**The CRISPR Journal**

## RESEARCH ARTICLE

# Redkmer: An Assembly-Free Pipeline for the Identification of Abundant and Specific X-Chromosome Target Sequences for X-Shredding by CRISPR Endonucleases

Philippos Aris Papathanos[1] and Nikolai Windbichler[2]

## Abstract

CRISPR-based synthetic sex ratio distorters, which operate by shredding the X-chromosome during male meiosis, are promising tools for the area-wide control of harmful insect pest or disease vector species. X-shredders have been proposed as tools to suppress insect populations by biasing the sex ratio of the wild population toward males, thus reducing its natural reproductive potential. However, to build synthetic X-shredders based on CRISPR, the selection of gRNA targets, in the form of high-copy sequence repeats on the X chromosome of a given species, is difficult, since such repeats are not accurately resolved in genome assemblies and cannot be assigned to chromosomes with confidence. We have therefore developed the redkmer computational pipeline, designed to identify short and highly abundant sequence elements occurring uniquely on the X chromosome. Redkmer was designed to use as input minimally processed whole genome sequence data from males and females. We tested redkmer with short- and long-read whole genome sequence data of *Anopheles gambiae*, the major vector of human malaria, in which the X-shredding paradigm was originally developed. Redkmer established long reads as chromosomal proxies with excellent correlation to the genome assembly and used them to rank X-candidate kmers for their level of X-specificity and abundance. Among these, a high-confidence set of 25-mers was identified, many belonging to previously known X-chromosome repeats of *Anopheles gambiae*, including the ribosomal gene array and the selfish elements harbored within it. Data from a control strain, in which these repeats are shared with the Y chromosome, confirmed the elimination of these kmers during filtering. Finally, we show that redkmer output can be linked directly to gRNA selection and off-target prediction. In addition, the output of redkmer, including the prediction of chromosomal origin of single-molecule long reads and chromosome specific kmers, could also be used for the characterization of other biologically relevant sex chromosome sequences, a task that is frequently hampered by the repetitiveness of sex chromosome sequence content.

## Introduction

Every year, more than half a million people die globally through the activities of blood-feeding insects. Insects also undermine our global prosperity significantly through losses in agricultural productivity. With a growing population, a changing climate, and spreading insecticide resistance, the burden these species pose on human health and society is expected to increase even further. The long-term sustainability of current approaches to insect control is now, more than ever, questionable, urging the need for the development of new tools. Genetic control is an alternative method of insect control based on the release of laboratory-modified insects, which, through mating with the wild population, transmit materials that directly reduce their potential to do harm. Over the last few years, there has been a steep rise in the research, development, and application of a number of new genetic control strategies, including several involving CRISPR[*]-

---

[*]**C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats.

[1]*Department of Experimental Medicine, Section of Genomics and Genetics, University of Perugia, Perugia, Italy;* [2]*Department of Life Sciences, Imperial College London, Sir Alexander Fleming Building, South Kensington Campus, London, United Kingdom.*

Address correspondence to: *Philippos Aris Papathanos, Department of Experimental Medicine, Section of Genomics and Genetics, University of Perugia, Via Gambuli, Edificio D, 3° Piano, 06132 Perugia, Italy,* E-mail: p.papathanos@gmail.com

based gene drives. One of these are synthetic sex ratio distorters, and we have previously shown in the malaria mosquito *Anopheles gambiae* that these can be rationally engineered using endonuclease-mediated X-chromosome shredding during spermatogenesis.[1,2] Sex ratio distortion is a useful phenotype to engineer in pest or vector insects, as it can lead to population suppression in a manner that, as predicted,[3] is more efficient than the classical sterile insect technique. Furthermore, X-shredders are attractive options for vector control because physically linking these to a Y chromosome can provide both the allele itself and the entire Y chromosome that harbors it a competitive advantage in inheritance against the X.[4] Through this benefit, such an engineered Y chromosome can display genetic drive, increasing in frequency within a population rapidly, starting from low frequencies, biasing the sex ratio toward males as it increases. As a result, the reproductive potential of the population diminishes for the lack of females. Because a large number of sites are targeted simultaneously through X-shredding, the development of resistance mechanisms is significantly impaired, unlike other strategies involving endonucleases for population suppression.[5,6]

Significant interest has developed in the genetic control community to engineer such synthetic sex ratio distorters in additional insect pests or disease vectors, since X-shredding exploits and manipulates the near-universal significance of paternal chromosome inheritance on the sex of an individual. There are a number of potential vector or agricultural pest species that are being considered for this, including those in the *Bactrocera* or *Anastrepha* genuses, the Mediterranean fruit-fly *Ceratitis capitata*, grain beetles, *Drosophila suzukii*, and other mosquito species, including those in the *Anopheles* and *Aedes* genuses. There are five essential requirements required to do so: (1) an XY male karyotype, (2) genetic transformation, (3) regulatory elements (promoters) that can drive expression of the X-shredding nuclease during spermatogenesis, (4) an endonuclease platform such as the CRISPR-Cas9 system that can be directed against X-chromosome-specific sequences, and finally (5) the existence of sequences on the X chromosome that are both specific and abundant to it.

In our previous work in the malaria mosquito, we were able to build an X-shredder because this mosquito's rDNA genes are exclusively located on the X chromosome. However, this arrangement is exceptional, and the vast majority of insects do not to share it. Furthermore, knowledge of naturally occurring multi-copy X-specific sequences, for example X-specific satellite DNA, is limited because such repetitive DNA sequences are *ipso facto* excluded from genome assemblies and because few studies deal with such elements, particularly in non-model organisms. Indeed, even after 15 years since the publication of the first genome assembly of *An. gambiae*,[7] the rDNA cluster is still not correctly represented in the current genome assembly. Knowledge of the rDNAs' specificity to the X chromosome in *An. gambiae* came from studies of mosquito population genetics using cytology. To address this limitation, we have developed a bioinformatic pipeline called redkmer for repeat extraction and detection based on k-mers, which, by utilizing long- and short-read sequencing technologies, is able to identify highly abundant, X-specific sequences in the absence of a genome assembly.

## Methods

### Data requirements

Redkmer requires as input whole genome sequencing (WGS) data based on long single-molecule (e.g., PacBio) and short (e.g., Illumina 100 bp) reads. For the long reads, WGS must be performed using male-only or mixed-sex samples, and reads must be self-error corrected, for example using Canu,[8] and provided in fasta format. For the short-read libraries, data must be generated from both male and female samples independently and provided in fastq format as a single file (paired-end reads can be merged into one file for each sex). A fasta file with the mitochondrial reference genome is also required to remove short reads derived from it. Redkmer is optimized to run with data sets that achieve at least a 10× genome coverage of the long-read library and a 20× genome coverage of each short-read library.

### Software and hardware requirements

Redkmer has been implemented to run on UNIX HPC systems scheduled by SGE or PBS. All steps were run using the cx1 general purpose cluster at the Imperial College High Performance Computing Service. All alignment steps allow splitting of the long read input data via the [NODES] parameter for parallel execution, and have been run on standard 12 and 24 core nodes with 32GB of memory. Processing steps require up to 120GB of memory depending on the size of the input libraries. The following third-party modules are required and loaded by redkmer: bowtie1,[9] bowtie2,[10] FastQC,[11] jellyfish,[12] samtools,[13] BLAST,[14] and the R environment for plotting, which utilizes ggplot2 and data table modules.

### Pipeline overview

Redkmer is designed to identify short 25 bp sequences (kmers) occurring abundantly and specifically on X chromosomes of species with XY male karyotypes. The pipeline

is designed around three basic principles for target-site identification and selection: (1) X-chromosome kmers are first identified by assessing differential representation in female versus male WGS data, also known as chromosome quotient (CQ);[15,16] (2) kmers that occur on other chromosomes are subsequently eliminated; and (3) those that are the most abundant within those displaying X-chromosome specificity are selected. In the first phase of the pipeline, redkmer uses unassembled error-corrected reads from single-molecule WGS (called PacBio from here for simplicity) as chromosomal proxies, by mapping to them short reads from separate female and male WGS libraries (called Illumina from here for simplicity).

The ratio of mapping female and male Illumina reads, also known as the CQ, is used to predict chromosomal origin, where reads originating from the X chromosome typically display CQ around 2, whereas reads originating from autosomes (equally represented in female and males typically display a CQ of $\sim 1$) and reads from the Y that are uniquely represented in male data have an average CQ of $\sim 0$. Thus, based on the CQ of each PacBio read, it is assigned to one of four possible "chromosomal bins" (autosomal, X, Y, and genome amplification [GA]), which are populated by the reads themselves and that act as chromosomal proxies in the absence of a genome assembly.

Because CQ is calculated through the mapping of many Illumina reads to a relatively longer sequence, the confidence in the predicted chromosomal origin of PacBio reads is higher compared to CQ calculated over the span of only the target site. Furthermore, because the PacBio read length is long (depending on the quality of the data set), redkmer can distinguish between reads derived entirely from the X chromosome from those that are autosomal in origin but which harbor shorter sequences homologous to those on the X chromosome, through flanking autosomal sequences that reduce the overall average CQ of the read. In the second phase, kmers generated from the Illumina libraries are mapped to PacBio reads of each chromosomal bin separately. Kmers mapping above a defined threshold to non-X derived Pacbio reads compared to X-reads are tagged as unsuitable for X-shredding.

Redkmer also generates data regarding off-targeting potential of candidate X-specific kmers by searching for degenerate target sites in non-X-chromosome-derived reads. The final output of the pipeline, a list of suitable kmers along with their specificity and abundance profiles, can then be used for the purpose of building and testing of X-shredder constructs, by running the selected kmers through external computational tools designed to classify and predict their suitability for RNA-guided nuclease platforms, for example for CRIPR-Cas9 by searching for suitable PAM sites with the 25bp kmers (see below).

## Implementation and module functionality

Redkmer is freely available in Github (github.com/genome-traffic/redkmer-hpc) under the GNU General Public License v3, June 29, 2007. Redkmer implements 10 core modules (Fig. 1) and is designed to run on a high-performance computing platform. All configurable parameters and run settings are controlled from the file *redkmer.cfg* that sets redkmer behavior. Redkmer first runs quality control and filtering of the input data. Reads from Illumina libraries are mapped to the mitochondrial genome using bowtie2,[10] and aligned reads are removed from each library. This step is required because mitochondrial sequences can behave similarly to X-chromosomal sequences when redkmer calculates CQ in some species.

We found this when running redkmer with *Drosophila melanogaster* WGS data (data not shown), whereby reads or kmers derived from the mitochondrial genome had a higher coverage in females compared to males. Because we did not observe this in the mosquito data, we believe that this is most likely due to the inclusion of germline tissues, in which active oogenesis in females can result in higher total number of mitochondria, unlike for example unfed female mosquito samples in which oogenesis is arrested until blood-feeding. The quality of the filtered Illumina reads is then checked with FastQC,[11] and reports are generated for the user.

PacBio reads are then filtered for read length, removing those of insufficient length to predict chromosomal origin reliably (Module 1). Redkmer then assigns the PacBio reads to chromosomal bins by separately mapping Illumina reads from the male and female libraries to the filtered PacBio reads using bowtie1,[9] allowing no mismatches throughout the length of the alignment, reporting all possible alignments, and normalizing the number of mapping reads for the library sizes (Modules 2–3). The initial kmer sets are generated from the female and male Illumina WGS libraries separately using jellyfish,[12] with which redkmer counts kmer abundance in both libraries and calculates their ratio of counts (effectively kmer-CQ) normalized by library size (Modules 4–5). All kmers are then mapped to reads in each chromosomal bin using bowtie1.[9]

The number of hits to each chromosomal bin is counted, and through this, redkmer derives the X-specific index (XSI) as the ratio of X-chromosome over non-X hits. Therefore, redkmer selection does not exclude kmers with perfect off-targets (100% matches of the kmer to the non-X bins) using an arbitrary cutoff, but instead uses the proportional cutoff that accounts for the
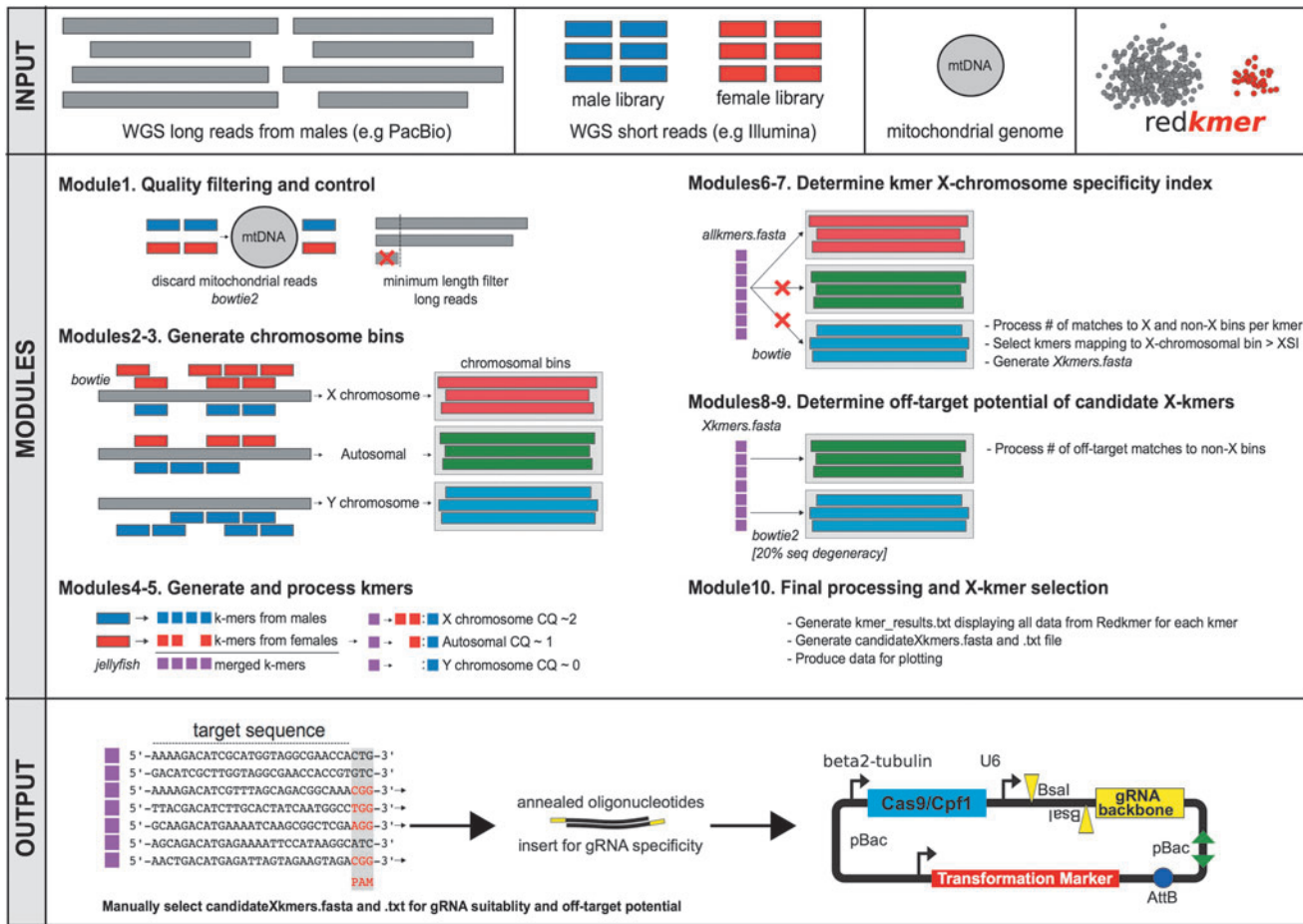
**FIG. 1.** Schematic overview of the redkmer pipeline, showing required input data, module main functions, and potential downstream activities.

total number of kmer hits to non-X reads (Modules 6–7). Kmers with a XSI that pass a selected threshold (e.g., 0.99% to <1% of hits tolerated to non-X reads) are then remapped to all non-X long reads, this time allowing 20% mismatches over the length of the alignment to identify ''degenerate'' off-targets (Modules 8–9). Redkmer finally processes the outputs of all modules and generates fasta and tab-delimited list of candidate kmers for X-shredding (Module 10). There are a number of supplementary modules that have been built to evaluate redkmer output, for example against an available genome assembly or set of reference genes, located in the Supplementary Data (Supplementary Data are available online at www.liebertpub.com/crispr).

## Results

### Redkmer target selection in *An. gambiae*

To test redkmer target selection, we ran it using WGS data from the Pimperena strain of *An. gambiae*, which are publicly available at the Sequence Read Archive (SRS667972, SRR1509742, and SRR1508169). Annotated genomic data, for example the mitochondrial genome and the genome assembly AgamP4, were based on the PEST strain and retrieved from Vectorbase. We selected *An. gambiae* for redkmer evaluation because the rDNA gene cluster and the repetitive elements residing within it are already known to be both X-specific and abundant and have been experimentally shown to be well-suited for X-shredding using both homing endonucleases as well as CRISPR-based endonucleases. In addition, separate male and female Illumina WGS data are also available for the Asembo1 strain of *An. gambiae* (SRR1504990 and SRR1504983), which was used here as a control, as this strain harbors the ribosomal gene cluster on both sex chromosomes, likely as a result of a rare X–Y recombination event.[17]

### Pacbio reads as chromosome proxies

Of the initial ∼4.3 million error-corrected input PacBio reads, redkmer retained ∼2 million reads that passed

the minimum length cutoff (default setting of 2 kbp), resulting in a total of $7.4 \times 10^9$ sequenced nucleotides (25× coverage assuming a 300 Mbp genome). The filtered PacBio reads were assigned to one of the four chromosome bins based on read CQ from Modules 2–3 (Fig. 2A). PacBio read CQ and coverage (measured in length normalized sum of Illumina reads from males and females mapping to each PacBio read – LSum) indicated that both autosomes and the sex chromosomes harbor numerous repetitive elements (Fig. 2B). Importantly, the high density of PacBio reads with high LSum and with a CQ close to two indicated an abundance of repeats on the X chromosome not shared with other chromosomes (Fig. 2B). A significant co-occurrence of X-linked repeats on the autosomes or the Y chromosome would result in significant shift of the CQ toward one, which was not apparent for the repeat-containing PacBio reads of the X-bin (Fig. 2B). Overall, PacBio reads from the Y-chromosome bin displayed the highest level of repetitive DNA content, followed by X chromosomes and then the autosomes (Fig. 2C), consistent with published data.[15]

Reads belonging to the GA bin had the lowest LSum values, consistent with the explanation that reads with CQ significantly higher than two would be expected to result from either sample-specific sequencing artifacts or bacterial contamination (Fig. 2C). Additional redkmer generated plots, providing data on basic statistics of the input PacBio reads library, including the distribution of CQ, LSum, Sum (sum of mapping reads from males and females prior to normalization), and read length, can be found in Supplementary Figures S1–S7. These redkmer-generated plots and additional useful data sets, for example those useful for the characterization of other sex chromosome sequences, such as read IDs and sequences assigned to each of the PacBio chromosomal bins, are collated and deposited by redkmer in the pac-bio_bins/fasta folder (Supplementary Table S1).

To evaluate CQ-based prediction of chromosome origin, we mapped PacBio reads from each bin by BLAST[14] to the latest *An. gambiae* genome assembly (Table 1 and Fig. 3) in a non-exclusive manner, reporting all matches of ≥2 kb. Hits to the 42.39 Mbp long ''unknown chromosome,'' which contains all repeat-rich scaffolds that have not been anchored to chromosomes by physical mapping, accounted for 89.5% all hits between the PacBio library and the assembly. Of the remaining alignments, 81.3% of hits from X-bin reads were to the X-chromosome assembly and 18% to autosomal arms. From the reads in the A-bin, 94.6% had hits to autosomal arms and only 5.4% to the X-chromosome assembly, with most of
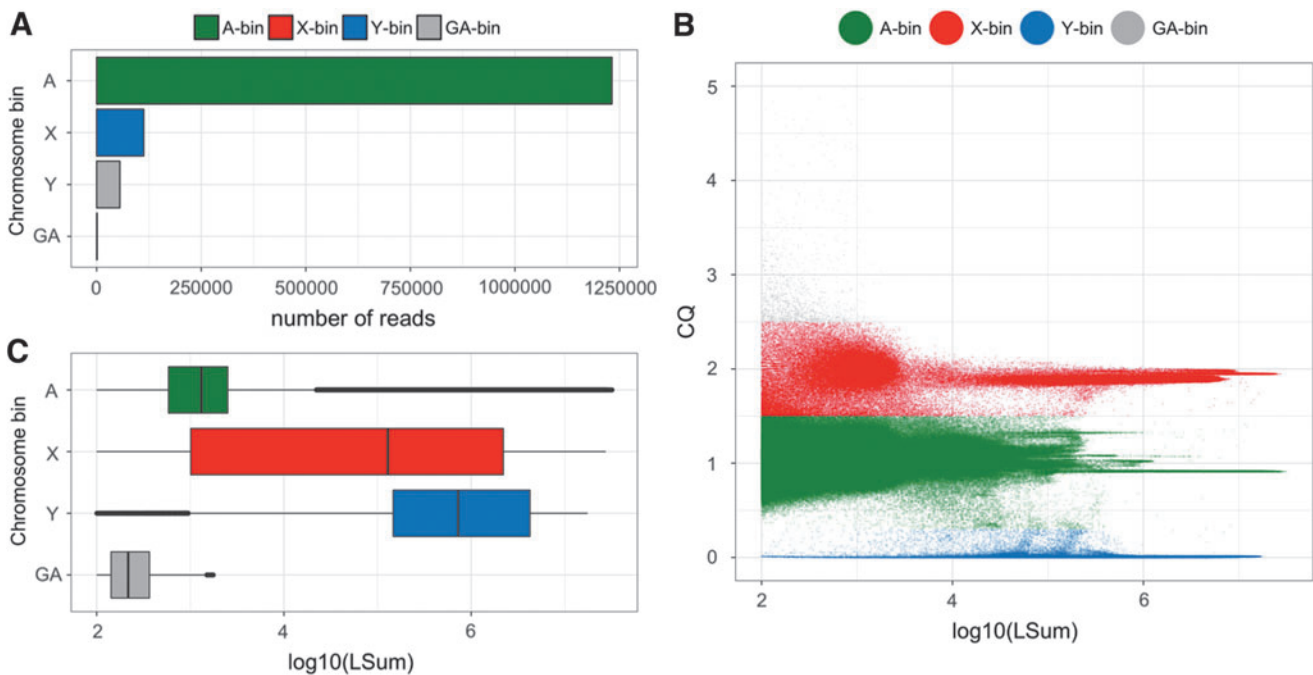


**FIG. 2.**   Long-read assignment into chromosomal bins using the *Anopheles gambiae* Pimperena strain. **(A)** Number of PacBio reads assigned to each chromosome bin. **(B)** PacBio read CQ (ratio of female/male data) over log10 of LSum (total number of mapping reads from male and female Illumina data) showing chromosomal bins and chromosome repetitiveness. **(C)** Box plots of PacBio coverage/repetitiveness for each chromosomal bin.

**Table 1. Number of hits of PacBio reads to *An. gambiae* genome assembly**

|  | # reads | 2R | 2L | 3R | 3L | X | Mt | UN | Y |
|---|---|---|---|---|---|---|---|---|---|
| A bin | 1,231,987 | 426,618 | 339,163 | 413,311 | 2,767,42 | 82,397 | 5 | 11,747,985 | 481 |
| X bin | 112,643 | 3,513 | 2,107 | 7,091 | 2,526 | 66,134 | 0 | 4,710,906 | 1 |
| Y bin | 55,541 | 8,695 | 1,231 | 1,716 | 623 | 6,970 | 1 | 12,378 | 293,974 |
| GA bin | 1,300 | 6 | 6 | 12 | 2 | 947 | 0 | 3,482 | 0 |
| X bin masked[a] | 112,643 | 299 | 6 | 2 | 9 | 34,248 | 0 | 9,438 | 0 |

[a]X bin after masking genome with reads from A bin.

these mapping in the 20–25 Mbp repeat-rich region of the X assembly, which is known to contain repeats that are shared between the X and the Y chromosome, which drives CQ to autosomal levels (Fig. 3).[15] Since redkmer excludes target sequences that also occur on autosomes or the Y, we next repeated the blast of the X-bin PacBio reads against the genome assembly, after first masking regions of the assembly that have significant similarity to PacBio reads assigned to the A bin (Table 2). After masking, there remained a significant proportion of hits (27.3%) from reads of the X bins against the ''unknown chromosome,'' highlighting that a significant portion of this pseudo-assembly is composed of non-autosomal se-

quences, or autosomal repeats not represented in the assembly. Of the hits between the X-bin reads, 99% mapped to the X-chromosome assembly, and only 0.7% mapped to the autosomal assembly (Table 2). These results indicate that CQ-based prediction of chromosomal origin combined with the exclusion of sequences that overlap with reads assigned to the autosomal bin can reliably identify X-chromosome-specific sequences.

## Candidate X-kmer selection

Redkmer generated 270 million kmers that passed the minimum occurrence threshold in the Illumina data from both males and females (kmernoise = 5). The CQ
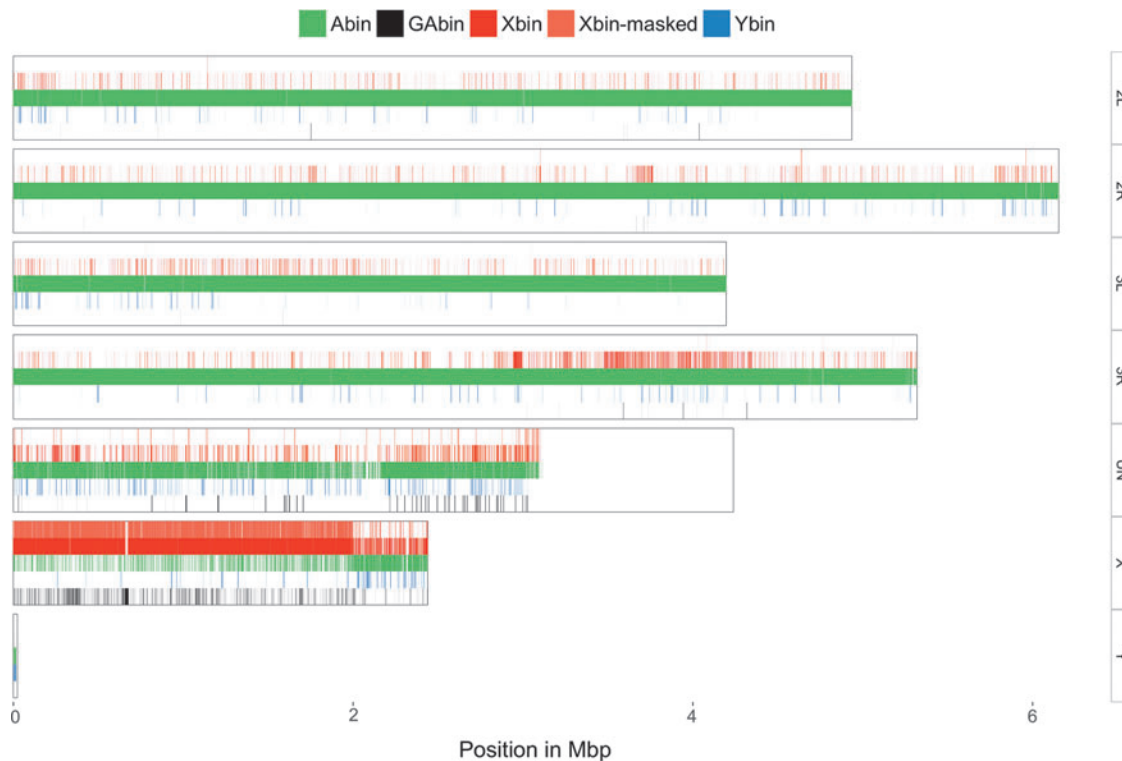


**FIG. 3.** Validation of the long-read bin assignment of the Pimperena strain using the AgamP4 genome assembly. Painting of the *An. gambiae* chromosome assembly with blastn matches >2 kbp from reads of each chromosomal bin. The mitochondrial genome is not shown here. The apparent absence of matches between PacBio reads and the "unknown chromosome" downstream of ~3 Mbp results from scaffold length being shorter than the minimum 2 kbp alignment length cutoff.

**Table 2. Number of alignments to the *An. gambiae* PEST assembly**

|  | Candidate X-kmers | Sum off-targets >0 | Sum off-targets = 0 |
|---|---|---|---|
| All | 25,742 | 19,683 | 6,059 |
| 2R | 916 | 912 | 4 |
| 2L | 418 | 416 | 2 |
| 3R | 1,038 | 781 | 257[a] |
| 3L | 249 | 249 | 0 |
| Mt | 0 | 0 | 0 |
| UN | 20,295 | 17,488 | 2,807 |
| X | 16,035 | 11,312 | 4,723 |
| Y | 0 | 0 | 0 |

[a]All 257 kmers matches to 3R map exclusively within genes *AGAP029007* and *AGAP029004*, whose position in the Pimperena strain on chromosome 3 could not be supported. This is possibly a mis-assembly, even in the PEST strain.

and kmer abundance patterns indicated chromosomal repetitiveness profiles similar to those of the PacBio reads (Fig. 4A). Additional redkmer-generated plots providing data on basic statistics of the kmer selection can be found in Supplementary Figures S8–S14. Redkmer selected 64,619 kmers (0.02%) as being specific to the X chromosome and within the top 99.5 percentile in kmer abundance (Fig. 4B). To identify among these kmers containing suitable gRNA target sequences for the CRISPR-Cas9 and Cpf1 nucleases, we used the FlashFry tool[18] (github.com/aaronmck/FlashFry), using the candidateXkmers.fasta file as input, and all PacBio reads from the A, Y, and GA-bins combined (1,288,828 reads) as a reference set for performing the CRISPR off-target analysis. Of the candidate X-kmers, 20,566 (32%) were selected as containing suitable gRNA targets for either enzyme, and 669 kmers had sequences suitable for both (Fig. 4B and C). Among the 25-mers passing the final selection step, five contained the 20 bp long gRNA target sequence T1, and 17 contained the 15 bp long I-*Ppo*I recognition sequence sites. We have already validated of these target sequences in independent studies as suitable for inducing sex ratio distortion by X-shredding in *An. gambiae*.[1,2]

To evaluate the selection of candidate X-kmers, we mapped these using BLAST to each arm of the *An. gambiae* genome assembly (Table 2). Of the 64,619 X-kmers, only 25,742 (40%) had hits to the genome assembly, in line with our expectation that many high-copy sequence repeats are poorly represented within genome assemblies. Of the kmers that did have hits within the assembly, 63% mapped to the X chromosome, 78% matched sequences in the unknown chromosome, and 10% had hits to the autosomal arms. Hard filtering of kmers allowing neither perfect nor degenerate off-targets hits to non-X long-read bins reduced the hits to autosomal arms in the as-

sembly to 4.3% (Table 2). Closer inspection of the hits between the X-kmers and the autosomes identified two hotspots on chromosome 3R. All 257 kmers mapped exclusively within two genes, *AGAP029007* and *AGAP029004*, both of which represent partial sequences of the 28S ribosomal RNA locus. This is likely an annotation or assembly artifact of the AgamP4 PEST assembly, as we could find no supporting evidence for either gene at the putative positions on chromosome 3R within the Pimperera PacBio reads—none spanned the chromosomal assembly where these two genes are located, and none of the kmers mapping to these genes mapped at the junctions where genes meet the flanking assembly.

Based on our use of the *An. gambiae* data set, we concluded that redkmer selection, especially after considering the off-target data, is reliable in providing high-confidence target sequences that occur exclusively within the X chromosome. To illustrate this further, we also mapped the 64,619 candidate X-kmers were also mapped against a set of reference sequences using BLAST. As reference sequences, we included the entire X-chromosome assembly, the repeats library of *An. gambiae* from Vectorbase, and the ribosomal DNA cluster. Of these kmers, 61% had no hits to any of the reference sequences, despite being present within the PacBio reads, highlighting that such sequences are underrepresented within annotated data sets. The ribosomal locus and the X-assembly each accounted for 46% of hits (Fig. 5A). With the exception of AgX367, which is a known X-specific satellite in *An. gambiae*,[15,19] all other hits to the reference sequences (22%) were to two families of the site-specific retrotransposons of the R1 and R2 clade that occupy specific positions with 28S rDNA of *An. gambiae* (Fig. 5A and B).[20–22] These results confirm that redkmer is able to identify sequences that are both X-chromosome specific and abundant.

## Validation of the results using control data from the Asembo1 strain

To test redkmer sensitivity to biological variation, we next re-ran the pipeline using Illumina WGS libraries from males and female of the Asembo1 strain of *An. gambiae*. Colonized in 1997 in Asembo, Kenya, the rDNA is believed to have introgressed onto the Y chromosome in this strain, forming a Mopti/Savanna hybrid in males.[17] Cytological evidence confirms that this strain harbors the ribosomal gene cluster not only on the X chromosome but also on the Y.[15] Long-read single-molecule sequencing is not available for the Asembo1 strain, so we evaluated its Illumina data against the Pimperena PacBio data set. We expected that CQ-based prediction of chromosomal origin would result in significantly different chromosome bins, particularly for reads corresponding
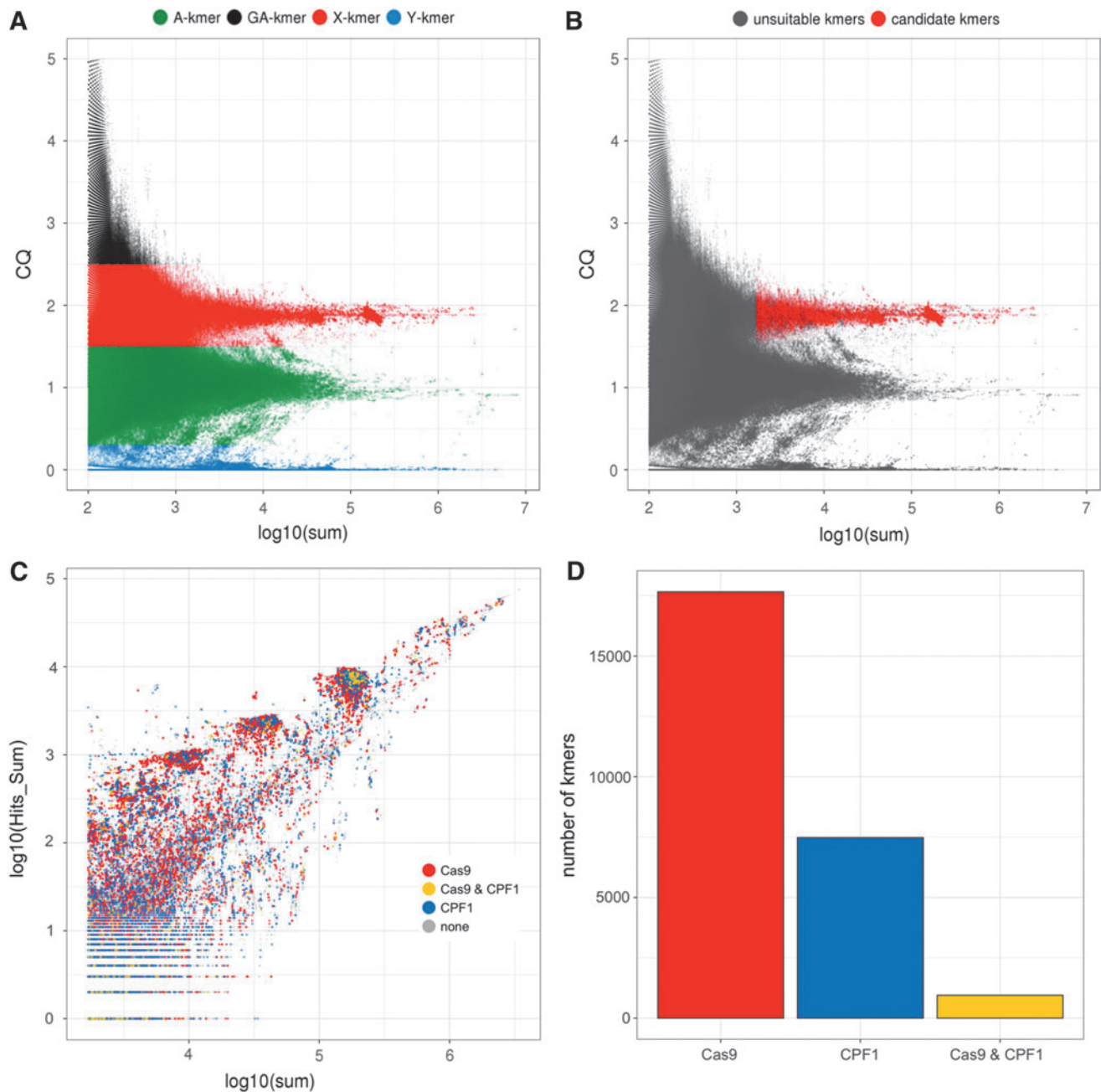
**FIG. 4.** kmer analysis and selection in the Pimperena strain. **(A)** kmer-CQ versus abundance (in log10 of sum) in both male and female Illumina libraries for all 270 million redkmer-generated kmers colored by chromosomal bins. **(B)** Redkmer selection plot showing kmer-CQ versus abundance for predicted X-chromosome-specific and abundant kmers (red dots) versus the unsuitable kmers (gray dots). **(C)** Identification of kmers containing sequences predicted to be suitable for targeting by CRISPR endonucleases showing kmer abundance in the Illumina data (*x*-axis) versus their occurrence within PacBio data (*y*-axis). Kmers lacking suitable characteristics for CRISPR (e.g., the absence of a PAM motif) are shown as gray dots, kmers suitable for Cas9 in red, kmers suitable for Cpf1 in blue, and those harboring sequences suitable for both Cas9 and Cpf1 are shown in yellow. **(D)** Number of kmers containing sequences suitable for CRISPR nuclease platforms.
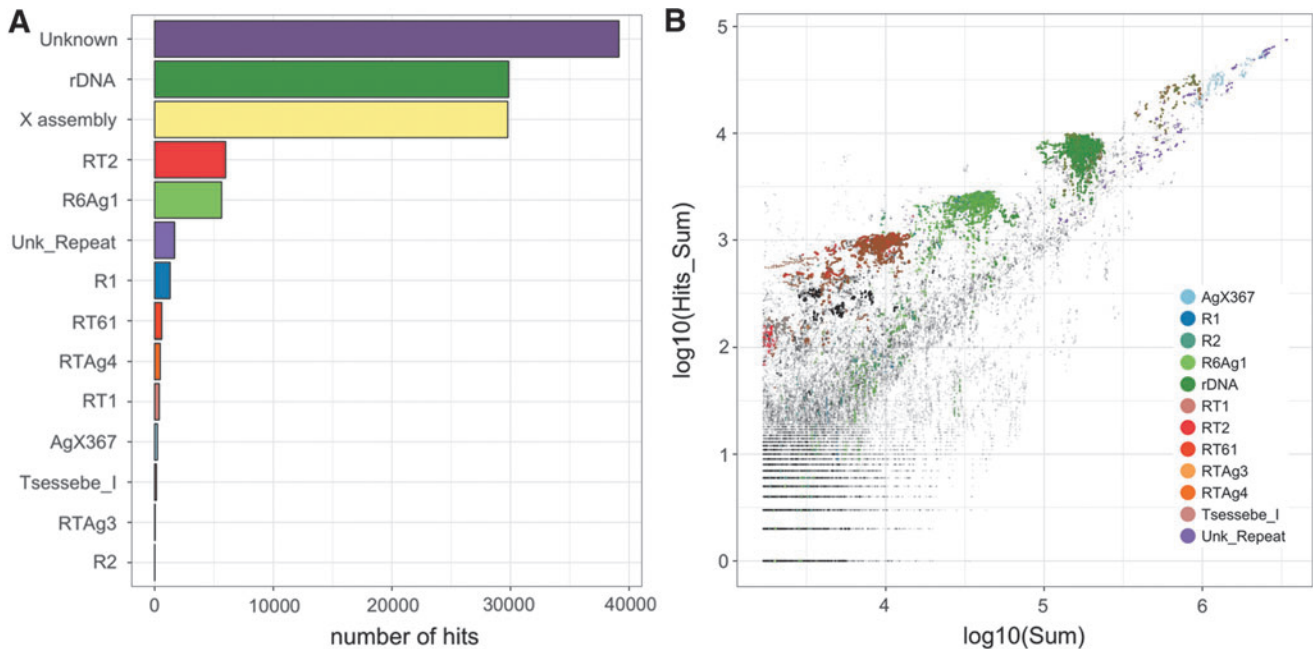
**FIG. 5.** Analysis of the X-candidate kmers of the Pimperena strain. **(A)** Blast results showing matches between selected X-kmers and *An. gambiae* reference sequence collection. **(B)** Coverage of selected X-kmers in the Illumina (log10 of Sum) and PacBio data (log10 of Hits_Sum). Each kmer is colored based on the locus from which it derives. Because X-kmers corresponding to the X-chromosome assembly do not cluster on the plot, their position is shown but not colored separately.

to the rDNA cluster and its associated repeats. The highly repetitive cloud of PacBio reads previously assigned to the X bin using Pimperena Illumina data (Fig. 2B) was absent when mapping was done using the Asembo1 reads (Fig. 6A). Effectively, PacBio reads corresponding to the X-heterochromatic region were now being assigned to the autosome bin, as co-occurrence of sequences on the X and Y chromosome drives CQ to autosomal levels.

Interestingly, the repeat content and abundance Y-chromosome-assigned reads did not indicate that the X–Y recombination event that transferred the ribosomal cluster from the X to the Y was reciprocal (Fig. 6A). Redkmer now identified only 210 kmers as specific to the X chromosome (Fig. 6B). We found by blasting these kmers to the references sequences used above,that 120 of these kmers (57%) had no hits. Of the remaining 90 kmers, 74 had hits to the X-chromosome assembly, 89 matched the RT2 or R7Ag1 retroposons that insert specifically within the rDNA locus, and 16 had hits to the rDNA cluster. The X and Y chromosomal rDNA loci of Asembo1 are known to be polymorphic and contain mixed arrays of both *An. gambiae* and *An. coluzzi* (called M and S in that study).[17] We reasoned that redkmer did not exclude these 94 kmers because these occurred specifically within

the X-chromosome array and not on the Y-chromosome rDNA array. To confirm this, we selected a subset of all kmers matching to the rDNA reference sequence (611 of 261743585 kmers), and found that the majority did indeed display kmer CQ indicative of linkage to both sex chromosomes, but that a small number did retain CQ values >1.5 (the X–CQ cutoff; Fig. 6B).

## Discussion

Knowledge of X-chromosome-specific sequences is an essential component for the development of synthetic sex ratio distorters based on X-chromosome shredding. However, for the vast majority of species, this type of information is not available because of the difficulty of characterizing repetitive, heterochromatic DNA, whose characteristics make genome assembly and scaffolding unreliable. The advent of next-generation sequencing and, more recently, single-molecule sequencing is beginning to provide the technologies required to study the makeup and properties of heterochromatic sequences. For example, we have previously shown that combining long single-molecule data with short Illumina sequencing can be a successful strategy to characterize both the genic and repetitive content of the Y chromosome of mosquitoes,[15] and similar methods are now
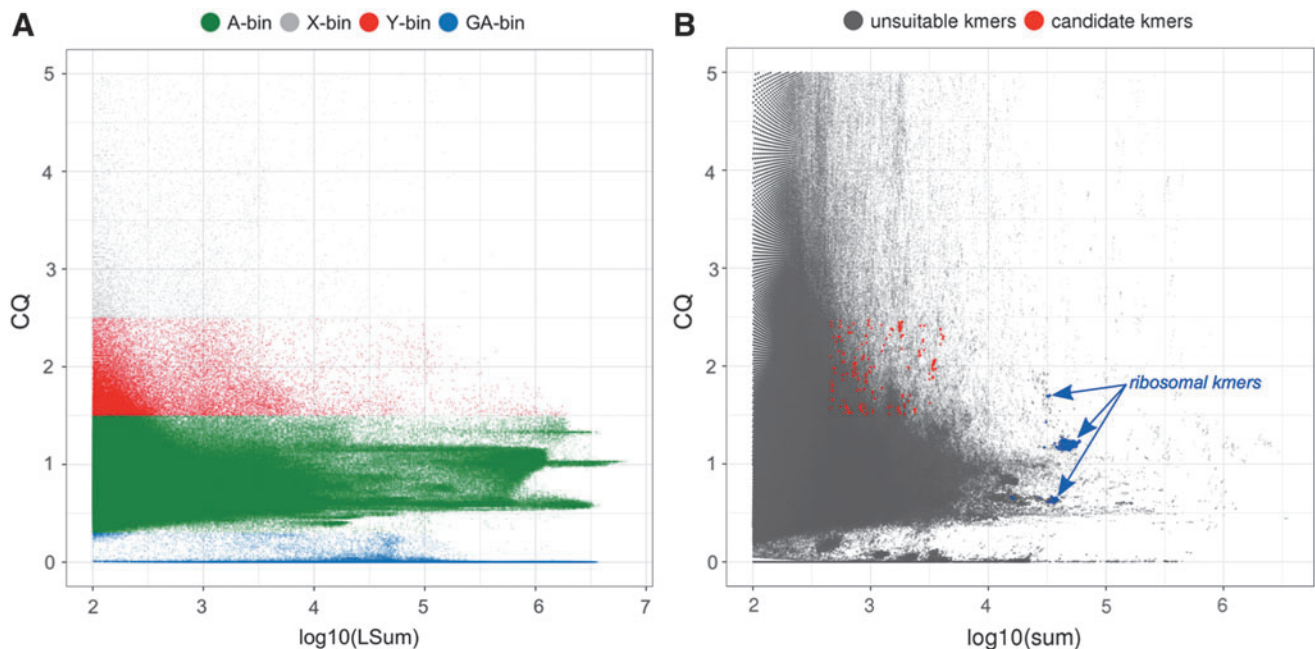
**FIG. 6.** Analysis of the Asembo1 strain. **(A)** PacBio read CQ (ratio of female/male data) over log10 of LSum (total length-normalized number of mapping reads from male and female illumina data) showing absence of X-specific repeat cloud in the Asembo strain. **(B)** Plot showing kmer-CQ versus abundance for selected X-chromosome specific and abundant kmers (red dots—candidate kmers) and those not suitable passing selection (gray dots). Blue dots indicate kmers mapping to the ribosomal array confirming that this locus is not X-specific in the Asembo1 strain (CQ—autosomal) and thus unsuitable for X-shredding. The 89 kmers that are selected and map to this locus have also been highlighted (top blue arrow).

being applied to other species. However, an approach to tackle X-chromosome-specific repetitive sequences, which is complicated by the fact that X-chromosome sequences, unlike those on the Y chromosome, are not sex specific, has not been developed.

To address this gap, we developed the redkmer pipeline, which is designed to require as input only raw WGS data with minimal filtering and error correction. The main output is a tab-delimited file providing data on kmers selected as X-chromosome specific and abundant, including kmer-CQ, coverage, X-chromosome specificity, and off-targeting data, along with a fasta file for downstream analysis. In addition, redkmer data can be used to infer and identify X- and Y-chromosome sequences within the ''chromosomal bins'' of PacBio reads. Plots describing the PacBio reads and the kmers are also produced at the end of the redkmer pipeline (Supplementary Figs. S1–S14). While neither the assignment of Pacbio reads to chromosomal bins nor the calculation of kmer CQ values are fully reliable on their own, we show that the combination of these two strategies can be used to identify X-chromosome-specific kmer sequences efficiently and reliably.

The *An. gambiae* Pimperena Illumina data set consists of ~140M 100 bp Illumina reads per sex and ~2M Pacbio reads that passed length filtering. The combination of these reads results in >$10^{14}$ possible cross-alignments, which is why redkmer is implemented for parallel execution by splitting of the Pacbio input data. To test the limits of the pipeline, we have also tested redkmer on another data set featuring ~400M Illumina reads per sex and unpublished ~8M Pacbio reads (data not shown), and it was confirmed that the pipeline is able to handle the larger data sets that are now becoming available. We showed that running redkmer with data from *An. gambiae* correctly identified known X-chromosome-specific and abundant sequences, some of which we used previously to develop the first synthetic sex ratio distorters by X-chromosome shredding.[1] We also showed using the control strain Asembo1, in which a large fraction of X-chromosome repetitive sequences are shared with the Y, that redkmer target prediction differs in line with our expectations. Therefore, running redkmer with data from different species in the future may help to identify those that are suitable for the development of X-shredder-based genetic control strategies, and identify target sites for doing so.

## Conclusions

The data presented in this study have shown that a combination of single-molecule and long-read sequencing when combined with short-read WGS data from males and females can be used to identify X-chromosome-specific sequences efficiently and reliably, which can be used to develop X-shredder-based sex ratio distortion systems.

## Author Disclosure Statement

No competing financial interests exist.

## References

1. Galizi R, Doyle LA, Menichelli M, et al. A synthetic sex ratio distortion system for the control of the human malaria mosquito. *Nat Commun* 2014;5:3977. DOI: 10.1038/ncomms4977.
2. Galizi R, Hammond A, Kyrou K, et al. A CRISPR-Cas9 sex-ratio distortion system for genetic control. *Sci Rep* 2016;6:31139. DOI: 10.1038/srep31139.
3. Schliekelman P, Ellner S, Gould F. Pest control by genetic manipulation of sex ratio. *J Econ Entomol* 2005;98:18–34.
4. Hamilton WD. Extraordinary sex ratios. A sex-ratio theory for sex linkage and inbreeding has new implications in cytogenetics and entomology. *Science* 1967;156:477–488. DOI: 10.1126/science.156.3774.477.
5. Hammond AM, Kyrou K, Bruttini M, et al. The creation and selection of mutations resistant to a gene drive over multiple generations in the malaria mosquito. *PLoS Genet* 2017;13:e1007039. DOI: 10.1371/journal.pgen.1007039.
6. Unckless RL, Clark AG, Messer PW. Evolution of resistance against CRISPR/Cas9 gene drive. *Genetics* 2017;205:827–841. DOI: 10.1534/genetics.116.197285.
7. Holt RA, Subramanian GM, Halpern A, et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 2002;298:129–149. DOI: 10.1126/science.1076181.
8. Koren S, Walenz BP, Berlin K, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017;27:722–736. DOI: 10.1101/gr.215087.116.
9. Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10:R25. DOI: 10.1186/gb-2009-10-3-r25.
10. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359. DOI: 10.1038/nmeth.1923.
11. Andrews S, Others. FastQC: a quality control tool for high throughput sequence data, www.bioinformatics.babraham.ac.uk/projects/fastqc/ (last accessed January 26, 2018).
12. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;27:764–770. DOI:10.1093/bioinformatics/btr011.
13. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079. DOI: 10.1093/bioinformatics/btp352.
14. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.
15. Hall AB, Papathanos P-A, Sharma A, et al. Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. *Proc Natl Acad Sci U S A* 2016;113:E2114–2123. DOI: 10.1073/pnas.1525164113.
16. Hall AB, Qi Y, Timoshevskiy V, et al. Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. *BMC Genomics* 2013;14:273. DOI: 10.1186/1471-2164-14-273.
17. Wilkins EE, Howell PI, Benedict MQ. X and Y chromosome inheritance and mixtures of rDNA intergenic spacer regions in *Anopheles gambiae*. *Insect Mol Biol* 2007;16:735–741. DOI: 10.1111/j.1365-2583.2007.00769.x.
18. McKenna A, Shendure J. FlashFry: a fast and flexible tool for large-scale CRISPR target design. *bioRxiv* 2017;189068. DOI: 10.1101/189068.
19. Krzywinski J, Sangaré D, Besansky NJ. Satellite DNA from the Y chromosome of the malaria vector Anopheles gambiae. *Genetics* 2005;169:185–196. DOI: 10.1534/genetics.104.034264.
20. Paskewitz SM, Collins FH. Site-specific ribosomal DNA insertion elements in *Anopheles gambiae* and *A. arabiensis*: nucleotide sequence of gene-element boundaries. *Nucleic Acids Res* 1989;17:8125–8133. DOI: 10.1093/nar/17.20.8125.
21. Besansky NJ, Paskewitz SM, Hamm DM, et al. Distinct families of site-specific retrotransposons occupy identical positions in the rRNA genes of *Anopheles gambiae*. *Mol Cell Biol* 1992;12:5102–5110. DOI: 10.1128/MCB.12.11.5102.
22. Kojima KK, Fujiwara H. Evolution of target specificity in R1 clade non-LTR retrotransposons. *Mol Biol Evol* 2003;20:351–361. DOI: 10.1093/molbev/msg031.