Revisiting Y-chromosome detection methods: R-CQ and KAMY efficiently identify Y chromosome sequences in Tephritidae insect pests

Dimitris Rallis [1], Konstantina T Tsoumani [1], Flavia Krsticevic [2], Philippos Aris Papathanos [2], Kostas D Mathiopoulos [1] & Alexie Papanicolaou [3]

[1] Department of Biochemistry and Biotechnology, University of Thessaly, Larissa, Greece.

[2] Department of Entomology, Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, Rehovot, Israel.

[3] Hawkesbury Institute for the Environment, Western Sydney University, Locked Bag 1797, Penrith NSW 2751, Australia.

Communicating authors: Dimitris Rallis drallis@uth.gr ; Kostas Mathiopoulos kmathiop@bio.uth.gr ; Alexie Papanicolaou a.papanicolaou@westernsydney.edu.au

Abstract

Background: The repetitive and heterochromatic nature of Y chromosomes poses challenges for genome assembly methods which can lead to fragmented or misassembled scaffolds. While new sequencing technologies and assembly techniques becoming popular, tools for improving the generation of an accurate Y chromosome are limited, especially for species, such as insects, with a frequent occurrence of heterochromatic chromosomes.

Results: Two novel Y-detection methods are presented here, R-CQ and KAMY, that revisit the ratio-based Chromosome Quotient and kmer-based Y-Genome Scan methods, respectively. We benchmark R-CQ and KAMY methods against their predecessors, over their ability in identifying Y-derived regions in genome assemblies of two important insect pests of the Tephritidae genus: the olive fruit fly *Bactrocera oleae* and the Mediterranean fruit fly *Ceratitis capitata.* These species are characterised by different Y-chromosome

morphologies and their genomes were sequenced with different methodologies. We also evaluated the efficiency and generic applicability of these methods using suitable *Drosophila melanogaster* genomic data, whose Y-chromosome is the best studied among insects. Furthermore, KAMY was assessed for the capability of identifying Y-derived transcripts in the absence of a reference-Y sequence and effectively identified the Tephritid maleness factor *MoY* in a set of mixed-sex transcriptomic data. Through our work, we describe a methodology for manually curating the computational results, through which the performance of different Y detection methods is determined, together with the size and quality of assembled Y sequences.

Conclusions: We find a variability in the performance of Y-detection methods, that is highly dependent on the sequencing approach used and on the sequence of the Y. Our benchmarking suggests an improved overall efficiency of KAMY and R-CQ compared to their predecessors, while our analysis highlights the importance for manually curating the algorithmic outputs towards describing the accuracy and quality of identified Y sequences. Based on our results, we provide some recommendations for future sequencing efforts in insects to best support downstream Y assembly steps.

Keywords

Sex chromosomes, Insect sex chromosomes, Y chromosome, Genomics, Computational biology, Tephritidae, Sterile Insect Technique

Background

Studying the genome of non-haploid species with sex chromosomes can be complicated by a high incidence of repetitive regions and large areas of haploidy [1]. It is common for the haploid, male-limited Y to comprise mainly by repetitive and heterochromatic regions that are known to confuse DNA sequencing and assembly technologies [2, 3]. *Drosophila melanogaster*, the first eukaryotic organism to have its genome sequenced[4] and assembled, has had extensive efforts towards resolving the content and structure of its Y chromosome[5]. Nonetheless, parts of the *D. melanogaster* Y remains unresolved and, despite the development of new DNA sequencing technologies, a single contiguous Y chromosome assembly is absent. Therefore, it is not surprising that our understanding over sex chromosomes is limited to families of model species, such as the Drosophilidae, as the analysis of those elements in non-model organisms is a painstakingly slow process.

Beyond sequence composition, another challenge is the haploid nature of heteromorphic sex chromosomes [6, 7]. While in diploid autosomes the DNA is mostly found in two copies, the sex chromosomes can have sequence and therefore sequencing reads that originate for haploid loci. Most DNA sequence assemblers do not cater for this complexity as they generally assume that the genome is entirely of the same ploidy. Further complications arise from repetitive sequences that are shared between one sex chromosome and autosomes [8] which can lead to sequences being significantly mis-assembled. Typically, manual work from experts (“manual curation” or “biocuration”) is required to correctly assemble and annotate such regions. However, such resources are rarely available and usually genome assemblies are dependent on algorithm limitation that cause Y sequences to appear fragmented in many small contigs or erroneously assembled [9].

A crucial step towards the curation of sex chromosome sequences is their identification in an assembly. The Chromosome Quotient (CQ) algorithm is a key method for identifying regions of sex-specific ploidy [9]. It relies on aligning sex-specific DNA sequencing reads to a sequence (e.g. a genome assembly or its constituent contigs) and comparing the ratio of aligned read counts between the two sexes (typically female to male). In its most basic implementation, CQ does not account for complex genome sequences. In

69 species with imperfect genome assemblies or more complicated and repetitive sequence composition than
70 *Drosophila melanogaster*, read alignments might fail to resolve ploidy differences among sex chromosomes
71 due to shared regions and mis-assemblies. Further, the CQ method requires the generation and alignment
72 of high-coverage sex-specific DNA sequencing datasets, which are both a computational and economical
73 constrain.

74 Alternative methods have been described, such as the Y-Genome Scan (YGS;[10]) that circumvent
75 alignment by making use of "kmer-isation" of DNA sequencing datasets. A "kmer" is a sequence of length *k*
76 and is often between 15 and 31 bp for computational efficiency in 64-bit computers. Kmerisation is a
77 process by which all possible kmers are computed and counted for each sequence in the dataset. This
78 allows the comparison of sequence datasets without computationally expensive algorithms that use read
79 alignments. Therefore, an approach is to generate kmer databases from sex-specific readsets. However, the
80 YGS software only considers kmers from female readsets and therefore fails to characterise sequences that
81 are shared between male and female genomic datasets.

82 The utility of improving methods and procedures used to characterize the sequence of sex
83 chromosomes, and especially of the male limited Y, extends beyond basic science. The long-standing
84 alternative population suppression method of Sterile insect technique (SIT), coupled with the genetic sexing
85 of Y-bearing male insects, has been proved valuable in the fight against pest species [11]. Tephritids are
86 major invasive pests in global agriculture [12] over which SIT has effectively been applied and optimized,
87 rendering them as models for the improvement and expansion of the method[13]. The Tephritid Y is known
88 for housing the maleness factor and its largely unexplored sequence is expected to contain elements related
89 to male survival and reproduction [7, 14]; two crucial aspects for the efficient implementation of SIT
90 programs[15].

91 The CQ and YGS approaches have been widely used in genomic projects involving sex chromosomes
92 [16, 17]. Here, we assess these two popular Y-region finding algorithms and develop two complementary
93 approaches, R-CQ and KAMY, which extend on the work of CQ and YGS, respectively. Coupled with *D.*
94 *melanogaster*, our test dataset includes *Ceratitis capitata* and *Bactrocera oleae* and is oriented towards the

characterization of the unexplored Tephritid Y chromosome. The Mediterranean fruit fly *Ceratitis capitata* is a polyphagous species and one of the most destructive members of the family, while the olive fruit fly, *Bactrocera oleae,* is a monophagous Tephritid pest targeting solely olives[18]. For both of these species, there is active work on exploiting their heteromorphic Y for the development of sexing strains for SIT programs. As a result, significant research activity has led into the cytogenic characterisation of their sex chromosomes and discovery of the Y-linked maleness factor gene *MoY* [14, 19]. Here, we extend the methodological toolkit for Y chromosome analysis and suggest an approach to support manual curation towards identifying significant assembly errors, characterizing Y-chromosome regions, and understanding the nuances behind sex chromosome assembly.

Methods

Species datasets

In this work, we used the publicly available reference genome assemblies of *Ceratitis capitata*, *Bactrocera olea and Drosophila melanogaster* (GCA_905071925.1, GCA_001188975.4, GCF_000001215.4). For the Tephritids, scaffolds were split to contigs in instances where >100 N characters were found. For *D. melanogaster* the scaffolded genome was used directly. Contigs smaller than 3,000 bp were filtered out. DNA-Seq Illumina sex-specific readsets were downloaded from NCBI's SRA (*B. oleae*: SRR8788622, SRR8788979; *C. capitata*: ERR4026336, ERR4026337; *D. melanogaster:* ERR2163720, ERR2163721).

For the transcriptome of *B. oleae* the Nanopore RNA-seq readsets SRR8819878, SRR8819879, SRR8819880, SRR8819881, SRR8819884, SRR8819885, SRR8819886, SRR8819887, SRR8820069, SRR8820070, SRR8820073, SRR8820074, SRR8820075, SRR8820076 were downloaded from NCBI's SRA. The Nanopore RNA-seq reads were error corrected using Falcon [20]. An assembly was first performed using Trinity RNA-seq[21] (Illumina using the Nanopore for scaffolding) and then by using Falcon (Nanopore only). These two assemblies were then merged using Mira 3 [22].

Implementation of CQ

The Chromosome Quotient (CQ) method uses separate female and male DNA-read alignments to a reference sequence and utilizes the female-to-male ratio of read counts to infer putative Y contigs. Here, the algorithm was implemented using the same criteria suggested by Hall et.al ([9]) for prior masking of

genomic repeats and putative Y selection. A hard-masked genome version was generated for *B. oleae* and *C. capitata* using the repeat masking module of JAMg (https://github.com/genomecuration/JAMg). The repeat masked *D. melanogaster* genome was downloaded from FlyBase [23]. Sequences with a CQ value of less than 0.3 were marked as putatively Y-chromosome derived.

Implementation of R-CQ

Reverse-CQ (R-CQ) was implemented in Perl through reverting the CQ calculation as male/female and extending the CQ approach with additional steps to account for repetitive regions and differences in DNA sequencing library depth. The algorithm is described further in the Results section and is freely available from https://zenodo.org/doi/10.5281/zenodo.10209595. It uses as input the reference genome sequence and the coverage data from male- and female-specific readsets. Whole genome DNA-seq alignments were performed using GSNAP (version 2023-02-17; [24]) with the uniquely-mapping read pairs (concordant-unique) post-processed to BigWig files. R-CQ implements a user-defined window to calculate female to male ratio statistics; a size of 50 bp was found to perform best. R-CQ also allows for an autosomal, high quality contig to be selected for normalisation; here the contigs of CAJHJT010000034.1.84, LGAM02035380.1, and NT_033779.5 were selected for *C. capitata, B. oleae* and *D. melanogaster* respectively. We empirically set ln(median(R-CQ)) > 1 as the threshold for marking putative Y contigs.

Implementation of YGS

The Y Genome Scan (YGS) method counts kmers from female DNA-sequencing datasets and, after accounting for repetitive sequences, it estimates the proportion of each contig that is not covered by female kmers. Male genomic data can be added to account for sequencing errors and rare polymorphisms by excluding female kmers that are absent in males; in the implementation of Carvalho & Clark [10], however, the lack of coverage by a kmer is assumed to be due to male sequence. Here, both male and female datasets were used as input for the YGS algorithm, performing also the optional male-validation step. In the original implementation, if a sequence is at least 95 % composed of kmers not found in the female dataset, it is marked as a putative Y sequence. Our observations (Figure S. 5, Figure S. 9, Figure S. 13) found that this cutoff is too aggressive and underperform, hence we set the threshold to 60 %. As suggested

by the original publication, the kmer size was set to 15 and kmers occurring 5 of fewer times in the female readset or only 1 in the male readset were filtered out.

Implementation of KAMY

The "Kmer-counting Algorithm Meryl identifies the Y" (KAMY) estimates the proportion of a sequence that is supported by male, female, and shared (common) kmers. Further details are available from the Results section and it is freely available from https://zenodo.org/doi/10.5281/zenodo.10209589. Prior to processing, male and female DNA-seq datasets were quality trimmed and had putative contamination removed via the justpreprocessmyreads of JAMg (https://github.com/genomecuration/JAMg). Kmer counting used Meryl r2013 (https://kmer.sourceforge.net) with a k=21. Kmers occurring only once (singletons) were filtered out. KAMY marks putative Y contigs as those with a DNA sequence that is covered by less than 90 % by common kmers and more than 3 % by unique male-specific kmers.

Manual curation and benchmarking of methods

Manual curation is the process by which an expert reviews computationally generated information from multiple sources and makes decisions based on their expertise. Here, the manual curation of Y predictions was performed through contrasting the results of all the methods with the observed read coverage depth on a genome browser, JBrowse2 [25]. Genome browser tracks of DNASeq read alignments were created for read pairs that aligned once (concordant-unique) and for read pairs that aligned multiple times (concordant-mult) (GSNAP; [24]) as well as identified repeats, putative genes, transcriptome and long-molecule sequencing reads. Exploration was further conducted through graphs created in R [26].

Each putative contig was inspected for the contiguity, depth and consistency of coverage of male and female DNA-seq reads and assigned to one of the following categories:

- **Y/C**: contiguous assembly of Y region with half the depth coverage of autosomal regions
- **Y/R**: contiguous assembly of Y with regions containing uniquely aligned male reads with more than twice the expected Y depth (putative falsely collapsed Y repeats)
- **Y/M**: male-specific region that has been clearly mis-assembled
- **R:** undefined repeats

- **M:** undefined mis-assembly
- **A**: autosomal region consistently covered by male and female reads at the expected depth coverage
- **C**: region with very low DNA-seq coverage validated for non-fly origin through BLASTn

The Y/C, Y/R, and Y/M categories depict the assembly state of the contig and are dependent on the genome sequencing and assembly methods. When a contig is heterochromatic or highly repetitive – common for heteromorphic Y chromosomes – it is prone to mis-assemblies and our approach allows for characterising the ability of each method to work with assembly inconsistencies. The Y/R category focuses on repetitive regions of the Y that have been sequenced independently but assembled as a single region. The Y/M category is for contigs with coverage inconsistencies suggesting mis-assemblies of Y-Y or Y-autosomal regions. Such contigs can be characterised by: 1) sharp peaks and drops of coverage from half the library depth (as expected for a heteromorphic Y chromosome) to the expected sequencing depth of autosomal regions; and 2) regions not supported by the DNA sequencing data.

Therefore, one manually curated core Y dataset was derived for each organism. These core datasets allowed us to infer the Y-sensitivity and Y-specificity of each method for each organism. Y-sensitivity was estimated as the percentage of contigs from the core Y dataset of the organism that were correctly predicted by each respective method ($Y_{sensitivity}= Y_{method}/Y_{total}$ %). Y-specificity represented the proportion of correct Y predictions over the total number of predictions for each method ($Y_{specificity}=Y_{method}/Total_{method}$).

PCR validation

A set of curated Y contigs from *B. oleae* and *C. capitata* were selected for PCR validation based on their size and a sequence complexity suitable for primer design. Primer pairs were designed for 46 and 12 *B. oleae* and *C. capitata* contigs respectively using the primer3 plugin on Geneious 2022. For most contigs a single primer pair was designed but for those with substantial gaps (which could denote a cross-chromosomal mis-assembly) one primer pair was designed for every ungapped region. Primer design sought to generate PCR products of 400 to 1,000 bp and was assessed for its uniqueness using the Primer BLAST. Template DNA was derived from males and females individually using pools of 4 insects so that multiple haplotypes can be

captured. Virgin *B. oleae* insects were collected from the "Demokritos" population and virgin *C. capitata insects* were collected from the "EGII laboratory" populations reared at the University of Thessaly. Genomic DNA was extracted using the protocol described in Calderón-Cortés et al. 2010 and quantified using a Quawell Q3000 spectrophotometer so that 30 ng/µl were used in each reaction. PCR reactions were conducted using a KAPA Taq (Merck) according to manufacturer's protocol. PCR reactions were carried out at a Bio-Rad T100 thermal cycler and the PCR products were visualized on 1 % agarose gel. When a band was present in only males it deemed that the primer pair was from a Y-linked contig.

Results

Table 1 Feature comparison of CQ, R-CQ, YGS and KAMY algorithms.

	CQ	*R-CQ*	*YGS*	*KAMY*
Approach	Ratio	Ratio	Kmers	kmers
Sex chromosomes identified	Y & X	Y & X	Y	Y
Required datasets	Female/Male sequencing	Female/Male sequencing	Female sequencing (male optional)	Female/Male sequencing
Pre-processing required	Read sampling, Genome repeat masking	Read alignment	No	No
Read depth normalization	User dependent	Yes	No	No
Search method	Query size	Window size	Kmer size	Kmer size
Shared regions	Excluded if repeat masked	Excluded	Considered	Considered

Creation of the R-CQ algorithm

Reverse-CQ algorithm detects putative Y sequences by comparing the depth-coverage differences between male and female datasets. The sex-specific datasets are independently mapped onto the reference genome and only uniquely aligned read pair sequences are considered, overcoming the need for any repeat masking. Further, the algorithm searches the genome sequences via non-overlapping windows of a

predetermined size (defaulting to 100 bp) so that the mean and median depth-coverage ratios are derived for each window. Third, to minimise the bias from sequencing depth variation between the sex-specific sequencing readsets, R-CQ uses the sequencing depth off a user-defined autosomal contig to derive a normalisation factor and uses it to normalise the female depth-coverage. Fourth, the normalisation factor is used to derive a repeat cutoff that is employed to account for mis-assembled regions where multi copy repeats have been collapsed into a single region. The repeat cutoff can also be set by the user. Crucially, CQ values are derived for each window as the ratio of male to normalised female mean depth (the reverse of the original CQ implementation) and a contig-wide CQ is then estimated by taking the median of all the mean CQ values across all windows. As our research is focused on heteromorphic Y chromosomes, we expect that Y genomic data are mainly uniquely present in the male genome and therefore mark contigs with an lnMedian(R-CQ) value of 1+ as putative Y sequences. Autosomal contigs are expected to cluster near lnMedian(R-CQ) = 0 while X linked sequences are expected to cluster near lnMedian(R-CQ) = -0.7.

Creation of the KAMY method

The KAMY method identifies Y-linked regions in an assembly by assessing their "breadth of coverage" (i.e. the proportion of a sequence) for sex-specific kmers. No alignment or repeatmasking is required. The Meryl algorithm is employed to generate 3 kmer databases from sex-specific genomic data: male, female and common. Subsequently, the contigs from a genome of interest are assessed based on how much of their sequence is covered by kmers from each database. The number of times a kmer is found – proportional to depth of coverage - is not used even though that could be a future improvement. The user can then exclude contigs covered mainly by common kmers (90 % in this work) which are likely to be autosomal or regions common in the X and Y chromosomes. Contigs exhibiting male-unique coverage (3 % in this work) can then be selected for downstream analysis. The cutoffs need to be determined for each project due to the variation in sequencing depth, genome composition, and assembly quality. In XX/XY organisms, the X chromosome is present in both male and female sequencing data and therefore KAMY cannot be used to identify such X-linked contigs.

Identification of Y sequences

Ceratitis capitata

For the reference genome of *C. capitata,* the four assessed Y-detection methods suggested a total of 251 putative Y contigs, that were subjected to manual curation (R-CQ: 104, KAMY: 248, YGS: 106, CQ: 60; Table S. 2). A total of 111 contigs were found to be probable Y-linked and their assembly status was curated as follows: 60 were well-assembled as Y (Y/C; Figure 4g); 37 contained falsely collapsed repeats (Y/R); 14 contained mis-assemblies (Y/M); and 140 contigs were false positives (A: 132, M: 7, R: 1; Table S. 2). Fifty-seven of some of the largest (more than half of them being larger than 150 kb) Y contigs were commonly identified by all four methods (Figure 4a/b). The second largest intersection contained 45 Y contigs, that were missed by CQ but consistently found by the other three methods. These groups contained smaller Y regions with higher repeat content, which might be the reason for the CQ method failing to mark them as putative Y, since the total of those sequences is expected to have been repeat-masked. Overall, KAMY, YGS and R-CQ methods were effective in handling the *C. capitata* Y and despite its repetitive character (~94 %

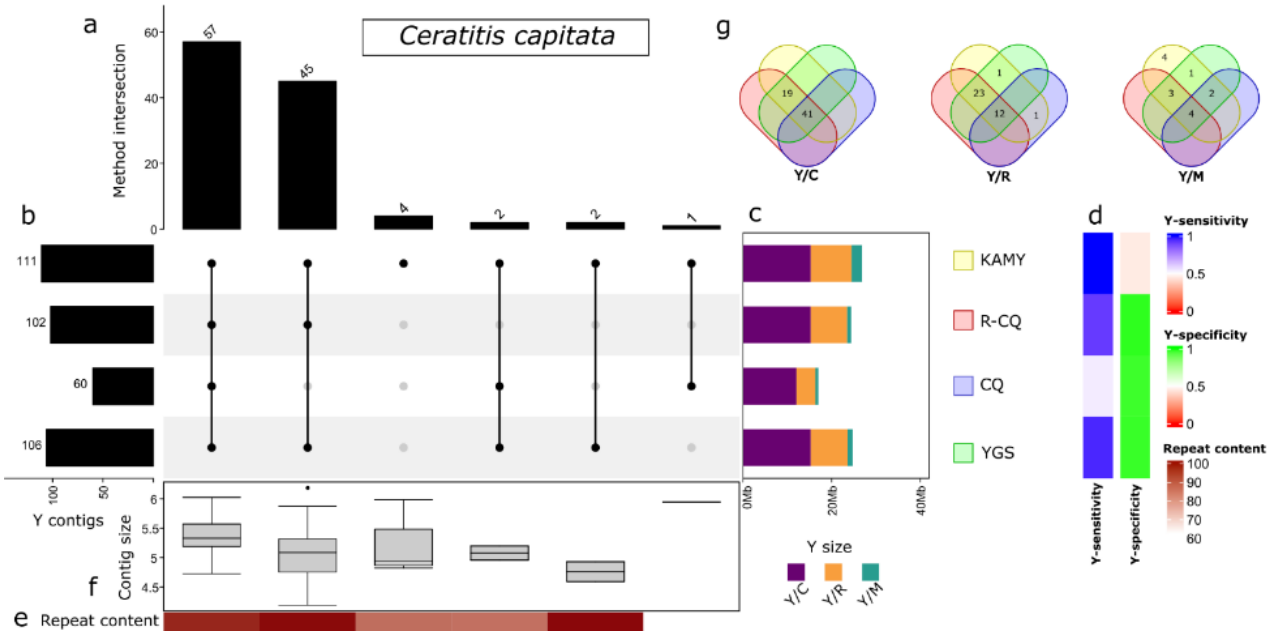

Figure 1 Benchmarking results on the Y-recall efficiency of the four assessed methods over the genome of Ceratitis capitata. Upset plot is built based on the intersections of method predictions. a) Total counts of each intersection indicated in the base of each bar plot. b) Total number of correct Y predictions of each method. c) Total size of Y predictions achieved by every method, split in the three Y categories (Y/C, Y/R, Y/M). d) Heat maps represent the Y-sensitivity (blue) and Y-specificity (green) values for every method. e) Heat map represents the mean repeat content (brown) of the contigs included in each of the intersection categories indicated above. f) Boxplots contain information regarding the log10(Size) of the contigs that fall in each of the intersection categories indicated above. g) Venn diagrams visualize the number of Y contigs commonly/uniquely identified by the four methods as per assembly status.

253 repetitive content based on RepeatMasker) the *C. capitata* Y had been surprisingly well-assembled.
254 Relatively few contigs were found to contain assembly errors (14 Y/M), with KAMY uniquely identifying 4 of
255 those.

256 The KAMY method had a 100 % Y-sensitivity rate (explained in Materials & Methods), having identified all
257 111 of the curated Y contigs (Figure 1b/d, Figure S. 4, Table S. 2) and 44.76 % Y-specificity with 137 of 248
258 predictions curated as false positives. YGS had the second highest Y-sensitivity (95.5 %), reflecting 106
259 correct Y contigs accompanied by a 97.25 % Y-specificity rate (3 false positives; Figure 4d, Figure S. 5, Table
260 S. 2, Table S. 3). The R-CQ algorithm achieved 91.89 % Y-sensitivity (102 Y contigs) and 98.08 % Y-specificity
261 (2 false positives; Figure 4d, Figure S. 3, Table S. 2, Table S. 3). The CQ method was the least efficient for *C.*
262 *capitata* with a 54.05 % Y-sensitivity (60 Y contigs) and 96.77 % Y-specificity (2 false positives; Figure 4d,
263 Figure S. 2).

264 Collectively, the 111 Y contigs identified across all 4 methods summed up to a total ~26.8 Mb of
265 curated Y sequence (Table S. 4). Out of that, ~15.3 Mb were considered as well-assembled Y regions, ~9.2
266 Mb included falsely collapsed repeats (Y/R) and ~2.3 Mb included mis-assemblies of Y regions (Y/M). The
267 higher sensitivity of KAMY in identifying contigs of the Y with assembly peculiarities was reflected in a ~2.1
268 Mb difference in the size of identified Y/R and Y/M contigs, compared to the second best (Figure 4c, Table S.
269 4). Crucially, KAMY found all ~9.2 and ~2.3 Mb of Y/R and Y/M regions while the respective numbers for YGS
270 were ~8.4 and ~1.1 Mb. However, it should be noted that due to the nature of the mis-assemblies the sizes
271 of those regions may not reflect biological reality (an underestimation for the Y/R category and an
272 overestimation in the Y/M).

273 *Bactrocera oleae*

For *B. oleae,* the four Y detection methods combined labelled 792 contigs as putatively Y-chromosome derived (R-CQ: 246, KAMY: 148, YGS: 579, CQ: 48) (Table S. 6). The manual curation characterised 83 contigs as Y, distributed in the three Y-categories as follows: 40 well-assembled Y (Y/C), 26 contained falsely collapsed repeats (Y/R) and 17 contained mis-assemblies. False positives accounted for the majority of the 709 contigs: 344 were contaminants of non-fly DNA (C); 268 were autosomal regions (A); 88 were undefined mis-assemblies (M); and 9 were assemblies of repeats that could not be classified as autosomal or shared with the sex chromosomes. Subsequently, this dataset of 83 curated Y contigs was used to assess each method. The two novel methods, R-CQ and KAMY, presented the highest Y-sensitivity rate (77.11 %) with both identifying 64 curated Y contigs, though not all were the same (Figure 5b, Table S. 6, Table S. 7, Figure S. 7, Figure S. 8). No contigs were commonly identified by all four methods, rather the largest intersection was 29 common Y predictions found in KAMY, R-CQ and CQ (Figure 5a).

Interestingly, the second largest intersection was derived from our two new methods commonly identifying 21 curated Y contigs. More than half of the Y sequence found within the two major intersections appeared

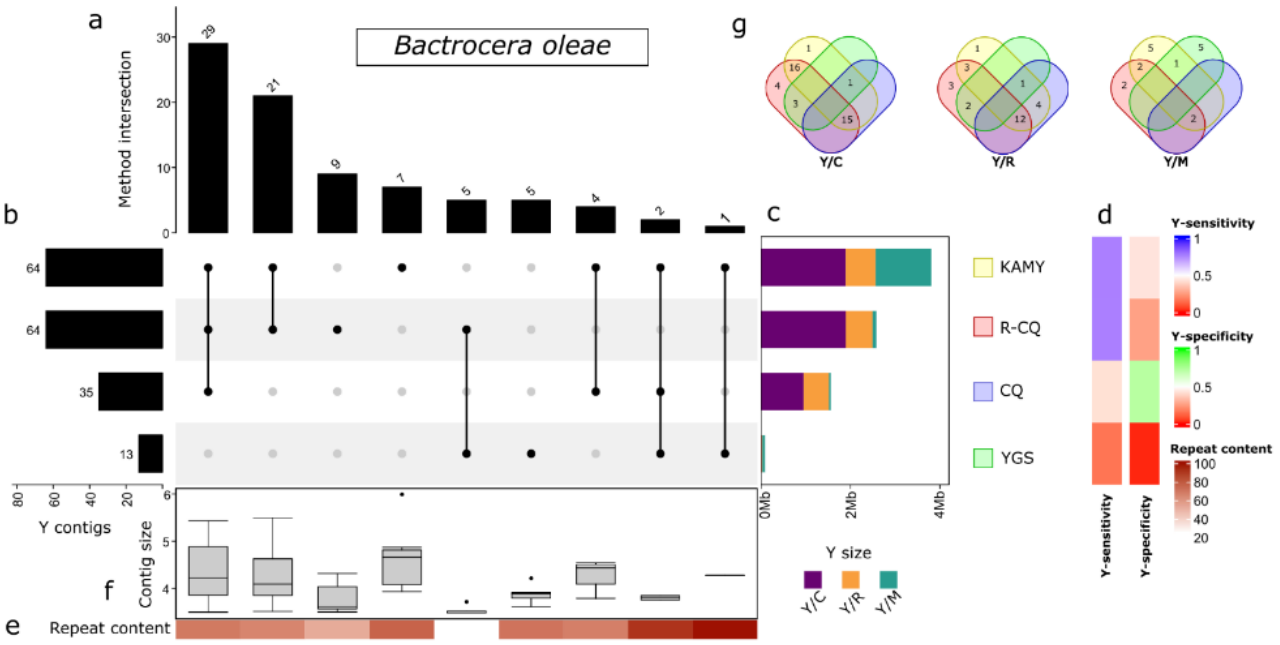

Figure 2 Benchmarking results on the Y-recall efficiency of the four assessed methods over the genome of Bactrocera oleae. Upset plot is built based on the intersections of method predictions. a) Total counts of each intersection indicated in the base of each bar plot. b) Total number of correct Y predictions of each method. c) Total size of Y predictions achieved by every method, split in the three Y categories (Y/C, Y/R, Y/M). d) Heat maps represent the Y-sensitivity (blue) and Y-specificity (green) values for every method. e) Heat map represents the mean repeat content (brown) of the contigs included in each of the intersection categories indicated above. f) Boxplots contain information regarding the log10(Size) of the contigs that fall in each of the intersection categories indicated above. g) Venn diagrams visualize the number of Y contigs commonly/uniquely identified by the four methods as per assembly status.

287 to be larger than ~18 kb in size (Figure 5f) and with a ~63 % repeat content, significantly lower than the
288 both *C. capitata*'s 94 % repeat content and ~ 100kb in size. Additionally, among the R-CQ unique predictions
289 were 9 Y contigs of relatively low size. This suggests an expected advantage of the window-based R-CQ
290 approach when assessing fragmented regions. On the other hand, KAMY was superior in uniquely
291 identifying Y regions containing mis-assemblies. Those contigs were characterised by their large size and
292 higher repeat content compared to the first two intersections (Figure 2f), suggesting the robustness of this
293 method over challenging Y regions.

294 Indeed, KAMY exhibited an improved performance in terms of Y-specificity at 43.24 %, however
295 including 84 false positives in its predictions, compared to R-CQ that was second best, achieving a Y-
296 specificity of 26.02 % with 182 false positives (Figure 5, Table S. 6, Table S. 7). The CQ method ranked third,
297 demonstrating a Y-sensitivity of 42.17 % (35 Y contigs) and a commendable Y-specificity of 72.92 % marked
298 by 13 false positives (Figure 5, Figure S. 6, Table S. 6, Table S. 7). In contrast, YGS exhibited a significant
299 underperformance, correctly predicting 13 Y contigs, representing a Y-sensitivity of 15.66 % (Figure 5,
300 Figure S. 9, Table S. 6, Table S. 7). The majority of YGS predictions comprised false positives, including 344
301 contamination (C), 142 autosomal (A), 77 undefined mis-assemblies (M) and 3 undefined repetitive contigs
302 (R), yielding a Y-specificity of 2.25 %. Despite its low Y-recall efficiency, YGS uniquely contributed 5 Y/M
303 contigs in the curation dataset that were missed by the other three.

304 The cumulative size of the curated olive fruit fly Y chromosome, resulting from the combination of
305 all four assessed methods was 3.9 Mb, with the anticipated size for the dot Y chromosome estimated at ~4
306 Mb[28] (Table S. 8). The improved performance of KAMY is once again highlighted by the total size of
307 identified Y, that summed up at ~3.8 Mb. This is notably ~1.2 Mb more than the second-ranking method, R-
308 CQ (Figure 5, Table S. 8). This difference is primarily attributed to the detection of large mis-assembled
309 contigs by KAMY, which were overlooked by R-CQ. The cumulative size of Y predictions from the CQ method
310 summed up at ~1.6 Mb while those of YGS amounted to ~0.9 Mb. It is crucial to note again that the
311 cumulative size of Y/M and Y/R might deviate from biological reality due to the nature of mis-assemblies in
312 those regions.

313 *Drosophila melanogaster*

314 For the Diptera model species *D. melanogaster,* the sex chromosomes have been characterized in
315 the reference assembly through extensive curation and manual annotation [29, 30]. The scaffolded
316 reference genome of *D. melanogaster* was used as input for the four algorithms. We chose not to split the
317 chromosome-level scaffolds since this introduced many small contigs in the analysis that would be either
318 filtered out (< 3 kb) or their assembly status would be hard to curate. The methods yielded a total of 202
319 putative Y regions (R-CQ: 110, KAMY: 151, YGS: 190, CQ: 100) (Table S. 11). The ~3.7 Mb annotated Y
320 scaffold was selected, along with 139 additional unmapped contigs that were curated as Y and 61 false
321 positives (A: 29, M: 17, R: 15). In contrast to observations in Tephritid genomes, the YGS method
322 demonstrated better performance in identifying regions of the *D. melanogaster* Y relative to the other three
323 methods. It achieved a 94.33 % Y-sensitivity rate having 133 correct Y predictions (Y/C: 52, Y/R: 54, Y/M: 27).
324 However, the Y-specificity of the method was compromised reaching 70 % (57 false positives) (Figure S. 1,
325 Table S. 11, Table S. 12, Figure S. 13). KAMY ranked as second-best in Y-specificity (91.49 %) accurately
326 predicting 129 Y regions (Y/C: 52, Y/R: 53, Y/M: 24), It also presented the highest Y-specificity of the four
327 (85.43 %) with 22 false positive predictions (Figure S. 1, Figure S. 12, Table S. 11, Table S. 12). R-CQ identified
328 92 curated Y regions (65.25 % Y-sensitivity) with a Y-specificity of 83.64% (18 false positives) (Figure S. 1,
329 Figure S. 11, Table S. 11, Table S. 12). CQ ranked as the least efficient, achieving a 58.87 % Y-sensitivity (83
330 correct predictions) with 83 % Y-specificity (17 false positives) (Figure S. 1, Figure S. 10, Table S. 11, Table S.
331 12). The underperformance of CQ on D. melanogaster's Y chromosome was apparent from the low Y-recall
332 rate and the absence of the large Y-annotated scaffold in the method's predictions.

333 All four methods commonly identified 52 curated Y regions (Figure S. 1). The second largest intersection
334 contained 29 Y regions missed by CQ but found by the other three, while the third largest comprised of 24 Y
335 regions missed only by R-CQ. As observed previously, the kmer-based methods exhibited better
336 performance with parts of the Y containing assembly peculiarities (Figure S. 1g). KAMY and YGS collectively
337 identified 16 such regions falling in the Y/R and Y/M categories. The repeat content of *D. melanogaster* Y
338 regions was approximately ~36 %, notably lower compared to the evaluated Tephritidae Y chromosomes. In
339 contrast to the assemblies of the other two species, the non-scaffolded Y of *D. melanogaster* displayed a

high degree of fragmentation, with the majority of regions being smaller than 12 kb. However, a set of 16 contigs uniquely contributed to the dataset by kmer methods exibited larger sizes, indicating a bias in the R-CQ window-based approach towards smaller Y regions.

A total ~5.5 Mb of Y sequence was curated from the predictions, with ~3.7 Mb constituting the scaffold annotated as Y in the chromosome-level assembly (Table S. 13). The remaining ~1.8 Mb of curated Y sequence contained annotated Y regions that were not assembled within the large Y scaffold (~517 Kb), centromeric repeats (~294 Kb), repeats shared between X;Y (~52 Kb) and ~934 Kb of sequence not previously annotated, considered as newly identified Y (Table S. 16). The differences in Y-sensitivity are consistently reflected in the total size of Y identified by each method. YGS exhibited the highest Y recall, encompassing the largest portion of assembled *D. melanogaster* Y (~5.4 Mb) followed by KAMY (~5.3 Mb), R-CQ (~4.9 Mb) and lastly CQ (~780 Kb) (Table S. 13). It is noteworthy that the existing Y chromosome annotations of contigs generally aligned with our prediction, with only a few exceptions where the contigs were curated as R or M due to the difficulty in confidently inferring the chromosomal origin (Table S. 14).

PCR validation

The dataset of curated Y contigs from *C. capitata* and *B. oleae* was subjected to PCR validation to assess the accuracy of our curations. In *C. capitata,* a 12 manually curated Y contigs were selected for PCR validation, of which all 12 resulted in male specific amplification from genomic DNA. Due to the highly fragmented and mis-assembled nature of the *B. oleae* Y chromosome, 46 curated Y contigs were chosen for PCR validation in this species, of which 36 were confirmed as Y-chromosome derived Cross validation of the Y curations on *D. melanogaster* utilized reference sequence annotations derived from cytological or genetical mapping [23].

The PCR validation results on *C. capitata* confirmed our observations regarding the high quality of assembly and supported the accuracy of our manual curation. All 12 assessed contigs were validated as Y-linked through PCR (Table S. 5). Collectively, the 12 *C. capitata* Y contigs summed up to ~7.4 Mb of PCR-validated *C. capitata* Y sequence. Conversely, the PCR validation of the olive fruit fly's Y curations revealed a mis-assembled and inconsistent assembly state for the Y. Out of the 46 contigs subjected to PCR, 36 were

validated, while the other 10 were considered false positives (Table S. 9, Table S. 10). Among them, the majority (7/10) were characterised as Y/C by the curation, while 2 were Y/M and 1 Y/R. The recall of PCR-validated Y contigs was higher for the KAMY method, which included all 36 validated Y contigs in its predictions, while R-CQ correctly predicted slightly fewer validated Y contigs (34/36). Additionally, KAMY demonstrated a slightly better performance in avoiding PCR-validated false positives, characterizing 9/10 false positives as Y, compared to R-CQ which included all 10 false positives in its Y predictions. Overall, the total size of PCR-validated *B. oleae* Y chromosome sequence summed up to ~2.4 Mb.

The existing annotations of *D. melanogaster* Y contigs generally aligned with our curations. All 37 Y and 4/8 X;Y-annotated regions that exceeded the size filter of 3 kb were included in our putative Y dataset (Table S. 14, Table S. 15). Among the 37 Y-annotations, 6 were curated as Y/C, 18 as Y/R and 7 as Y/M. Additionally, 4 annotated Y regions were curated as undefined repeats (R) and 1 as undefined mis-assembly (M) as the chromosomal origin could not be confidently inferred for these regions due to high inconsistency in depth and width of read coverage. Kmer-based methods demonstrated better efficiency in predicting the annotated Y regions of *D. melanogaster*. Specifically, KAMY recalled 35 Y and 3 X;Y annotations, while YGS recalled 35 Y and 4 X;Y annotations. Overall, ~4.4 Mb of *D. melanogaster* Y sequences are annotated in the reference assembly.

Assessing KAMY with transcriptomic datasets

We extended KAMY's application on transcripts as a potential alternative – and lower cost and effort – way to derive Y-linked sequences. First, we generated a reference-free assembly of the RNA-Seq Nanopore datasets obtained from mixed-sex *B. oleae* embryos. Subsequently, we evaluated the Y-linkage of transcripts using male and female DNA-seq KAMY databases. The maleness factor gene *MoY*, previously identified on the Y chromosome of *B.oleae*, with a known peak expression during early embryonic stages

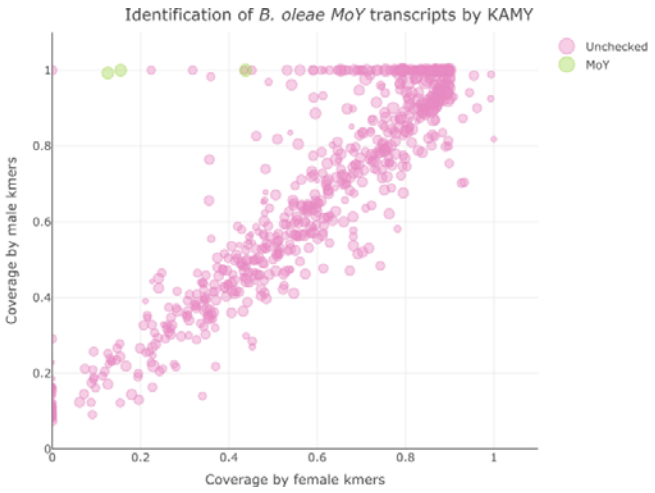

Figure 3 Scatterplot of male VS female kmer coverage over putative Y B. oleae transcripts as identified by KAMY. Transcript assemblies representing sequences of MoY are plotted in green.

392 (Meccariello et al. 2019), served as a control for our assessment. KAMY successfully identified a set of
393 putative-Y embryonic transcripts. Notably, among these were three sequences corresponding to the *B.*
394 *oleae MoY* gene (Figure 3). Such results underline the capability of KAMY to identify important Y-linked
395 elements, such as the maleness factor, even in the absence of a reference sequence for the Y chromosome.

396 Discussion

397 In this work, we evaluate and provide methods for identifying, curating, and improving the genome
398 sequence of heteromorphic sex chromosomes. Identifying Y-linked regions is far more problematic than the
399 literature on model species suggests[9, 10]. For example, after more than forty years of effort to identify the
400 *C. capitata* Y-linked genes [31] only one gene has so far been successfully characterised (Meccariello et al.
401 2019) within a chromosome considered by the research community to be around 40 Mb [19].

402 In spite of large synteny within the Diptera class [32], sex chromosomes vary widely even within the same
403 genus [19], much more within the same taxonomic class. Despite *C. capitata* and *B. oleae* sharing the same
404 number of chromosomes with the model species *D. melanogaster*, the Y chromosomes among those
405 organisms vary in morphology. The heteromorphic Y of *B. oleae* appears significantly smaller than the other
406 autosomes and is estimated to be around 4 Mb, while those of *C. capitata* and *D. melanogaster* are
407 significantly larger and are postulated to be ~40 Mb [19, 33, 34]. Here, we assessed the discrepancies
408 among different Y-detection methods in the correct identification of Y-linked sequences from those
409 morphologically different Y chromosomes.

410 The two considered Tephritid species have significant whole genome sequence projects employing
411 different sequencing approaches. The olive fly assembly included a 10X Genomics linked-read male library
412 with ~300× sequencing depth and a female Illumina shotgun library with ~100× coverage, Oxford nanopore
413 and PacBio [17], whilst the *C. capitata* was based on PacBio, Phase Genomics HiC and accompanied by two
414 Illumina libraries of ~40× depth [35]. We originally found that the different nature and assembly strategies
415 limited our ability to correctly identify the Y. Therefore, we opted to assess the performance of established
416 methods and develop extensions that take into account the variability in biological sequence and genomic
417 data.

418 We observed that our R-CQ implementation performs better on Tephritidae Ys over the established CQ
419 algorithm. The repeat cutoff handled with success the highly repetitive *C. capitata* Y while the non-
420 overlapping windows were efficient over the fragmented, smaller Y contigs of the *B. oleae* Y. Regarding the
421 performance across different genomic datasets, the specificity of the method was compromised for *B.*
422 *oleae*. There was a significant number of false positives, primarily associated with sequences exhibiting
423 consistently low female coverage throughout the entire length of the contig. Such differences could not be
424 counteracted by the depth normalizing step, therefore were falsely identified as Y-linked by the R-CQ and
425 were removed from the dataset through the process of manual curation and PCR validation. The library
426 preparation approaches followed in this organism (10× Genomics single molecule vs Illumina shotgun) may
427 have caused the observed variation in sequencing depth. In contrast, the results of R-CQ in *C. capitata*, for
428 which male and female libraries were prepared using the same methodology, suggest that the depth
429 normalizing step effectively adjusted for small differences between libraries. This adjustment led to more
430 accurate Y, autosomal and X predictions compared to the CQ algorithm.

431 On the side of kmer-based methods, KAMY exhibited improved Y-sensitivity compared to YGS. However,
432 the interaction between the highly repetitive *C. capitata* Y and the filtering parameters employed by KAMY
433 for the common proportion of each contig resulted in the inclusion of numerous false positive predictions
434 for this organism, reducing the respective Y-specificity of the method. Importantly, though, the Y-sensitivity
435 was not compromised, and KAMY yielded the most correct Y predictions. Conversely, YGS proved to be the
436 most effective method on *D. melanogaster* Y, but its Y-identification ability was diminished on the olive fly.
437 The variability in YGS performance across the assessed genomes can be attributed to the differences in
438 biological sequence characteristics among different Y chromosomes. YGS identifies Y regions through
439 assessing the percentage of a sequence that is not present in a female-only kmer database. During the
440 initial development of this algorithm, the authors optimized the method using a *Drosophila* species with Y
441 regions found to have 95-100 % of their sequence unmatched by female kmers[10]. This observation aligns
442 with the results presented here for *D. melanogaster*. However, the percentage of Tephritidae Y sequence
443 unmatched by female kmers ranged at lower levels, indicating the presence of shared kmer regions
444 between Y and autosomes- a putative characteristic of that Y lineage. Notably, previous studies imply an

independent evolution between *D. melanogaster* and *B. oleae* Y chromosomes [36]. The case of the olive fly's Y exemplifies poor method performance, as the majority of contigs from this dot-like Y had more than 40 % of their sequence present in female kmer database. Consequently, Y contigs could not be independently clustered by the YGS algorithm, and the majority of Y predictions in this organism consisted of false positives, primarily non-fly contaminants (Figure S. 9). This functionality of the method was initially pinpointed by its authors[10] and can contribute to improving the overall quality of assemblies. However, concerning Y-identification, the sensitivity over exogenous DNA sequences is a limiting factor on the performance and general applicability. Overall, kmer-based methods were found to perform better with regions of the Y containing mis-assemblies or falsely collapsed repeats, an important parameter for improving our knowledge over quirky Y chromosomes. Notably though, if it was for a wide sex chromosome analysis, kmer-based methods would have been unable to identify the X since reads originating from that chromosome would be present in both male and female datasets.

Assessing Y-transcripts with KAMY

Despite the emerging efficiency of KAMY with genomic data, several factors, including cytological or technical limitation, may hinder the assembly of contiguous Y sequences. Consequently, this can impede the annotation of functional regions on Y chromosomes, limiting the comprehensive characterization of Y-linked genes. Furthermore, the study of developmentally important Y elements, such as maleness factors, which exert their functional role during early embryonic stages, faces additional hurdles. The generation of sex-specific RNA-seq data during these stages is a technical constrain. Using KAMY with sex specific DNA-seq kmer databases over mixed-sex transcriptomic data from *B. oleae* embryos identified *MoY*, among a set of putative-Y embryonic genes. This highlights how KAMY can overcome assembly software limitations, particularly in the context of Y-linked genes. By doing so, KAMY provides flexibility in the datasets and resources required for Y-related studies.

Curation is required

Our benchmarking suggests that there is no magic bullet when it comes to Y-identification methods, different methods were found to perform with different efficiencies among different species, while the

inclusion of false positives was a common instance. The employment of more than one Y-detection method and the subsequent manual curation of the collective outputs resulted in the generation of high-trust core Y datasets and removed many false positive predictions that could otherwise misguide downstream analysis. Such curation can be performed by non-specialized users on a regular desktop by using files already generated during the method implementation step. Considering the weaknesses of Y detection methods, one should always review the algorithmic outputs towards a more accurate assessment of the content and size of Y chromosomes. The PCR results from *C. capitata* and the existing Y-annotations of *D. melanogaster* prove the validity of our curation approach. However, the experimental validation of *B. oleae* Y contigs indicated the problematic nature of this assembly and suggests that PCR should always be considered for a conclusive validation of Y regions in such cases.

Conclusions

Overall, two novel Y-detection methods, R-CQ and KAMY, were developed and compared to the established methods. Two Tephritidae Y chromosomes were used as models for the improvement of the methods, while their efficiency and generic applicability was cross validated with the well annotated Y of *Drosophila melanogaster*. We concluded to the superiority of kmer-based methods in the identification of Y regions in an assembly, especially of regions containing assembly peculiarities, with KAMY appearing as the most sensitive in Tephritidae Y chromosomes, presenting also the ability of identifying Y-derived transcripts from mixed-sex transcriptomic datasets without the use of a reference genome. However, we suggest that more than one method is employed in a project since we observed that Y-identification methods perform differently depending on the sex chromosome characteristics and assembly quality. Additionally, we suggest that the same sequencing depth and library preparation approach is used for both males and females, as the opposite might generate depth inconsistencies between the two sexes that can be falsely identified as Y-derived by the algorithms. False positives are expected in every method and therefore we highly recommend the manual curation of algorithmic outputs. In this way, a significant number of false Y predictions can be removed, although PCR verification can be the only conclusive method.

Declarations

Data availability

All non-published algorithms used in this study are available at https://zenodo.org/doi/10.5281/zenodo.10209595 & https://zenodo.org/doi/10.5281/zenodo.10209589 . The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Funding

Funding for this research was provided by the European Union's Horizon Europe Research and Innovation Programme REACT (Grant agreement 101059523).

Competing interests

The authors declare no competing interests.

Authors' contributions

RD and AP performed the data analysis and developed the methods. RD and FK performed the wet lab validation. RD performed the desktop validation and curation. KTT co-created the first iteration of the KAMY algorithm. KDM led the biological interpretation. PAP and KDM contributed to the draft of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Carey SB, Lovell JT, Jenkins J, Leebens-Mack J, Schmutz J, Wilson MA, et al. Representing sex chromosomes in genome assemblies. Cell Genomics. 2022;2.

2. Furman BLS, Metzger DCH, Darolti I, Wright AE, Sandkam BA, Almeida P, et al. Sex Chromosome Evolution: So Many Exceptions to the Rules. Genome Biol Evol. 2020;12:750–63.

3. Bachtrog D. Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet. 2013;14:113–24.

4. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The Genome Sequence of *Drosophila melanogaster*. Science (1979). 2000;287:2185–95.

5. Chang CH, Larracuente AM. Heterochromatin-enriched assemblies reveal the sequence and organization of the drosophila melanogaster Y Chromosome. Genetics. 2019;211:333–48.

525 6. Bachtrog D, Mahajan S, Bracewell R. Massive gene amplification on a recently formed Drosophila Y
526 chromosome. Nat Ecol Evol. 2019;3:1587–97.

527 7. Choo A, Nguyen TNM, Ward CM, Chen IY, Sved J, Shearman D, et al. Identification of Y-chromosome
528 scaffolds of the Queensland fruit fly reveals a duplicated gyf gene paralogue common to many Bactrocera
529 pest species. Insect Mol Biol. 2019;28:873–86.

530 8. Tsoumani KT, Drosopoulou E, Bourtzis K, Gariou-Papalexiou A, Mavragani-Tsipidou P, Zacharopoulou A, et
531 al. Achilles, a new family of transcriptionally active retrotransposons from the olive fruit fly, with y
532 chromosome preferential distribution. PLoS One. 2015;10:18–38.

533 9. Hall AB, Qi Y, Timoshevskiy V, Sharakhova M V., Sharakhov I V., Tu Z. Six novel Y chromosome genes in
534 Anopheles mosquitoes discovered by independently sequencing males and females. BMC Genomics.
535 2013;14.

536 10. Bernardo Carvalho A, Clark AG. Efficient identification of y chromosome sequences in the human and
537 drosophila genomes. Genome Res. 2013;23:1894–907.

538 11. Dyck VA, Hendrichs J, Robinson AS. Sterile Insect Technique Principles and Practice in Area-Wide
539 Integrated Pest Management. 2021.

540 12. Hoy A. Marjorie. Insect Molecular Genetics. 4th edition. 2018.

541 13. Plá I, García de Oteyza J, Tur C, Martínez MÁ, Laurín MC, Alonso E, et al. Sterile insect technique
542 programme against mediterranean fruit fly in the valencian community (Spain). Insects. 2021;12:1–15.

543 14. Meccariello A, Salvemini M, Primo P, Hall B, Koskinioti P, Dalíková M, et al. Maleness-on-the-Y (MoY)
544 orchestrates male sex determination in major agricultural fruit fly pests. Science (1979). 2019;365:1457–60.

545 15. Dyck VA, Hendrichs J, Robinson AS. Sterile Insect Technique; Principles And Practice In Area-Wide
546 Integrated Pest Management. Dordrecht, The Netherlands: Springer; 2005.

547 16. Zhou Y, Zhan X, Jin J, Zhou L, Bergman J, Li X, et al. Eighty million years of rapid evolution of the primate
548 Y chromosome. Nat Ecol Evol. 2023;:16–9.

549 17. Bayega A, Djambazian H, Tsoumani KT, Gregoriou ME, Sagri E, Drosopoulou E, et al. De novo assembly of
550 the olive fruit fly (Bactrocera oleae) genome with linked-reads and long-read technologies minimizes gaps
551 and provides exceptional y chromosome assembly. BMC Genomics. 2020;21:1–21.

552 18. Sagri E, Reczko M, Tsoumani KT, Gregoriou M-E, Harokopos V, Mavridou A-M, et al. The molecular
553 biology of the olive fly comes of age. BMC Genom Data. 2014;15:S8.

554 19. Zacharopoulou A, Augustinos AA, Drosopoulou E, Tsoumani KT, Gariou-Papalexiou A, Franz G, et al. A
555 review of more than 30 years of cytogenetic studies of Tephritidae in support of sterile insect technique and
556 global trade. Entomol Exp Appl. 2017;164:204–25.

557 20. Falcon. A set of tools for fast aligning long reads for consensus and assembly.
558 <https://github.com/PacificBiosciences/FALCON>.

559 21. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome
560 assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.

561 22. Chevreux B, Wetter T, Suhai S. Genome Sequence Assembly Using Trace Signals and Additional Sequence
562 Information. In: German Conference on Bioinformatics. 1999.

23. Dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, et al. FlyBase: Introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 2015;43:D690–7.

24. Wu Thomas D. and Reeder J and LM and BG and BMJ. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. In: Mathé Ewy and Davis S, editor. Statistical Genomics: Methods and Protocols. New York, NY: Springer New York; 2016. p. 283–334.

25. Diesh C, Stevens GJ, Xie P, De Jesus Martinez T, Hershberg EA, Leung A, et al. JBrowse 2: a modular genome browser with views of synteny and structural variation. Genome Biol. 2023;24:74.

26. R Core Team. R: A Language and Environment for Statistical Computing. 2022.

27. Calderón-Cortés N, Quesada M, Cano-Camacho H, Zavala-Páramo G. A simple and rapid method for DNA isolation from xylophagous insects. Int J Mol Sci. 2010;11:5056–64.

28. Locke J, McDermid HE. Analysis ofDrosophila chromosome4 using pulsed field gel electrophoresis. Chromosoma. 1993;102:718–23.

29. Zhang J, Luo J, Chen J, Dai J, Montell C. The role of y chromosome genes in male fertility in drosophila melanogaster. Genetics. 2020;215:623–33.

30. Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, et al. Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. 2007;:1823–36.

31. Willhoeft U, Franz G. Identification of the Sex-Determining Region of the Ceratitis capitata Y Chromosome by Deletion Mapping. Genetics. 1996;144:737–45.

32. Sved JA, Chen Y, Shearman D, Frommer M, Gilchrist AS, Sherwin WB. Extraordinary conservation of entire chromosomes in insects over long evolutionary periods. Evolution (N Y). 2016;70:229–34.

33. Hoskins RA, Smith CD, Carlson JW, Bernardo Carvalho A, Halpern A, Kaminker JS, et al. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. 2002.

34. Gatti M, Pimpinelli S. Cytological and genetic analysis of the Y chromosome of Drosophila melanogaster. Chromosoma. 1983;88:349–73.

35. Ward CM, Aumann RA, Whitehead MA, Nikolouli K, Leveque G, Gouvi G, et al. White pupae phenotype of tephritids is caused by parallel mutations of a MFS transporter. Nat Commun. 2021;12.

36. Gabrieli P, Gomulski LM, Bonomi A, Siciliano P, Scolari F, Franz G, et al. Interchromosomal duplications on the Bactrocera oleae Y chromosome imply a distinct evolutionary origin of the sex chromosomes compared to Drosophila. PLoS One. 2011;6.

Figure S. & Table S. captions

Figure 4 Benchmarking results on the Y-recall efficiency of the four assessed methods over the genome of Ceratitis capitata. Upset plot is built based on the intersections of method predictions. a) Total counts of each intersection indicated in the base of each bar plot. b) Total number of correct Y predictions of each method. c) Total size of Y predictions achieved by every method, split in the three Y categories (Y/C, Y/R, Y/M). d) Heat maps represent the Y-sensitivity (blue) and Y-specificity (green) values for every method. e) Heat map represents the mean repeat content (brown) of the contigs included in each of the intersection categories indicated above. f) Boxplots contain information regarding the log10(Size) of the contigs that fall in each of the intersection categories

601 *indicated above. g) Venn diagrams visualize the number of Y contigs commonly/uniquely identified by the four methods as per*
602 *assembly status.*

603 *Figure 5 Benchmarking results on the Y-recall efficiency of the four assessed methods over the genome of Bactrocera oleae. Upset*
604 *plot is built based on the intersections of method predictions. a) Total counts of each intersection indicated in the base of each bar*
605 *plot. b) Total number of correct Y predictions of each method. c) Total size of Y predictions achieved by every method, split in the*
606 *three Y categories (Y/C, Y/R, Y/M). d) Heat maps represent the Y-sensitivity (blue) and Y-specificity (green) values for every method.*
607 *e) Heat map represents the mean repeat content (brown) of the contigs included in each of the intersection categories indicated*
608 *above. f) Boxplots contain information regarding the log10(Size) of the contigs that fall in each of the intersection categories*
609 *indicated above. g) Venn diagrams visualize the number of Y contigs commonly/uniquely identified by the four methods as per*
610 *assembly status.*

611 *Figure S. 1 Benchmarking results on the Y-recall efficiency of the four assessed methods over the genome of Drosophila*
612 *melanogaster. Upset plot is built based on the intersections of method predictions. a) Total counts of each intersection indicated in*
613 *the base of each bar plot. b) Total number of correct Y predictions of each method. c) Total size of Y predictions achieved by every*
614 *method, split in the three Y categories (Y/C, Y/R, Y/M). d) Heat maps represent the Y-sensitivity (blue) and Y-specificity (green) values*
615 *for every method. e) Heat map represents the mean repeat content (brown) of the scaffolds included in each of the intersection*
616 *categories indicated above. f) Boxplots contain information regarding the log10(Size) of the scaffolds that fall in each of the*
617 *intersection categories indicated above. g) Venn diagrams visualize the number of Y scaffolds commonly/uniquely identified by the*
618 *four methods as per assembly status.*

619 *Figure S. 2 Violin plots of CQ values for C. capitata contigs, horizontal line indicates the 0.3 cutoff used. A) Contigs are split into four*
620 *quantiles based on size. B) Distribution of CQ values on the different curation categories.*

621 *Figure S. 3 Violin plots of ln(R-CQ) values for C. capitata contigs, horizontal line indicates the ln=1 cutoff used. A) Contigs are split*
622 *into the four quantiles based on size. B) Distribution of CQ values on the different curation categories.*

623 *Figure S. 4 Scatterplot of male VS female kmer coverage of C. capitata contigs as identified by KAMY. A) KAMY kmer coverage of*
624 *total C. capitata contigs. B) Curation of total C. capitata contigs over KAMY kmer coverage. C) Kmer coverage of putative Y C.*
625 *capitata contigs. Identified by KAMY. D) Curation of putative Y contigs identified by KAMY.*

626 *Figure S. 5 Scatterplot of the percentage of unmatched female kmers as identified by YGS over the size of C. capitata contigs. A)*
627 *Percentage of unmatched female kmers of total C. capitata contigs. B) Curation of total C. capitata contigs over percentage of*
628 *unmatched female kmers.*

629 *Figure S. 6 Violin plots of CQ values for B. oleae contigs, horizontal line indicates the 0.3 cutoff used. A) Contigs are split into four*
630 *quantiles based on size. B) Distribution of CQ values on the different curation categories.*

Figure S. 7 Violin plots of ln(R-CQ) values for B. oleae contigs, horizontal line indicates the ln=1 cutoff used. A) Contigs are split into four quantiles based on size. B) Distribution of CQ values on the different curation categories.

Figure S. 8 Scatterplot of male VS female kmer coverage of B. oleae contigs as identified by KAMY. A) KAMY kmer coverage of total B. oleae contigs. B) Curation of total B. oleae contigs over KAMY kmer coverage. C) Kmer coverage of putative Y B. oleae contigs. Identified by KAMY. D) Curation of putative Y B. oleae contigs identified by KAMY.

Figure S. 9 Scatterplot of the percentage of unmatched female kmers as identified by YGS over the size of B. oleae contigs. A) Percentage of unmatched female kmers of total B. oleae contigs. B) Curation of total B. oleae contigs over percentage of unmatched female kmers. C) Percentage of unmatched female kmers of small (< 500 kb) B. oleae contigs. D) Curation of small (< 500 kb) B. oleae contigs over percentage of unmatched female kmers.

Figure S. 10 Violin plots of CQ values for D. melanogaster scaffolded genome, horizontal line indicates the 0.3 cutoff used. A) Scaffolds are split into four quantiles based on size. B) Distribution of CQ values on the different curation categories.

Figure S. 11 Violin plots of ln(R-CQ) values for D. melanogaster scaffolds, horizontal line indicates the ln=1 cutoff used. A) Scaffolds are split into four quantiles based on size. B) Distribution of CQ values on the different curation categories.

Figure S. 12 Scatterplot of male VS female kmer coverage of D. melanogaster scaffoldss as identified by KAMY. A) KAMY kmer coverage of total D. melanogaster scaffolds. B) Curation of total D. melanogaster scaffolds over KAMY kmer coverage. C) Kmer coverage of putative Y D. melanogaster scaffolds.

Figure S. 13 Scatterplot of the percentage of unmatched female kmers as identified by YGS over the size of D. melanogaster scaffolds. A) Percentage of unmatched female kmers of total D. melanogaster scaffolds. B) Curation of total D. melanogaster scaffolds over percentage of unmatched female kmers. C) Percentage of unmatched female kmers of small (< 100 kb) D. melanogaster scaffolds. D) Curation of small (< 500 kb) D. melanogaster scaffolds over percentage of unmatched female kmers.

651 *Table S. 2 Manual curation of method outputs over C. capitata Y chromosome.*

652 *Table S. 3 Percentage of C. capitata manually curated contigs identified by every method over the total number of curations in the*
653 *respective category.*

654 *Table S. 4 Size (bp) of curated C. capitata Y contigs identified by every method and their respective Y category size.*

655 *Table S. 5 Number of PCR validated C. capitata Y contigs identified by every method, the respective Y category and total size (bp) of*
656 *PCR validated Y.*

657 *Table S. 6 Manual curation of method outputs over B. oleae Y chromosome.*

658 *Table S. 7 Percentage of B. oleae manually curated contigs identified by every method over the total number of curations in the*
659 *respective category.*

660 *Table S. 8 Size of curated B. oleae Y contigs identified by every method and their respective Y category size (bp).*

661 *Table S. 9 Number of PCR validated B. oleae Y contigs identified by every method, the respective Y category and total size (bp) of PCR*
662 *validated Y.*

663 *Table S. 10 Number of PCR validate B. oleae false positive contigs identified by every method, the respective Y category and total size*
664 *(bp) of PCR validated false positives.*

665 *Table S. 11 Manual curation of method outputs over D. melanogaster Y chromosome.*

666 *Table S. 12 Percentage of D. melanogaster manually curated contigs identified by every method over the total number of curations in*
667 *the respective category.*

668 *Table S. 13 Size of curated D. melanogaster Y contigs identified by every method and their respective Y category size (bp).*

669 *Table S. 14 Number of Y-annotated D. melanogaster scaffolds identified by every method, their respective curation and total size*
670 *(bp).*

671 *Table S. 15 Number of X;Y-annotated D. melanogaster scaffolds identified by every method, their respective curation and total size*
672 *(bp).*

673 *Table S. 16 Size (bp) of annotated D. melanogaster contigs identified by every method as putative Y and their respective*
674 *chromosomal annotation.*

675