.1126/science.aax1318.
14. Aryan A, Anderson MAE, Biedler JK, Qi Y, Overcash JM, Naumenko AN, Sharakhova MV, Mao C, Adelman ZN, Tu Z. Nix alone is sufficient to convert female Aedes aegypti into fertile males and myo-sex is needed for male flight. Proc Natl Acad Sci U S A. 2020;117(30):17702–9. https://doi.org/10.1073/pnas.2001132117.

15. Simoni A, Hammond AM, Beaghton AK, Galizi R, Taxiarchi C, Kyrou K, Meacci D, Gribble M, Morselli G, Burt A, Nolan T, Crisanti A. A male-biased sex-distorter gene drive for the human malaria vector *Anopheles gambiae*. Nat Biotechnol. 2020;38(9):1054–60. https://doi.org/10.1038/s41587-020-0508-1.

16. Hamilton WD. Extraordinary sex ratios. A sex-ratio theory for sex linkage and inbreeding has new implications in cytogenetics and entomology. Science. 1967;156(3774):477–88. https://doi.org/10.1126/science.156.3774.477.

17. Alcalay, Y., Fuchs, S., Galizi, R., Bernardini, F., Haghighat-Khah, R.E., Rusch, D.B., Adrion, J.R., Hahn, M. W, Tortosa, P., Papathanos, P.A. The potential for a released autosomal X-shredder becoming a driving-Y chromosome and invasively suppressing wild populations of malaria mosquitoes. https://www.biorxiv.org/content/10.1101/860551v3 (2019).

18. Liquido NJ, Cunningham RT, Nakagawa S. Host plants of Mediterranean fruit fly (Diptera: Tephritidae) on the island of Hawaii (1949–1985 survey). J Econ Entomol. 1990;83(5):1863–78. https://doi.org/10.1093/jee/83.5.1863.

19. Klassen W, Curtis CF. History of the sterile insect technique. In: Dyck VA, Hendrichs J, Robinson AS, editors. Sterile insect technique. Principles and practice in area-wide integrated pest management. Dordrecht: Springer; 2005. p. 3–36.

20. Shelly TE, Epsky N, Jang EB, Reyes-Flores J, Vargas RI, editors. Trapping and the detection, control and regulation of tephritid fruit flies. Netherlands: Springer; 2014.

21. Aumann RA, Häcker I, Schetelig MF. Female-to-male sex conversion in Ceratitis capitata by CRISPR/Cas9 HDR-induced point mutations in the sex determination gene transformer-2. Sci Rep. 2020;10(1):18611. https://doi.org/10.1038/s41598-020-75572-x.

22. Meccariello A, Monti SM, Romanelli A, Colonna R, Primo P, Inghilterra MG, del Corsano G, Ramaglia A, Iazzetti G, Chiarore A, Patti F, Heinze SD, Salvemini M, Lindsay H, Chiavacci E, Burger A, Robinson MD, Mosimann C, Bopp D, Saccone G. Highly efficient DNA-free gene disruption in the agricultural pest Ceratitis capitata by CRISPR-Cas9 ribonucleoprotein complexes. Sci Rep. 2017;7(1):10061. https://doi.org/10.1038/s41598-017-10347-5.

23. Papathanos PA, Windbichler N. Redkmer: an assembly-free pipeline for the identification of abundant and specific X- chromosome target sequences for X-shredding by CRISPR endonucleases. CRISPR J. 2018;1(1):88–98.

24. Hall AB, Qi Y, Timoshevskiy V, Sharakhova MV, Sharakhov IV, Tu Z. Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. BMC Genomics. 2013;14(1): 273. https://doi.org/10.1186/1471-2164-14-273.

25. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337(6096):816–21. https://doi.org/10.1126/science.1225829.

26. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, Volz SE, Joung J, van der Oost J, Regev A, Koonin EV, Zhang F. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell. 2015;163(3):759–71. https://doi.org/10.1016/j.cell.2015.09.038.

27. Anzalone AV, Koblan LW, Liu DR. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. Nat Biotechnol. 2020;38(7):824–44. https://doi.org/10.1038/s41587-020-0561-9.

28. Franz G, Gencheva E, Kerremans P. Improved stability of genetic sex-separation strains for the Mediterranean fruit fly, Ceratitis capitata. Genome. 1994;37(1):72–82. https://doi.org/10.1139/g94-009.

29. McKenna A, Shendure J. FlashFry: a fast and flexible tool for large-scale CRISPR target design. BMC Biol. 2018;16(1):74. https://doi.org/10.1186/s12915-018-0545-0.

30. Scolari F, Schetelig MF, Bertin S, Malacrida AR, Gasperi G, Wimmer EA. Fluorescent sperm marking to improve the fight against the pest insect Ceratitis capitata (Wiedemann; Diptera: Tephritidae). New Biotechnol. 2008; 25(1):76–84. https://doi.org/10.1016/j.nbt.2008.02.001.

31. Gomulski LM, Pitts RJ, Costa S, Saccone G, Torti C, Polito LC, Gasperi G, Malacrida AR, Kafatos FC, Zwiebel LJ. Genomic organization and characterization of the white locus of the Mediterranean fruitfly, Ceratitis capitata. Genetics. 2001;157(3):1245–55.

32. Ward, C. M., Aumann, R.A., Whitehead, M.A., Nikolouli, K., Leveque, G., Gouvi, G., Fung, E., Reiling, S.J., Djambazian, H., Hughes, M.A., Whiteford, S., Caceres-Barrios, C., Nguyen, T.N.M., Choo, A., Crisp, P., Sim, S., Geib, S., Marec, F., Häcker, I., Ragoussis, J., Darby, A.C., Bourtzis, K., Baxter, S.W., Schetelig, M.F. A generic white pupae sex selection phenotype for insect pest control. https://www.biorxiv.org/content/10.1101/2020.05.08.076158v2.full.pdf. (2020).

33. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117(17):9451–7. https://doi.org/10.1073/pnas.1921046117.

34. McCarty NS, Graham AE, Studena L, Ledesma-Amaro R. Multiplexed CRISPR technologies for gene editing and transcriptional regulation. Nat Commun. 2020;11(1):1281. https://doi.org/10.1038/s41467-020-15053-x.

35. Meiklejohn CD. Heterochromatin and genetic conflict. Proc Natl Acad Sci U S A. 2016;113(15):3915–7. https://doi.org/10.1073/pnas.1603015113.

36. Price TAR, Windbichler N, Unckless RL, Sutter A, Runge JN, Ross PA, Pomiankowski A, Nuckolls NL, Montchamp-Moreau C, Mideo N, Martin OY, Manser A, Legros M, Larracuente AM, Holman L, Godwin J, Gemmell N, Courret C, Buchman A, Barrett LG, Lindholm AK. Resistance to natural and synthetic gene drive systems. J Evol Biol. 2020;33(10):1345–60. https://doi.org/10.1111/jeb.13693.

37. Schetelig MF, Caceres C, Zacharopoulou A, Franz G, Wimmer EA. Conditional embryonic lethality to improve the sterile insect technique in Ceratitis capitata (Diptera: Tephritidae). BMC Biol. 2009;7(1):4. https://doi.org/10.1186/1741-7007-7-4.

38. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36. https://doi.org/10.1101/gr.215087.116.

39. Holmes DS, Bonner J. Preparation, molecular weight, base composition, and secondary structure of giant nuclear ribonucleic acid. Biochemistry. 1973; 12(12):2330–8. https://doi.org/10.1021/bi00736a023.

40. Eckermann KN, Ahmed HMM, KaramiNejadRanjbar M, Dippel S, Ogaugwu CE, Kitzmann P, Isah MD, Wimmer EA. Hyperactive piggyBac transposase improves transformation efficiency in diverse insect species. Insect Biochem Mol Biol. 2018;98:16–24. https://doi.org/10.1016/j.ibmb.2018.04.001.

41. Sagri E, Koskinioti P, Gregoriou ME, Tsoumani KT, Bassiakos YC, Mathiopoulos KD. Housekeeping in tephritid insects: the best gene choice for expression analyses in the medfly and the olive fly. Sci Rep. 2017;7(1): 45634. https://doi.org/10.1038/srep45634.

42. Ahmed M, Kim DR. pcr: an R package for quality assessment, analysis and testing of qPCR data; 2018. https://doi.org/10.7717/peerj.4473.

43. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Searching for SNPs with cloud computing. Genome Biol. 2009;10(11):R134. https://doi.org/10.1186/gb-2009-10-11-r134.

44. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12(8):1269–76. https://doi.org/10.1101/gr.88502.

45. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21(Suppl 1):i351–8. https://doi.org/10.1093/bioinformatics/bti1018.

46. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80. https://doi.org/10.1093/nar/27.2.573.

47. Noe L, Kucherov G. YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res. 2005;33(Web Server):W540–3. https://doi.org/10.1093/nar/gki478.

48. Smit A, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

49. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70. https://doi.org/10.1093/bioinformatics/btr011.

Publisher's Note