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.

Abstract

Despite the convergent evolution of sex chromosomes on multiple organisms, so far little is known regarding the structure and content of the different Y (or W) chromosomes. The sequence characteristics acquired by the Y throughout the divergence from its X pair interfere with computational methods, causing a difficulty in the correct assembly and characterization of Y-linked regions. Even in cases where the rest of the genome is resolved in chromosome-level, Y chromosome appears fragmented in many unmapped scaffolds or erroneously assembled with autosomal regions. This limits the insights on Y chromosomes and requires the development of computational tools for the efficient detection of male-only regions in the assembly. Two novel Y-detection methods are presented here, R-CQ and KAMY, that revisit two classical ratio- and kmer-based approaches, towards the efficient characterization of Y-linked assembly regions. The R-CQ and KAMY methods were benchmarked against their predecessors CQ and YGS for their performance on identifying Y-linked regions on the genomes of two Tephritidae species, *Bactrocera oleae* and *Ceratitis capitata,* that are characterized by different Y morphologies and are sequenced with different methodologies. The efficiency and generic applicability of the methods was further validated by using the existing Y annotations of *Drosophila melanogaster*. The results indicate an improved performance for R-CQ and KAMY over the characterization of Y chromosomes from different lineages compared to CQ and YGS. In addition, the extensive manual curation of the results reported here describes the size and quality of the assembled sequence from the two Tephritidae Y chromosomes, while evaluating the performance of methods when dealing with validated mis-assemblies in contigs. We conclude with recommendations for future sequencing efforts in insects aiming towards including a contiguous Y-chromosome assembly.

Introduction

Sex chromosomes, have evolved independently multiple times across organisms, driven by the presence of multiple genetic sex determination mechanisms[1,2]. Since the initial report of XY chromosomes from the beetle *Tenebrio molitor*[3], their study had been fascinating in terms of methodological and biological aspects.

In sexually dimorphic species, two different sex chromosome systems are found, the XY system where the males are the heterogametic sex possessing the Y chromosome, and the ZW system where the W chromosome is limited to the females[1]. Although species with non-differentiated (homomorphic) sex chromosomes exist[1,4,5], it is common for the sex-limited chromosome (Y or W) to be morphologically different (heteromorphic) to its respective pair due to a wide reformation taking place during their evolution[6,7]. Through this differentiation process, the Y converts into a highly repetitive, heterochromatic and gene-poor chromosome, with the genes remaining on the Y usually exerting a male-specific/fertility-related function[8,9]. Such convergent patterns of Y morphology and content are found on genomes as distantly related as humans to flies, although the genetic maleness factors themselves and sex determination cascades share little to no homology[7,10,11]. In fact, a great variability of sex determining factors is observed even in the same class, such as insects[12]. However, our knowledge regarding the evolutionary events leading to the different sex chromosome trajectories as well as their sequence, functional content and organization are limited to few lineages.

Despite the interest in providing answers to sex determination and inferring a universal role for sex chromosomes, studies on those elements have been methodologically challenging. In the case of homomorphic sex chromosomes, the XY pair is almost identical in sequence with the exception of the sex determining locus. For years, the presence of a genetic sex determining region (M) was reported for the mosquito species *Aedes aegypti.* However, the highly repetitive sequence of the M locus hampered the identification of the M factor, which was discovered only after the development and application of novel Y-detection methods[13]. Things are even more complicated for heteromorphic sex chromosomes, as the repeats extend almost throughout the length of the Y[7,9]. This repetitive character, together with the expansion of heterochromatin makes the correct assembly and downstream identification of those elements a complex task[14,15]. Insects were among the first organisms to be sequenced, and extensive efforts have been directed towards resolving the content and structure of the Y chromosome from the model species *Drosophila melanogaster*. Pioneer work provided new methods for resolving the sequence of Y[16]. Nonetheless, large part of the *D. melanogaster* Y chromosome still remains unknown and the prospect of generating a contiguous Y chromosome assembly still remote. Worse so, our knowledge regarding the Y content of Diptera species other than Drosophilidae is even more limited. An interesting example is that of the maleness factor of Tephritidae, which was recently characterized and is still the only functionally annotated gene on their Y[17].

The difficulty in studying the Y correlates to the evolutionary processes acting on them, which forms a unique sequence and structure. Although morphologically different to most of their extent, heteromorphic Y chromosomes still share sequence homology with autosomes and their respective X pair. The classical theory of sex chromosome evolution predicts that those elements evolve from a pair of autosomes that begin to differentiate upon acquisition of a sex determining role[18]. Due to the progressive divergence of X and Y chromosomes, common X;Y regions might persist known as pseudoautosomal, while the instance of X-linked genes being amplified/multi-copied on the Y is a common observation of heteromorphic sex chromosome pairs[19,20]. In addition to this, similar repetitive elements are mapped both on Y and autosomal regions[21], therefore increasing the non-unique part of the Y. Such sequence peculiarities, together with their highly repetitive and heterochromatic character, cause bioinformatic approaches to underperform. It is not rare for the sequence of heteromorphic Ys to appear fragmented in many small contigs or erroneously assembled in the final assembly[14,15], leading many scientific groups towards female-only whole genome projects to avoid the noise generated by quirky Y chromosomes.

Even in cases where males are introduced in the assembly, an intrinsic problem arises: how to identify contigs that include sequences of the sex chromosomes, and especially of the fragmented and mis-assembled Y. The Chromosome Quotient (CQ) algorithm had been the gold standard for the characterization of sex chromosome sequences[14]. Its implementation is based on calculating the ratio of female to male read counts that align on a specific sequence (eg. scaffold, contig, transcript), which is related to the variation of X and Y copy number between the two sexes. However, the aforementioned problems regarding non-unique regions of the Y requires the species-specific optimization of the algorithm to account for repetitive elements and shared X;Y regions. In addition, the implementation of the CQ method is based on the alignment of DNA-reads from male and female datasets on the reference genome, a task that is time consuming and computationally demanding. On the other hand, kmer based Y-detection algorithms have been developed, with Y-Genome Scan (YGS) being the more eminent implementation[15]. This approach is based on the generation of a female-specific kmer database from DNA sequencing reads, that is then used to identify regions of the genome unmatched by female kmers. The employment of a kmer database for Y-detection surpasses the need for whole genome read alignments and considers the Y-linkage of regions based on the precision of kmers derived from raw-reads, therefore detects even small variation between Y and autosomal regions. However, such methods fall behind on the characterization of X-linked regions since X-derived kmers are present both in male and female genomic datasets. Despite efforts for the development of more sophisticated Y-detection methods[22], it is common for those approaches to be used for Y-identification in recent genomic projects[23,24].

Tephritidae consist of a large family of insects with many economically important members[25]. The Mediterranean fruit fly *Ceratitis capitata* is considered one of the most destructive members of Tephritidae pest species, mainly due to the ability of female insects ovipositing in a large number of fruits and

104 vegetables[26]. On the other hand, the olive fruit fly, *Bactrocera oleae,* is a pest whose destructive action is
105 limited to olive trees since its larvae are monophagous and feed solely on the flesh of unripe olives[27]. All well-
106 studied members contain heteromorphic pairs of XY sex chromosomes with the Tephritidae maleness factor
107 being recently characterized on the Y chromosome of representative species[17,28]. Otherwise, the content and
108 structure of Tephritidae Y chromosomes is largely unknown. In this study we present major improvements
109 for understanding Tephritid Y. Initially, we employ the two published algorithms, CQ and YGS, on the genomes
110 of two Tephritidae species, *Ceratitis capitata* and *Bactrocera oleae*, and on the Diptera model *Drosophila*
111 *melanogaster*. Furthermore, we develop R-CQ and KAMY which are optimized versions of a ratio-based and
112 kmer-based Y-detection approaches, targeting in their trivial and generic applicability. We manually curate
113 the outputs of each method as for Y-linkage and assembly quality, built a core Y dataset and use it to
114 benchmark the Y-sequence identification efficiency for each of the four assessed methods.

115

116 Materials and methods

117 Species datasets

118 In this work, the publicly available reference genome assemblies of *Ceratitis capitata*, *Bactrocera olea and*
119 *Drosophila melanogaster* with accession numbers GCA_905071925.1, GCA_001188975.4, GCF_000001215.4
120 were used. To derive contigs from scaffolds, genome sequence files were split when 100 or more N bases
121 were found, except *D. melanogaster* where the scaffolded genome was used. After employment of each
122 method, small contigs of < 3 kb size were filtered out from downstream analysis. The DNA-Seq Illumina data
123 from male and female samples used in this work were acquired from NCBI's SRA: SRR8788622, SRR8788979
124 (*B. oleae);* ERR4026336, ERR4026337 (*C. capitata);* ERR2163720, ERR2163721 (*D. melanogaster)*.

125 Implementation of R-CQ

126 Reverse-CQ (R-CQ) is a Perl algorithm based on the traditional Chromosome Quotient (CQ) approach that
127 calculates the male/female ratio of depth coverage with additional steps to account for repetitive regions
128 and library depth variability (*Supplementary material: R-CQ.pl*). The input datasets are the reference genome
129 sequence and read coverage data from the two sexes of female and male. In this work, whole genome DNA-
130 seq alignments were performed using GSNAP [version 2023-02-17] and the resulting concordant-unique
131 BigWig files were used for R-CQ calculation. A window size of 50 bp was used as it was found to perform
132 best, and contigs CAJHJT010000034.1.84, LGAM02035380.1 and NT_033779.5 were used as normalizers for
133 *C. capitata, B. oleae* and *D. melanogaster* respectively. Based on empirical observations, a threshold of
134 ln(median(R-CQ)) > 1 was used to mark a putative Y.

135 Implementation of KAMY

136 "Kmer-count Algorithm Meryl identifies the Y" (KAMY) is a pipeline that predicts Y region in an assembly
137 through estimating the proportion of a contig sequence being supported by a male- or female-specific kmer

database (i.e. a sequence of length k). Example run of KAMY is available in Supplementary material. The male and female DNA-seq datasets were trimmed based on quality and putative contamination via justpreprocessmyreads of JAMg (https://github.com/genomecuration/JAMg) and an older version of Meryl (https://kmer.sourceforge.net) was used to generate the male/female common databases using a k=21 kmer size. To improve reliability, we filtered under-occurring kmers from the 3 databases and selected those with 2 or more counts. Putative Y contigs were selected based on the percentage of coverage by unique male-specific kmers, derived from subtracting values of common database coverage from the male-specific kmer coverage. Contigs having < 90 % coverage by common kmers and > 3 % unique male-specific kmer coverage.

Implementation of CQ

Chromosome Quotient (CQ) method calculates the number of female and male DNA-read alignments for a region of the assembly and uses the female-to-male ratio of read counts to infer Y-linkage [$CQ_{(Si)} = F_{(Si)}/M_{(Si)}$]. The CQ was calculated for the assemblies of the three organisms using the same horizontal criteria suggested by Hall et.al (2013) for masking of genomic repeats and putative Y selection. A soft-masked genome version was generated for *B. oleae* and *C. capitata* (*D. melanogaster* genome was downloaded as soft-masked version from FTP of https://flybase.org), and the masked repeats were removed from the assembly. The non-repetitive genome was used as input for the CQ algorithm and contigs (or scaffolds) having a CQ value of less than 0.3 were selected as putative Y.

Implementation of YGS

Y Genome Scan (YGS) method counts kmers from female DNA-sequencing datasets accounting for repetitive sequences and then calculates the percentage of each contig that is unmatched by female kmers. Male genomic data can be added to account for sequencing errors/rare polymorphisms. In this work, both male and female datasets were used as input for the YGS algorithm, performing also the optional male-validation step. The kmer size was set to 15 and filtering parameters for low-occurrence kmers were set to 5 for females and 1 for males, according to Carvalho & Clark (2013). The distribution of observations at the final YGS scatterplots of *B. oleae* and *C. capitata* implied that the suggested threshold of 95 % unmatched sequence would significantly underperform, therefore a threshold of 60 % was set for putative Y contigs.

Manual curation and benchmarking of methods

The manual curation of Y predictions was performed through observing the read coverage depth and consistency of each putative Y contig on a genome browser. For that purpose, the concordant-unique and concordant-multiple bam files outputted by the GSNAP algorithm were visualized using JBrowse2 (https://jbrowse.org/jb2). Continuously, the core Y dataset of each organism was employed to infer the Y-sensitivity and Y-specificity of each method against each of the tree assessed genomes. Y-sensitivity was calculated 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 ($Y_{specificity}=Y_{method}/Total_{method}$).

173 ### PCR validation

174 A set of curated Y contigs from *B. oleae* and *C. capitata* were selected for PCR validation based on their size.
175 A single primer pair was designed for every contig using the primer3 plugin on Geneious 2022, except for
176 contigs where large gaps were included in the contig's sequence. In that case, one primer pair was designed
177 for every contiguous region of the contig. Each primer pair was designed to amplify a region ranging from
178 ~400-1,000 bp and was assessed for its uniqueness using the Primer BLAST software against the respective
179 reference genome of the organism. PCR reactions were conducted using KAPA Taq (Merck) according to
180 manufacturer's protocol. Genomic DNA was extracted from males and females individually using pools of 4
181 insects to include more than one haplotype. Virgin *B. oleae* insects were collected from Demokritos
182 laboratory population and virgin *C. capitata* individuals were collected from the EGII laboratory populations.
183 An in house CTAB protocol was used for extracting genomic DNA, which was quantified using Quawell Q3000
184 spectrophotometer and 30ng/ul were used in each reaction. PCR reactions were carried out at a Bio-Rad T100
185 thermal cycler and the PCR products were visualized on 1 % agarose gel.

186

187 ## Results

188 Heteromorphic Y chromosomes are usually distinguished by their respective X pair through
189 karyotypic observation. At genome level, this morphological difference is translated into sequence divergence
190 between X and Y. Since X chromosomes are present in different copies among the two sexes, male and female
191 sequencing data differ: i) by the presence or not of Y-originated sequencing reads and ii) by the sequencing
192 depth on X-linked regions. Traditionally, Y-detection methods leverage male specificity, and search for
193 diverged representation of sequences between male and female genomic data.

194 ### Reverse-CQ (R-CQ)

195 The Reverse-CQ method is a re-implementation of the Chromosome Quotient (CQ) algorithm that
196 achieves Y-detection by searching for depth-coverage differences between male and female datasets among
197 assembly regions. The sexed genomic data are independently mapped on the reference genome, and only
198 uniquely aligned read pairs are considered by the method. The algorithm splits the sequences into non-
199 overlapping windows of predetermined size (default 100 bp) and the mean depth-coverage by male and
200 female data is calculated for each window. To avoid the inclusion of any bias by the variation of sequencing
201 depth between the two libraries, the method considers the depth-coverage from a user-defined autosomal
202 contig and calculates a respective normalization factor for the datasets. The female depth-coverage is
203 subjected to normalization, to avoid manipulating the depth of Y-originated coverage values. Additionally, a
204 repeat cutoff is derived from the normalized sequencing depth values, which is employed to exclude regions
205 with unresolved repeats that could disorient the final CQ calculation. Alternatively, the repeat cutoff can be
206 manually determined by the user. CQ values are calculated for each independent window as the ratio of

207 normalized male to female mean depth; the contig CQ is then determined as the median of CQ values from
208 the mean contig's windows. Since Y genomic data are ideally present only in male datasets, we select every
209 contig that exceeds an lnMedian(R-CQ) value of 1 as putative Y. Autosomal contigs are expected to cluster
210 near lnMedian(R-CQ) 0 while the X linked contigs are expected to appear near lnMedian(R-CQ) values of -0.7.

211 Kmer-counter Algorithm Meryl identifies the Y (KAMY)

212 The KAMY method identifies Y-linked regions in an assembly by assessing their coverage by male-
213 specific kmers. Meryl algorithm is employed to generate 3 kmer databases from sex-specific genomic data:
214 male, female and common. Subsequently, the contigs from a genome of interest are assessed for coverage
215 by the kmers of each database. Three values are assigned to each contig representing the percentage of the
216 sequence that is covered by kmers from the tree individual databases, which are then used for the
217 identification of Y regions. For the detection of Y contigs in this analysis, we removed contigs with more than
218 90% common kmers, subtracted the common from the male-kmer coverage (male-unique) and selected
219 contigs with more than 3% male-unique coverage. Those filtering parameters are recommended to be
220 customized based on species-specific empirical observations. Since the X chromosome is present in both male
221 and female sequencing data, the identification of X-linked contigs is not amenable with the KAMY method.

222 Manual curation of the results

223 Our newly developed R-CQ and KAMY, as well as the CQ and YGS algorithms were employed on the
224 reference genomes of two Tephritidae pest species *Bactrocera oleae, Ceratitis capitata* and on the
225 chromosome-level assembly of *Drosophila melanogaster.* The collective outputs of the methods were
226 combined and subjected to manual curation towards characterizing their chromosomal origin. Each putative-
227 Y contig was inspected on a genome browser for the contiguity, depth and consistency of coverage by male
228 & female DNA-seq reads and assigned to one of the following categories:

229 - **Y/C**: contiguous assembly of Y region with half the depth coverage of autosomal regions
230 - **Y/R**: contiguous assembly of Y with regions containing uniquely aligned male reads with more than
231 twice the expected Y depth (putative unresolved Y repeats)
232 - **Y/M**: male-specific region that has been clearly misassembled
233 - **R:** undefined repeats
234 - **M:** undefined mis-assembly
235 - **A**: autosomal region consistently covered by male and female reads at the expected depth
236 coverage
237 - **C**: region with very low DNA-seq coverage validated for non-fly origin through BLASTn

238 As suggested above, the curations of Y-regions were split into three categories based on the assembly
239 state of the contig. Since Y chromosomes are famous for being highly repetitive and prone to mis-assemblies,
240 we intended to characterize the efficiency of each method in recalling regions of the Y containing assembly

241 inconsistencies. The Y/R category focuses on repetitive regions of the Y that have been sequenced
242 independently but assembled as a single locus. On the other hand, the Y/M category was focused on contigs
243 with coverage inconsistencies suggesting mis-assemblies of Y-Y or Y-autosomal regions. Such contigs were
244 characterized by: 1) sharp peaks and drops of coverage from half the library depth (expected for the single
245 copy Y) to the expected sequencing depth (double copy autosomes) and 2) regions not supported by any
246 genomic data. Overall, the manual curation resulted in three high-trust core Y datasets that were employed
247 for benchmarking the Y-recall rate of the four assessed methods.

248 Identification of Y chromosome

249 *Ceratitis capitata*

250 The sex chromosomes of *C. capitata* comprises of an XY pair, where the X chromosome appears as
251 the largest chromosome of the karyotype while its respective Y pair is significantly smaller and contains a
252 centromere flanked by a long and a short chromosomal arm[29]. For years, the long arm of the Medfly Y was
253 known for containing the maleness factor[30], but it was only until recently that the gene responsible for
254 triggering the male-phenotype was characterized[17]. Apart from that, no other genes have been functionally
255 characterized on the medfly Y.

256 The employment of the four assessed methods on the reference genome of *C. capitata* resulted in
257 251 putative Y contigs (R-CQ: 104, KAMY: 248, YGS: 106, CQ: 60) that were subjected to manual curation
258 (Table 1). A total of 111 contigs were concluded as being Y-linked and their assembly status was curated as
259 follows; 60 well-assembled Y contigs (Y/C), 37 contigs containing unresolved repeat (Y/R) and 14 contigs
260 containing mis-assemblies (Y/M) (Table 1, Figure 1). The rest 140 contigs were considered as false positives
261 by the manual curation (A: 132, M: 7, R: 1) (Table 1). Based on this contig characterization the methods were
262 benchmarked for their performance. The newly developed KAMY method had a 100 % Y-sensitivity rate,
263 having identified all 111 of the curated Y contigs (Figure 1, Figure 6, Table 1, Table 2). However, the specificity
264 of the method was significantly compromised (44.76 %) since 137 out of the 248 predicted Y contigs were
265 curated as false positives. YGS had the second highest Y-sensitivity (95.5 %), reflecting 106 correct Y
266 predictions accompanied by a 97.25 % Y-specificity rate (3 false positive contigs) (Figure 1, Figure 7, Table 1,
267 Table 2). Similarly, R-CQ achieved 91.89 % Y-sensitivity (104 Y contigs) and 98.08 % Y-specificity (2 false
268 positive contigs) (Figure 1, Figure 5, Table 1, Table 2). The CQ method was the least efficient in recalling the
269 medfly Y, reaching a 54.05 % Y-sensitivity (62 Y contigs) with 96.77 % sensitivity (2 false positive contigs)
270 (Figure 1, Figure 4). All four methods were able to commonly identify 57 Y contigs accounting for almost half
271 of the core Y dataset (Figure 1). This group contained the largest Y contigs, with more than 50 % of those
272 being larger than 150 kb. The second intersection contained 45 Y contigs, that were missed by CQ but
273 consistently found by the other three methods. These groups contained smaller Y regions and with a higher
274 repeat content, which might be the reason for the underperformance of CQ. Generally, the KAMY, YGS and R-
275 CQ methods were effective in handling the highly repetitive Y contigs. Overall, the core Y dataset suggested

276 that despite the repetitive character (~94 % repetitive content) the Medfly Y had been surprisingly well-
277 assembled. Relatively few contigs containing assembly errors (14 Y/M), with the novel kmer method uniquely
278 contributing 4 of those.

279 Collectively, the 111 Y contigs summed up to a total ~26.8 Mb of curated Y sequence (Table 3). Out
280 of that, ~15.3 Mb were considered as well-assembled Y regions, ~9.2 Mb included unresolved repeats (Y/R)
281 and ~2.3 Mb included mis-assemblies of Y regions (Y/M). The highest sensitivity of KAMY in identifying contigs
282 of the Y containing assembly peculiarities was reflected in a ~2.1 Mb difference in the size of Y/R and Y/M,
283 compared to the second best (Figure 1, Table 3). Specifically, KAMY found all ~9.2 and ~2.3 Mb of Y/R and
284 Y/M regions while the respective numbers for YGS were ~8.4 and ~1.1 Mb. However, it should be noted that
285 due to the nature of the mis-assemblies in those regions an underestimation in contig size is expected for the
286 Y/R category and an overestimation in the Y/M.

287 *Bactrocera oleae*

288 Due to its monophagous character, the olive fruit fly is of limited economic importance compared to
289 other Tephritidae members. Nonetheless, the study of its sex chromosomes is of exceptional interest. The
290 karyotype of the olive fruit fly consists of six chromosomal pairs including a pair of heteromorphic sex
291 chromosomes[31]. Despite of Medfly and olive fly sharing the same number of chromosomes, the sex
292 chromosomes of *B. oleae* appear smaller compared to the rest[28]. Specifically, the olive fly X is approximately
293 half the size of the smaller autosome while the heteromorphic Y is highly degenerate, appearing as a dot-like
294 structure that houses the maleness factor[31]. Such sex chromosome morphology is also observed in other
295 important pest members of the Tephritidae family, like the oriental fruit fly *Bactrocera dorsalis* and the
296 Queensland fruit fly *Bactrocera tryoni*[28]. However, the limited availability of high-quality genomic data for the
297 above species constraints the study of their sex chromosomes.

298 The four Y detection methods were collectively applied on the reference genome of the olive fruit fly
299 and their outputs were combined, resulting in 792 putative Y contigs (R-CQ: 246, KAMY: 148, YGS: 579, CQ:
300 48) (Table 5). The manual curation characterized 83 contigs as Y, distributed in the three Y-categories as
301 follows: 40 well-assembled Y contigs (Y/C), 26 Y contigs containing unresolved repeats (Y/R) and 17 contigs
302 containing mis-assemblies. The rest 709 contigs were considered as false positives, with 344 of them being
303 found to represent non-fly DNA (C), 268 contained autosomal regions (A), 88 were undefined mis-assemblies
304 (M) and 9 were assemblies of undefined repeats. Subsequently, the core dataset of curated Y contigs was
305 employed for the assessment of Y-recall rates for every method. The two novel methods, R-CQ and KAMY,
306 presented the highest Y-sensitivity rate (77.11 %) having both identified 64 curated Y contigs (Figure 2, Table
307 5, Table 6, Figure 9, Figure 10). Notably, the Y contigs predicted by those two methods differed, despite the
308 total number of correct predictions being the same. Specifically, R-CQ identified 38 Y/C contigs, 20 Y/R and 6
309 Y/M while the respective numbers for KAMY were 33 Y/C, 21 Y/R and 10 Y/M. No contigs were commonly

310 identified by all four methods, rather an intersection of 29 common Y predictions between KAMY, R-CQ and
311 CQ (Figure 2). Interestingly, our newly developed methods commonly identified 21 curated Y contigs. More
312 than half of the Y contigs found within the two larger intersections appeared to be larger than 18 kb in size
313 and with a ~63 % repeat content, significantly lower than the Medfly's. Additionally, among the R-CQ
314 predictions were 9 Y contigs uniquely found by this method that appeared to be of relatively low size. This
315 suggests an advantage of the window-based R-CQ approach over the fragmented regions of Y. On the other
316 hand, KAMY was superior in uniquely identifying Y regions containing mis-assemblies. Those contigs were
317 characterized by their large size and higher repeat content compared to the first two intersections, suggesting
318 the robustness of this method over challenging Y regions. Nonetheless, in the case of the olive fruit fly
319 genome, kmer overperformed in terms of Y-specificity (43.24 %) by including 84 false positives in its
320 predictions, compared to R-CQ that was ranked as second best with 31.06 % Y-specificity and 246 false
321 positives (Figure 2, Table 5, Table 6). The CQ method ranked third in the benchmarking, achieving a 42.17 %
322 Y-sensitivity (35 Y contigs) accompanied by a 72.92 % Y-sensitivity (13 false positives) (Figure 2, Figure 8, Table
323 5, Table 6). A significant underperformance was observed for YGS, as it managed to correctly predict only 13
324 Y contigs (15.66 % Y -sensitivity) (Figure 2, Figure 11, Table 5, Table 6). Notably, most of its Y predictions
325 consisted of false positives, including 344 contamination (C), 142 autosomal (A), 77 undefined mis-assemblies
326 (M) and 3 undefined repetitive contigs (R), which concluded to a Y-specificity of only 2.25 %. However, despite
327 being the less efficient method in Y-recall, YGS uniquely contributed 5 Y/M contigs in the curation dataset that
328 were missed by the other three.

329 The size of curated olive fruit fly Y resulting from the combination of all four assessed methods
330 reached a total of ~3.9 Mb, with the expected size for the dot Y being ~4 Mb[32] (Table 7). The superiority of
331 the novel kmer method is again highlighted in the total size of identified Y, that sums at ~3.8 Mb, ~1.2 Mb
332 more than R-CQ that ranked as second best (Figure 2, Table 7). This difference is mainly attributed to the large
333 mis-assembled contigs found by kmer but missed by R-CQ. The respective Y predictions of the CQ method
334 summed up to a total of ~1.6 Mb while those of YGS were ~0.9 Mb. Again, the Y/M category is expected to
335 introduce an overestimation in the size of resolved Y, since mis-assemblies with autosomal regions were
336 identified, while on the other side, Y/R is expected to underestimate the size of resolved Y due to the
337 compiling of repetitive sequences.

338 *Drosophila melanogaster*

339 The fruit fly *Drosophila melanogaster* is a member of the Drosophilidae family of insects, being
340 considered for years as the model species for Diptera. The exhaustive curation and annotation of its genetic
341 content has provided valuable insights in the functional characterization of its genomic elements, including
342 sex chromosome[33,34]. The karyotype of *D. melanogaster* consists of five chromosomal pairs, two large
343 autosomes (chromosome 2-3), a dot-like autosome (chromosome 4) and a pair of heteromorphic sex
344 chromosomes (X-Y)[35]. Contrary to what is observed for the Medfly and olive fruit fly, the Y chromosome of *D.*

345 *melanogaster* does not house a male-determining factor. Rather, sex determination in *D. melanogaster* is
346 controlled by the expression levels of four X-linked signal elements, which is higher in individuals possessing
347 a double copy X that develop into females[36]. Therefore, it is the ratio of X:A that determines the male sex and
348 not the presence of the Y (or a maleness factor on the Y). In addition to their non-canonical role in sex
349 determination, the X chromosome of *D. melanogaster* is considered to have evolved after the reversion of an
350 ancestral X to an autosome[37]. Further studies regarding sex chromosomes suggest the independent evolution
351 between the *B. oleae* and *D. melanogaster* Y[38], which raises the question of how the four assessed Y-detection
352 methods perform on sex chromosomes in species other than Tephritidae.

353 The scaffolded reference genome of *D. melanogaster* was used as input in the four algorithms. We
354 chose not to split the chromosome-level scaffolds since this introduced many small contigs in the analysis
355 that would be either filtered out (< 3 kb) or their assembly status would be hard to curate. The collective
356 outputs of the methods included a total of 202 putative Y regions (R-CQ: 110, KAMY: 151, YGS: 190, CQ: 100)
357 (Table 10). The manual curation suggested that among the predictions was the ~3.7 Mb annotated Y scaffold,
358 an additional 139 unmapped Y contigs and 61 false positives (A: 29, M: 17, R: 15). In contrast to what was
359 observed for Tephritidae genomes, the YGS method overperformed the other three in identifying regions of
360 the *D. melanogaster* Y. Specifically, it achieved a 94.33 % Y-sensitivity rate having 133 correct Y predictions
361 (Y/C: 52, Y/R: 54, Y/M: 27) but with a rather compromised Y-specificity of 70 % (57 false positive predictions)
362 (Figure 3, Table 10, Table 11, Figure 15). KAMY ranked as second best in Y-specificity (91.49 %) by correctly
363 predicting 129 Y regions (Y/C: 52, Y/R: 53, Y/M: 24), having also the highest Y-specificity of the four (85.43 %)
364 with 22 false positive predictions (Figure 3, Figure 14, Table 10, Table 11). Comparing the CQ-based methods,
365 R-CQ was able to find 92 curated Y regions (65.25 % Y-sensitivity) with a Y-specificity of 83.64% (18 false
366 positive predictions) (Figure 3, Figure 13, Table 10, Table 11) while the CQ ranked as the least efficient by
367 achieving a 58.87 % Y-sensitivity (83 correct predictions) with 83 % Y-specificity (17 false positives) (Figure 3,
368 Figure 12, Table 10, Table 11). Apparently, the underperformance of CQ on *D. melanogaster's* Y was indicated
369 by the low Y-recall rate as well as from the absence of the large Y-annotated scaffold from the method's
370 predictions. Notably, the predictions included 52 curated Y regions that were commonly found by all four
371 methods, 29 Y regions that were missed by the underperforming CQ but found by the other three and 24 Y
372 regions commonly found by all four methods except R-CQ (Figure 3). As previously, the kmer-based methods
373 appeared to perform better with regions of the Y containing assembly peculiarities, with KAMY and YGS
374 collectively identifying 16 regions falling in the Y/R and Y/M categories. Surprisingly, the repeat content of *D.*
375 *melanogaster* Y regions was at ~36 %, significantly lower compared to the assessed Tephritidae Ys. In contrast
376 to the other two species, the non-scaffolded Y of *D. melanogaster* was highly fragmented with the majority
377 of regions being smaller than 12 kb. However, a group of 16 contigs that were found only by the kmer methods
378 were larger in size, indicating again the bias of the R-CQ window-based approach towards smaller Y regions.

379 A total ~5.5 Mb of Y sequence was curated from the predictions, the ~3.7 Mb of which comprised the
380 scaffold annotated as Y in the chromosome-level assembly (Table 12). The rest ~1.8 Mb of curated Y sequence
381 were split into annotated Y regions that were not assembled within the large Y scaffold (~517 Kb), centromeric
382 repeats (~294 Kb), repeats shared between X;Y (~52 Kb) and ~934 Kb of sequence that was not previously
383 annotated and could be considered newly identified Y (Table 15). The differences in Y-sensitivity are
384 consistently reflected at the total size of Y identified by every method, as YGS recalled the largest part of
385 assembled *D. melanogaster* Y (~5.4 Mb) followed by KAMY (~5.3 Mb), R-CQ (~4.9 Mb) and finally CQ (~780
386 Kb) (Table 12). It should be noted that the existing Y chromosome annotations of contigs were in agreement
387 with our prediction, except for few cases where the contigs were curated as R or M due to the difficulty in
388 confidently inferring the chromosomal origin (Table 13).

389 ## PCR validation

390 The core dataset of curated Y contigs from *C. capitata* and *B. oleae* was subjected to experimental
391 validation through PCR to determine the accuracy of our curations and provide a more trustworthy
392 characterization of the Y. A total of 12 curated Y contigs were subjected to PCR validation from the well-
393 assembled Y chromosome of *C. capitata.* On the other hand, due to the highly fragmented and mis-assembled
394 nature of the *B. oleae* Y, 46 curated Y contigs were selected for PCR validation in this species. The cross
395 validation of the Y curations on *D. melanogaster* was done by using the available refseq annotations derived
396 from cytological or genetical mapping[39]. The results of the PCR validation on Medfly confirmed our
397 observations regarding the high quality of the assembly and suggested the accuracy of our manual curation.
398 All of the 12 assessed contigs were concluded as being Y-linked through PCR (Table 4). Collectively, the 12
399 Medfly Y contigs summed up to a total ~7.4 Mb of PCR-validated Medfly Y sequence. Conversely, the PCR
400 validation of the olive fruit fly's Y curations indicated a mis-assembled and inconsistent assembly state for the
401 Y. Out of the 46 contigs subjected to PCR, 36 were validated while the other 10 were considered by PCR as
402 being false positives (Table 8, Table 9). The majority of those contigs (7/10) were characterized as Y/C by the
403 curation, while there were 2 Y/M and 1 Y/R additional false curations among the PCR validated dataset. The
404 recall of PCR validated Y contigs was found to be higher for the KAMY method, having included all 36 validated
405 Y contigs in its predictions, while R-CQ correctly predicted slightly less validated Y contigs (34/36). KAMY
406 performed also slightly better in avoiding PCR validated false positives, as it predicted 9/10 false positives
407 compared to R-CQ that included all 10 false positives in its predictions. Overall, the total size of PCR validated
408 *B. oleae* Y chromosome sequence summed up to ~2.4 Mb.

409 The existing annotations of *D. melanogaster* Y contigs were generally in agreement with our
410 curations. All 37 Y and 4/8 X;Y-annotated regions exceeding the size filter of 3 kb were included in our putative
411 Y dataset (Table 13, Table 14). Out of 37 Y-annotations, 6 were curated as Y/C, 18 as Y/R and 7 as Y/M.
412 Additionally, 4 annotated Y regions were curated as undefined repeats (R) and 1 as undefined mis-assembly
413 (M) since the chromosomal origin could not be confidently inferred for those regions due to high

414 inconsistency in depth and width of read coverage. Kmer-based methods were again more efficient in
415 predicting the annotated Y regions of *D. melanogaster*. Specifically, KAMY recalled 35 Y and 3 X;Y annotations
416 while YGS recalled 35 Y and 4 X;Y annotations. In total, ~4.4 Mb of *D. melanogaster* Y sequences are annotated
417 in the reference assembly.

418

419 Discussion

420 In this work, we intended to provide the methods for identifying the sequence of sex-limited
421 chromosomes. The study of sex chromosomes, and especially the sex-limited Y (or W), is a field of unique
422 interest due to the distinct evolutionary forces acting on them, while there are practical implications for
423 Integrated Pest Management approaches. Alternative population control strategies, like Sterile Insect
424 Technique (SIT), exploit the complete linkage of Y chromosome with the male phenotype to contain the
425 destructive action of pests through rearing, sterilization and mass-release in the field of sterile male-only
426 insects[40]. However, as we present here, identifying Y-linked regions is far more problematic than the literature
427 on model species suggests[14,15]. In fact, after more than forty years of effort to identify the Medfly Y-linked
428 maleness factor[30], the *Maleness-on-the-Y* (*MoY*) gene was recently identified and remains the only
429 characterized Y gene[17].

430 Biological content and dataset quality affects bioinformatic methods

431 Our work was focused on developing a methodology for studying heteromorphic Y chromosomes, with a
432 focus on Tephritidae. We selected two such species with recently published male genome assemblies for
433 which both male and female genomic data were available, the Mediterranean fruit fly *Ceratitis capitata* and
434 the olive fruit fly *Bactrocera oleae.* Despite belonging to the same family, the Y chromosomes of those species
435 vary in size: the Y of the olive fruit fly appears as a dot-like structure on the karyotype while that of the Medfly
436 is characterized by a centromeric structure and two distinct chromosomal arms[28]. Furthermore, two different
437 sequencing strategies were followed during the whole genome assembly of these species, with that of *B.*
438 *oleae* being specifically focused on resolving the sequence of the dot Y. The olive fly assembly included a 10X
439 linked-read male library with ~300× sequencing depth and a female Illumina library with ~100× coverage
440 whilst the *C. capitata* was based on two Illumina libraries of ~40× depth[41]. We hypothesized that the different
441 nature and assembly of those Ys could stress the limits of a Y identification algorithm and would reveal its
442 robustness and generic applicability. The classical ratio-based CQ approach is known to underperform on
443 repetitive Y regions and requires optimization over depth and repeat content of the input dataset[14]. On the
444 other hand, kmer-based approaches are more precise over Y repeats through the assessment of small
445 sequence regions (size of k) that might reveal polymorphisms between Y and autosomal repeats[15]. However,
446 their lability is the lack of predictions over the X chromosome. Here, two novel methods were built from
447 scratch that were named R-CQ and KAMY, which employed a ratio and a kmer approach, respectively.

448 The R-CQ method accounts for differences in the sequencing depth between libraries, automatically
449 excludes repetitive regions and assesses the male-to-female ratio in a non-overlapping window manner,
450 therefore accounts for putative mis-assemblies that could hinder Y-detection. This design appeared to
451 perform better on Tephritidae Ys over the non-optimized CQ algorithm. The repeat cutoff handled with great
452 success the highly repetitive Medfly Y while the non-overlapping windows are assumed to be responsible for
453 the identification of smaller Y contigs on the dot and fragmented olive fly Y (Figure 1, Figure 2, Figure 5, Figure
454 9). Regarding the different library size of *B. oleae,* this introduced several false positive contigs in the R-CQ
455 results therefore reducing its specificity (Figure 9, Table 5). The R-CQ predictions included regions with ~150×
456 male and ~10× female coverage throughout the length of the contig, which were curated as autosomal.
457 However, others having lower female coverage (2-3 reads) were falsely considered as Y-linked and were
458 removed only after the PCR validation. The different library preparation approach followed in this organism
459 (10× single molecule vs Illumina shotgun) is assumed to have biased such differences in sequencing depth,
460 which could not be counteracted by the normalization step. In contrast, the results of R-CQ in Medfly, for
461 which male and female libraries were prepared with the same method, suggest that the depth normalizing
462 step adjusted the small differences between libraries and resulted in a better clustering of Y, autosomal and
463 X predictions compared to the CQ algorithm (Figure 4, Figure 5).

464 Kmer-based methods are well-known for handling with great precision the repetitive and mis-assembled
465 Y regions, through taking advantage of the highly detailed sequence of short-reads to detect polymorphisms
466 deriving from the Y. Nonetheless, the implementation of YGS on Tephritidae genomes appeared to
467 underperform, with the method presenting a significant bias over Drosophila Y (Figure 1, Figure 2, Figure 3).
468 Therefore, KAMY was developed that considers the common and male-unique kmer proportion of a contig to
469 identify Y regions. Our benchmarking suggests that KAMY efficiently identifies Y chromosome sequences from
470 all three organisms, highlighting its improvement in generic applicability compared to YGS. Apparently, the
471 filtering parameters used for the common proportion of each contig resulted in the inclusion of many false
472 positive predictions on the highly repetitive Medfly Y, causing the reduced specificity of the method (Figure
473 6, Table 1). Importantly, though, the Y-sensitivity was not compromised, and KAMY yielded the most correct
474 Y predictions (Figure 1). Conversely, YGS was the best performing method on *D. melanogaster* Y, but its Y-
475 identification ability was diminished on the olive fly (Figure 1, Figure 3).

476 The variability in the performance of YGS across the three assessed genomes is attributed to the
477 difference in the sequence characteristics between different Y chromosomes. YGS detects Y regions through
478 assessing the percentage of a sequence that is not present in a female-only kmer database. In the initial
479 development of this algorithm, the authors optimized the method using a Drosophila species as a model, the
480 Y regions of which were found to have 95-100 % of their sequence unmatched by female kmers[15]. This
481 observation is consistent to the results presented here for *D. melanogaster* (Figure 15). However, the
482 percentage of Tephritidae Y sequence that was unmatched by female kmers ranged at lower levels, indicating

483 the presence of shared kmer regions between Y and autosomes, a putative characteristic of that Y lineage
484 (Figure 6, Figure 11). The case of the olive fly's Y is indicative of poor method performance, as the majority of
485 contigs from this dot-like Y had more than 40 % of their sequence present in female kmer database.
486 Consequently, Y contigs could not be independently clustered by the YGS algorithm, while the majority of Y
487 predictions in this organism comprised of false positives, mainly non-fly contaminants (Figure 11). This
488 functionality of the method was initially pinpointed by its authors[15] and can assist in improving the total
489 quality of assemblies. However, when it comes to Y-identification, the sensitivity over exogenous DNA
490 sequences is a limiting factor on the performance and general applicability. Overall, kmer-based methods
491 were found to perform better with regions of the Y containing mis-assemblies or unresolved repeats (Table
492 1, Table 5, Table 10), an important parameter for improving our knowledge over quirky Y chromosomes.
493 However, if it was for a wide sex chromosome analysis, kmer-based methods would have been unable to
494 identify the X since reads originating from that chromosome would be present in both male and female
495 datasets.

496 ## Curation as a solution

497 Our benchmarking suggests that there is no magic bullet when it comes to Y-identification methods,
498 different methods were found to perform with different efficiencies among different species, while the
499 inclusion of false positives was a common instance. The employment of more than one Y-detection method
500 and the subsequent manual curation of the collective outputs resulted in the generation of high-trust core Y
501 datasets and removed many false positive predictions that could otherwise misguide downstream analysis
502 (Table 1, Table 5, Table 10). Such curation can be performed by non-specialized users on a regular desktop by
503 using files already generated during the method implementation step. Considering the weaknesses of Y
504 detection methods, one should always review the algorithmic outputs towards a more accurate assessment
505 of the content and size of Y chromosomes. The PCR results from *C. capitata* and the existing Y-annotations of
506 *D. melanogaster* prove the validity of our curation approach. However, the experimental validation of *B. oleae*
507 Y contigs indicated the problematic nature of this assembly and suggests that PCR should always be
508 considered for a conclusive validation of Y regions in such cases.

509 ## Conclusion

510 Overall, two novel Y-detection methods, R-CQ and KAMY, were developed in this study. Two Tephritidae
511 Y chromosomes were used as models for the improvement of the methods and their efficiency and generic
512 applicability was assessed over the well annotated Y of *Drosophila melanogaster*. We concluded to the
513 superiority of kmer-based methods in the identification of Y regions in an assembly, especially of regions
514 containing assembly peculiarities, with the KAMY appearing as the most sensitive in Tephritidae Y
515 chromosomes. However, we suggest that more than one method is employed in a project since we observed
516 that Y-identification methods perform differently depending on the sex chromosome characteristics and
517 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.

References

1. Bachtrog, D. *et al.* Sex Determination: Why So Many Ways of Doing It? *PLoS Biol.* **12**, 1–13 (2014).
2. Vicoso, B. & Bachtrog, D. Numerous Transitions of Sex Chromosomes in Diptera. *PLoS Biol.* **13**, 1–22 (2015).
3. Stevens, N. Studies in spermatogenesis with especial reference to the 'Accessory chromosome'. (1905).
4. Presgraves, D. C. & Orr, H. A. Haldane's rule in taxa lacking a hemizygous X. *Science (80-.).* **282**, 952–954 (1998).
5. Fontaine, A. *et al.* Extensive genetic differentiation between homomorphic sex chromosomes in the mosquito vector, Aedes aegypti. *Genome Biol. Evol.* **9**, 2322–2335 (2017).
6. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philos. Trans. R. Soc. B Biol. Sci.* **355**, 1563–1572 (2000).
7. Bachtrog, D. Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124 (2013).
8. Mahajan, S. & Bachtrog, D. Convergent evolution of y chromosome gene content in flies. *Nat. Commun.* **8**, (2017).
9. Furman, B. L. S. *et al.* Sex Chromosome Evolution: So Many Exceptions to the Rules. *Genome Biol. Evol.* **12**, 750–763 (2020).
10. Masataka, S. G. *Sex Determination Cascade in Insects: A Great Treasure House of Alternative Splicing*. (Springer US, 2018).
11. Lints, R. & Emmons, S. W. Regulation of sex-specific differentiation and mating behavior in C. elegans by a new member of the DM domain transcription factor family. *Genes Dev.* **16**, 2390–2402 (2002).
12. Saccone, G., Pane, A. & Polito, L. C. Sex determination in flies, fruitflies and butterflies. *Genetica* **116**, 15–23 (2002).
13. Hall, A. B. *et al.* Insights into the preservation of the homomorphic sex-determining chromosome of Aedes aegypti from the discovery of a male-biased gene tightly linked to the M-locus. *Genome Biol. Evol.* **6**, 179–191 (2014).
14. Hall, A. B. *et al.* Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females. *BMC Genomics* **14**, (2013).
15. Bernardo Carvalho, A. & Clark, A. G. Efficient identification of y chromosome sequences in the human and drosophila genomes. *Genome Res.* **23**, 1894–1907 (2013).
16. Chang, C. H. & Larracuente, A. M. Heterochromatin-enriched assemblies reveal the sequence and organization of the drosophila melanogaster Y Chromosome. *Genetics* **211**, 333–348 (2019).
17. Meccariello, A. *et al.* Maleness-on-the-Y (MoY) orchestrates male sex determination in major agricultural fruit fly pests. *Science (80-.).* **365**, 1457–1460 (2019).
18. Bull, J. J. Sex determining mechanisms: an evolutionary perspective. *Experientia. Suppl.* **55**, 93–115 (1987).
19. Bachtrog, D., Mahajan, S. & Bracewell, R. Massive gene amplification on a recently formed Drosophila Y chromosome. *Nat. Ecol. Evol.* **3**, 1587–1597 (2019).
20. Choo, A. *et al.* Identification of Y-chromosome scaffolds of the Queensland fruit fly reveals a duplicated gyf gene paralogue common to many Bactrocera pest species. *Insect Mol. Biol.* **28**, 873–886 (2019).
21. Tsoumani, K. T. *et al.* Achilles, a new family of transcriptionally active retrotransposons from the olive fruit fly, with y chromosome preferential distribution. *PLoS One* **10**, 18–38 (2015).
22. Rangavittal, S. *et al.* DiscoverY: A classifier for identifying y chromosome sequences in male assemblies. *BMC Genomics* **20**, 1–11 (2019).
23. Zhou, Y. *et al.* Eighty million years of rapid evolution of the primate Y chromosome. *Nat. Ecol. Evol.* 16–19 (2023) doi:10.1038/s41559-022-01974-x.
24. Bayega, A. *et al.* De novo assembly of the olive fruit fly (Bactrocera oleae) genome with linked-reads and long-read technologies minimizes gaps and provides exceptional y chromosome assembly. *BMC Genomics* **21**, 1–21 (2020).
25. Hoy A. Marjorie. *Insect Molecular Genetics*. (2018).
26. Nicanor, L. J. *Host Plants of the Mediterranean Fruit Fly (Diptera: Tephritidae): An Annotated World Review*. (Entomological Society of America, 1991).
27. Daane, K. M. & Johnson, M. W. Olive fruit fly: Managing an ancient pest in modern times. *Annu. Rev. Entomol.* **55**, 151–169 (2010).
28. Zacharopoulou, A. *et al.* A review of more than 30 years of cytogenetic studies of Tephritidae in support of sterile insect technique and global trade. *Entomol. Exp. Appl.* **164**, 204–225 (2017).

29. Bedo, D. G. Polytene chromosome mapping in Ceratitis capitata (Diptera: Tephritidae) . *Genome* **29**, 598–611 (1987).

30. Willhoeft, U. & Franz, G. Identification of the Sex-Determining Region of the Ceratitis capitata Y Chromosome by Deletion Mapping. *Genetics* **144**, 737–745 (1996).

31. Canovai, R., Macelloni, G. & Galleni, L. Karyology of some Italian populations of the olive fly Bactrocera oleae (Gmelin) (Diptera: Tephritidae). *Caryologia* **49**, 247–255 (1996).

32. Locke, J. & McDermid, H. E. Analysis ofDrosophila chromosome4 using pulsed field gel electrophoresis. *Chromosoma* **102**, 718–723 (1993).

33. Zhang, J., Luo, J., Chen, J., Dai, J. & Montell, C. The role of y chromosome genes in male fertility in drosophila melanogaster. *Genetics* **215**, 623–633 (2020).

34. Lin, M. F. *et al.* Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. 1823–1836 (2007) doi:10.1101/gr.6679507.

35. Aulard, S., Monti, L., Chaminade, N. & Lemeunier, F. Mitotic and polytene chromosomes: Comparisons between Drosophila melanogaster and Drosophila simulans. *Genetica* **120**, 137–150 (2004).

36. Nöthiger, R. & Steinmann-Zwicky, M. Sex determination in Drosophila. *Trends Genet.* **1**, 209–215 (1985).

37. Vicoso, B. & Bachtrog, D. Reversal of an ancient sex chromosome to an autosome in Drosophila. *Nature* **499**, 332–335 (2013).

38. Gabrieli, P. *et al.* Interchromosomal duplications on the Bactrocera oleae Y chromosome imply a distinct evolutionary origin of the sex chromosomes compared to Drosophila. *PLoS One* **6**, (2011).

39. Dos Santos, G. *et al.* FlyBase: Introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.* **43**, D690–D697 (2015).

40. Dyck, V. A., Hendrichs, J. & Robinson, A. S. *Sterile Insect Technique; Principles And Practice In Area-Wide Integrated Pest Management*. *Sterile Insect Technique* (Springer, 2005). doi:10.1201/9781003035572-12.

41. Papanicolaou, A. *et al.* The whole genome sequence of the Mediterranean fruit fly, Ceratitis capitata (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species. *Genome Biol.* **17**, 1–31 (2016).

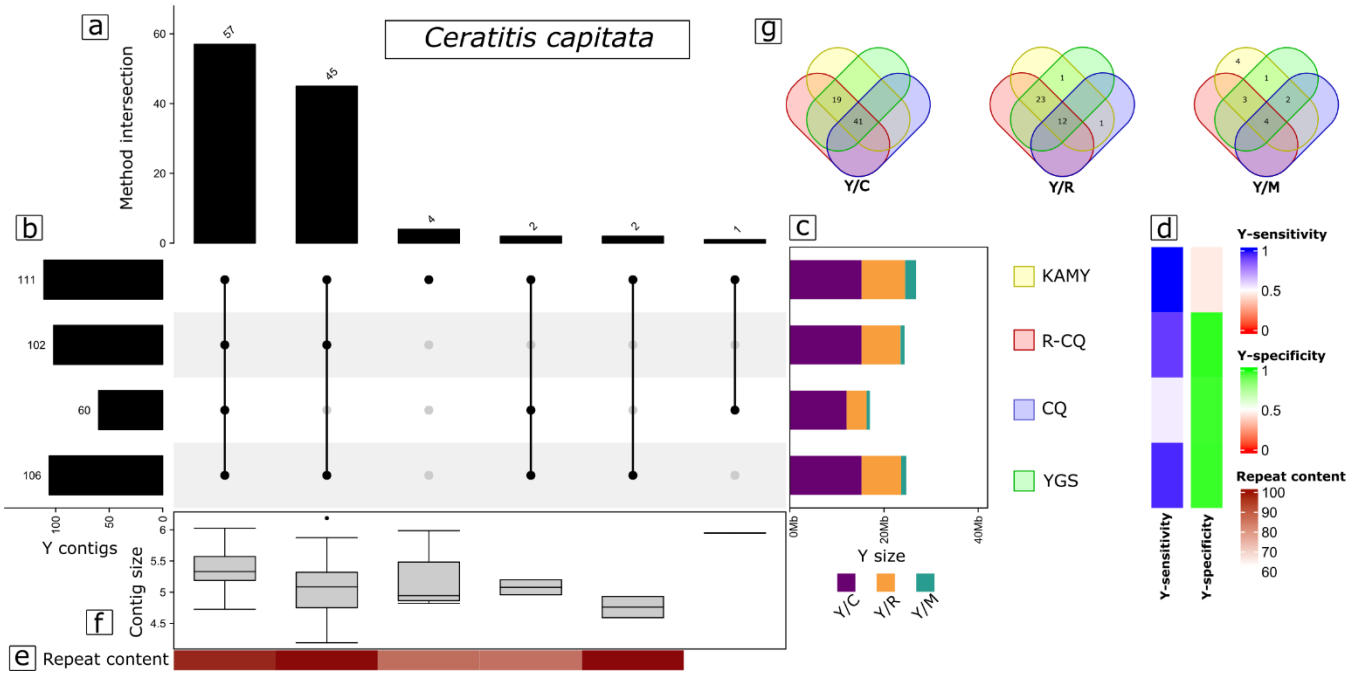

Figure 1 Benchmarking results on the Y-recall efficiency of the four assessed methods over the genome of Ceratitis capitata. Upset plot is built based on the intersections of method predictions. a) Total counts of each intersection indicated in the base of each bar plot. b) Total number of correct Y 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.

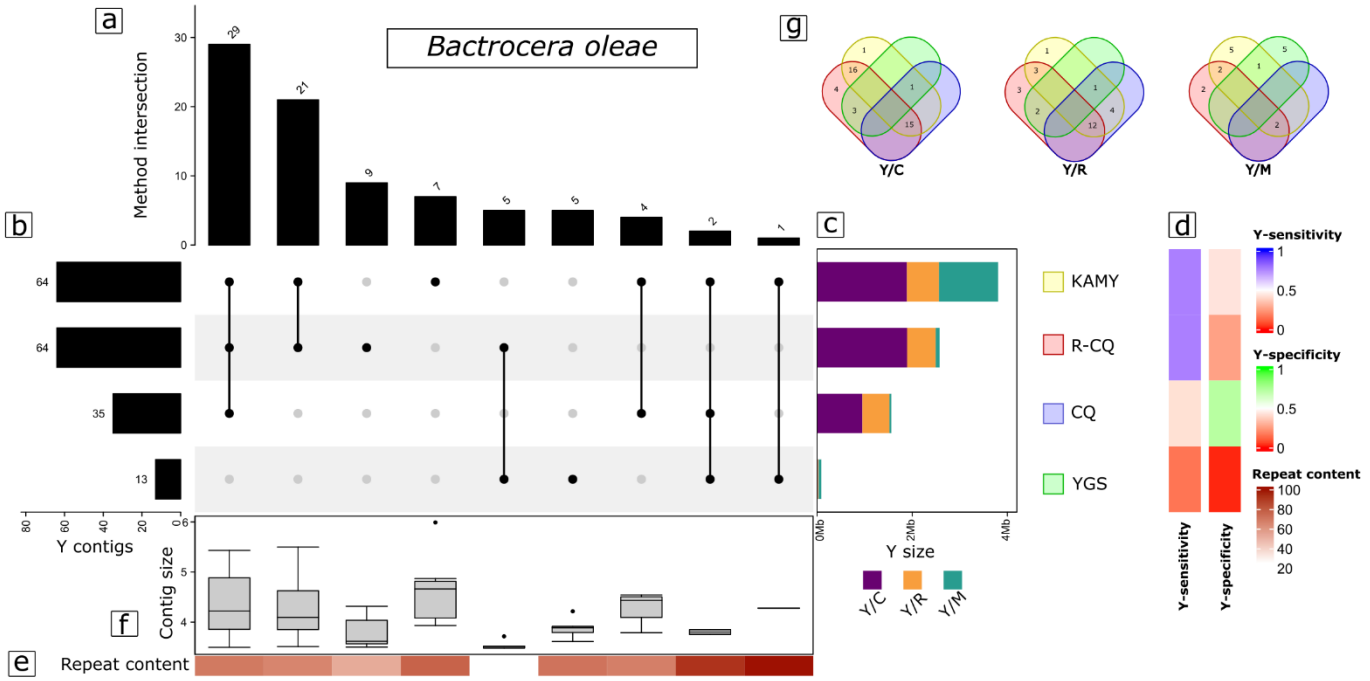

Figure 2 Benchmarking results on the Y-recall efficiency of the four assessed methods over the genome of Bactrocera oleae. Upset plot is built based on the intersections of method predictions. a) Total counts of each intersection indicated in the base of each bar plot. b) Total number of correct Y predictions 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.

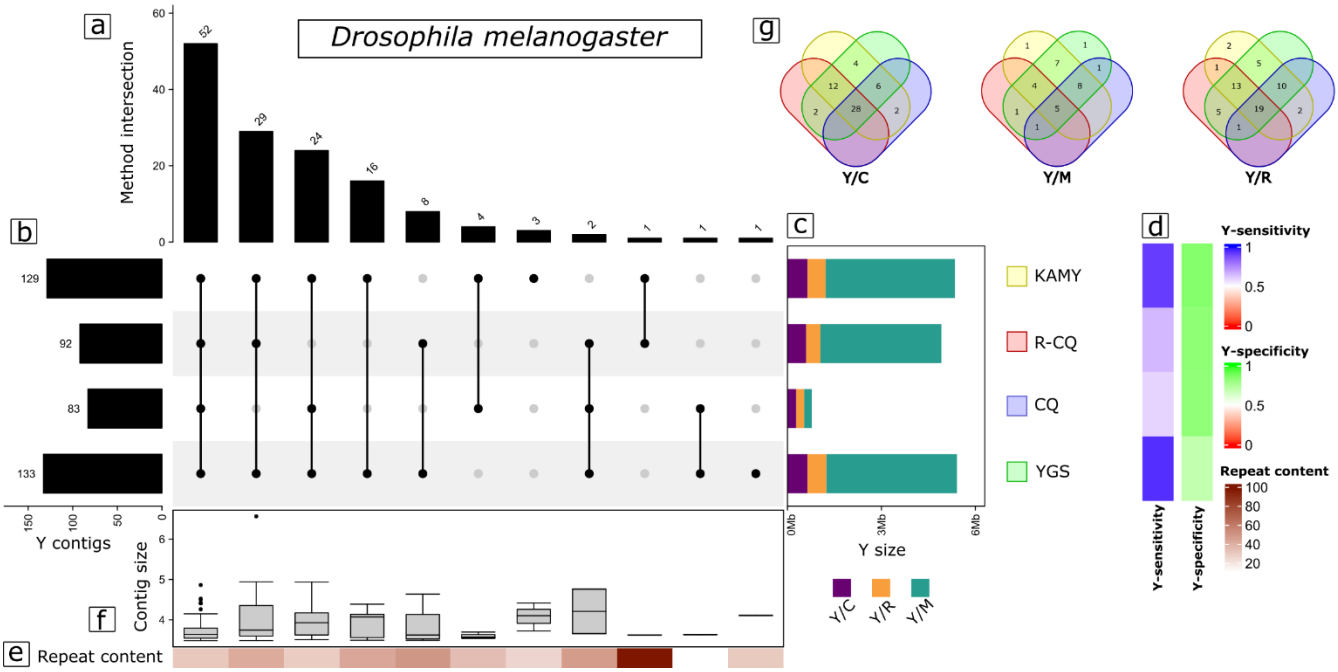

Figure 3 Benchmarking results on the Y-recall efficiency of the four assessed methods over the genome of Drosophila melanogaster. Upset plot is built based on the intersections of method predictions. a) Total counts of each intersection indicated in the base of each bar plot. b) Total number of correct 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 scaffolds included in each of the intersection categories indicated above. f) Boxplots contain information regarding the log10(Size) of the scaffolds that fall in each of the intersection categories indicated above. g) Venn diagrams visualize the number of Y scaffolds commonly/uniquely identified by the four methods as per assembly status.

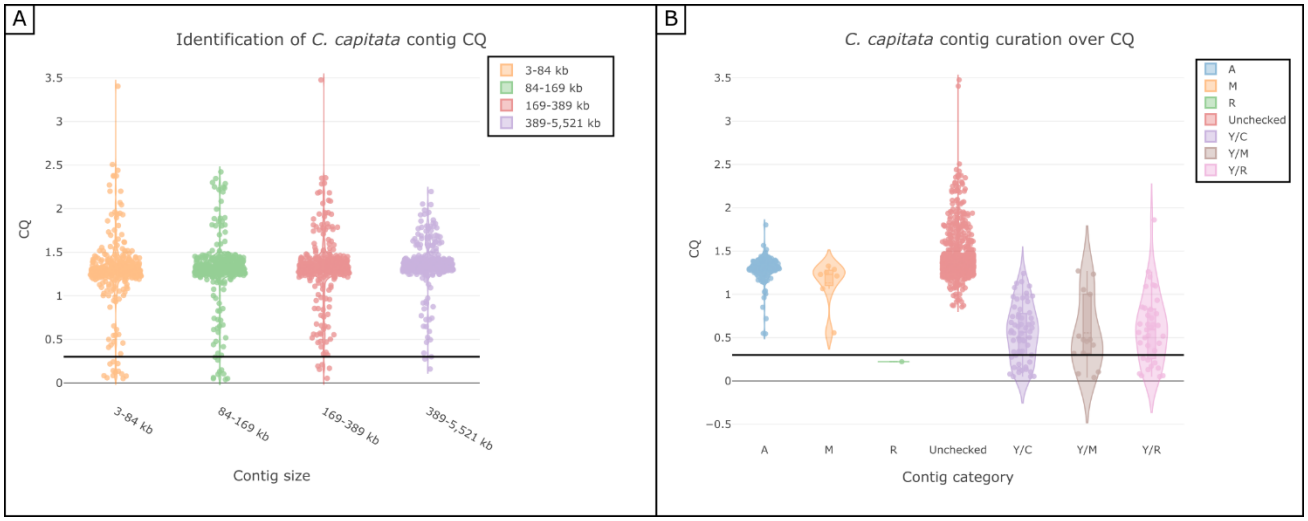

Figure 4 Violin plots of CQ values for C. capitata contigs, horizontal line indicates the 0.3 cutoff used. A) Contigs are split into four quantiles based on size. B) Distribution of CQ values on the different curation categories.

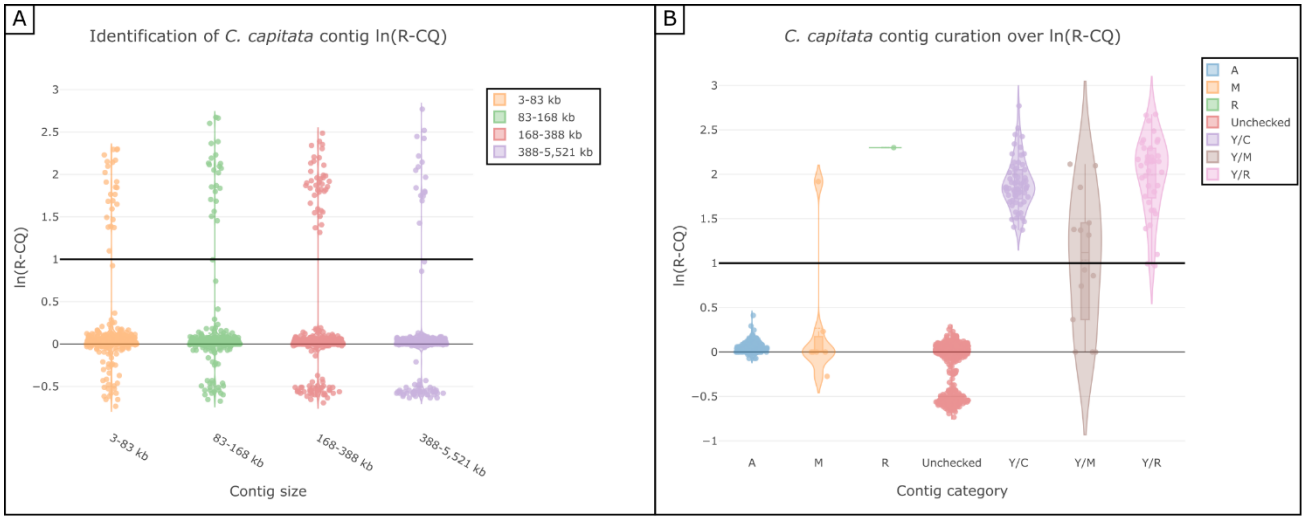

Figure 5 Violin plots of ln(R-CQ) values for C. capitata contigs, horizontal line indicates the ln=1 cutoff used. A) Contigs are split into the four quantiles based on size. B) Distribution of CQ values on the different curation categories.

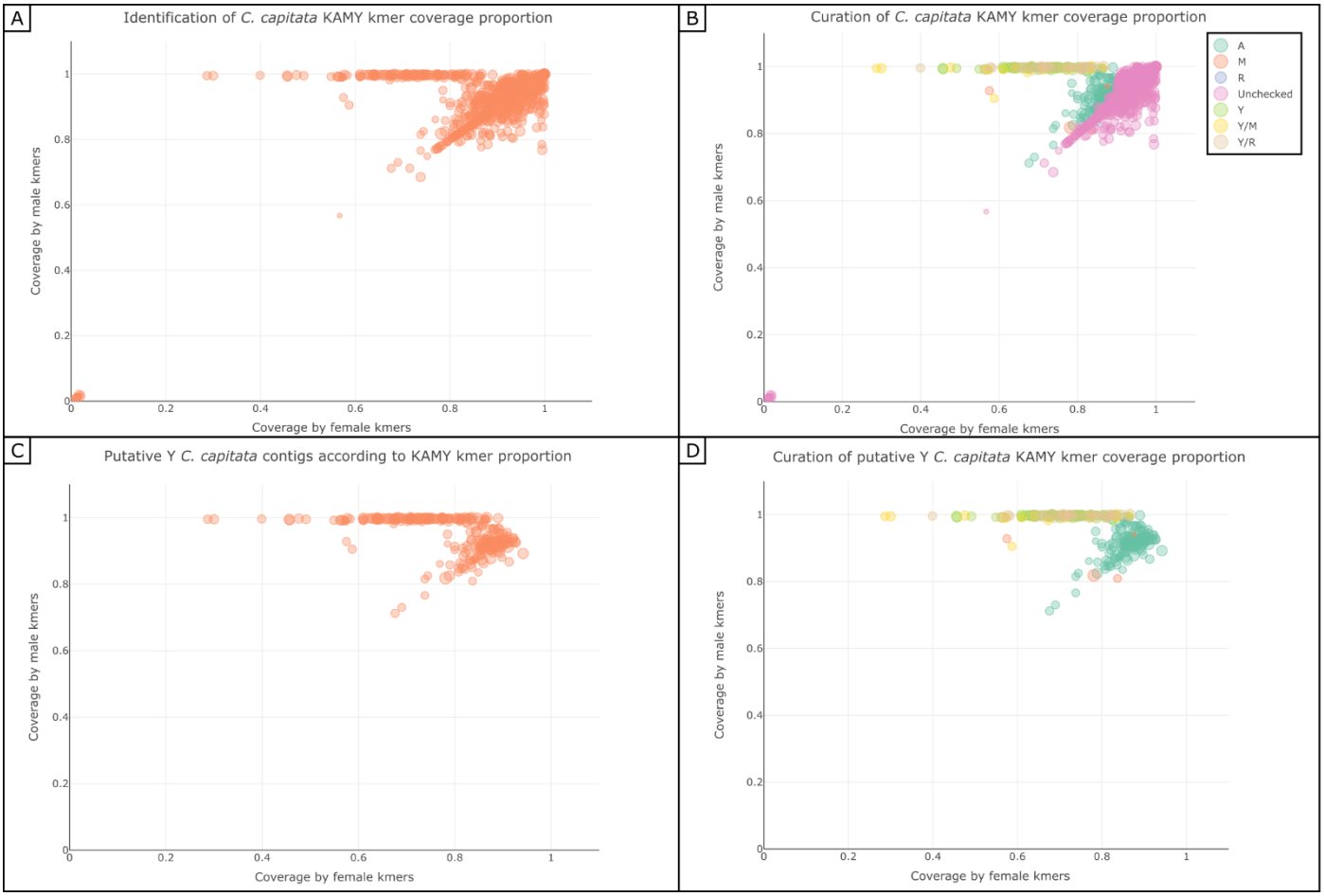

Figure 7 Scatterplot of male VS female kmer coverage of C. capitata contigs as identified by KAMY. A) KAMY kmer coverage of total C. capitata contigs. B) Curation of total C. capitata contigs over KAMY kmer coverage. C) Kmer coverage of putative Y C. capitata contigs. Identified by KAMY. D) Curation of putative Y contigs identified by KAMY.

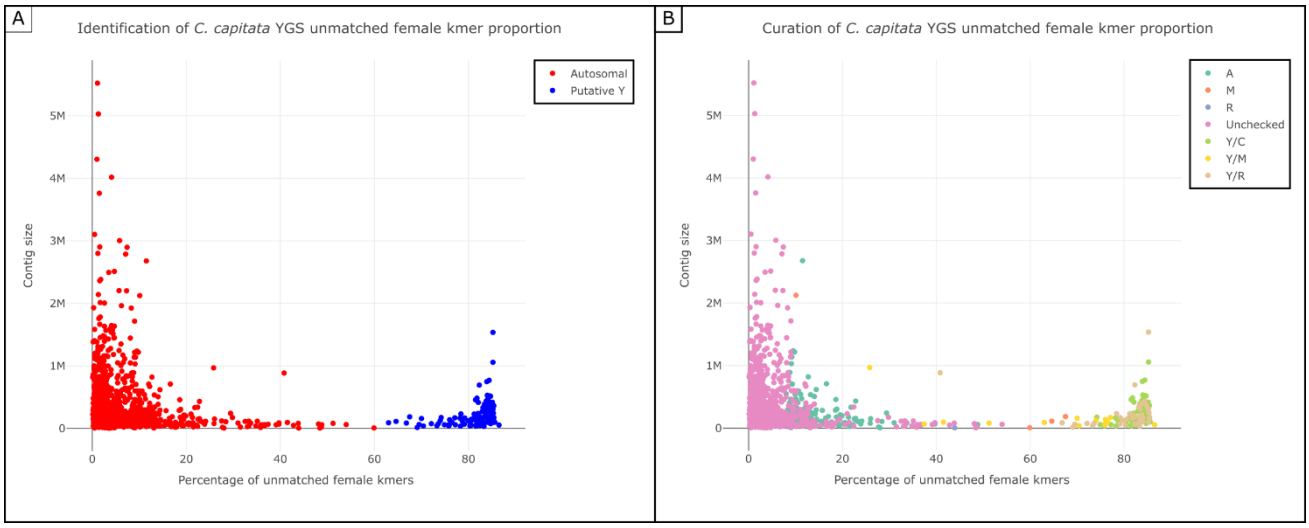

Figure 6 Scatterplot of the percentage of unmatched female kmers as identified by YGS over the size of C. capitata contigs. A) Percentage of unmatched female kmers of total C. capitata contigs. B) Curation of total C. capitata contigs over percentage of unmatched female kmers.

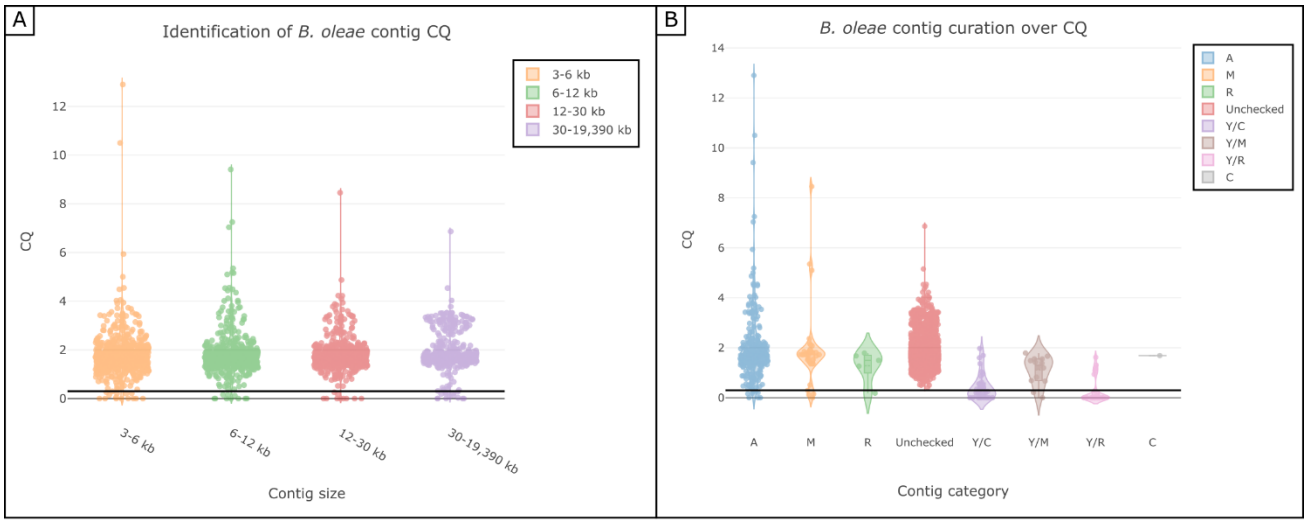

Figure 8 Violin plots of CQ values for B. oleae contigs, horizontal line indicates the 0.3 cutoff used. A) Contigs are split into four quantiles based on size. B) Distribution of CQ values on the different curation categories.

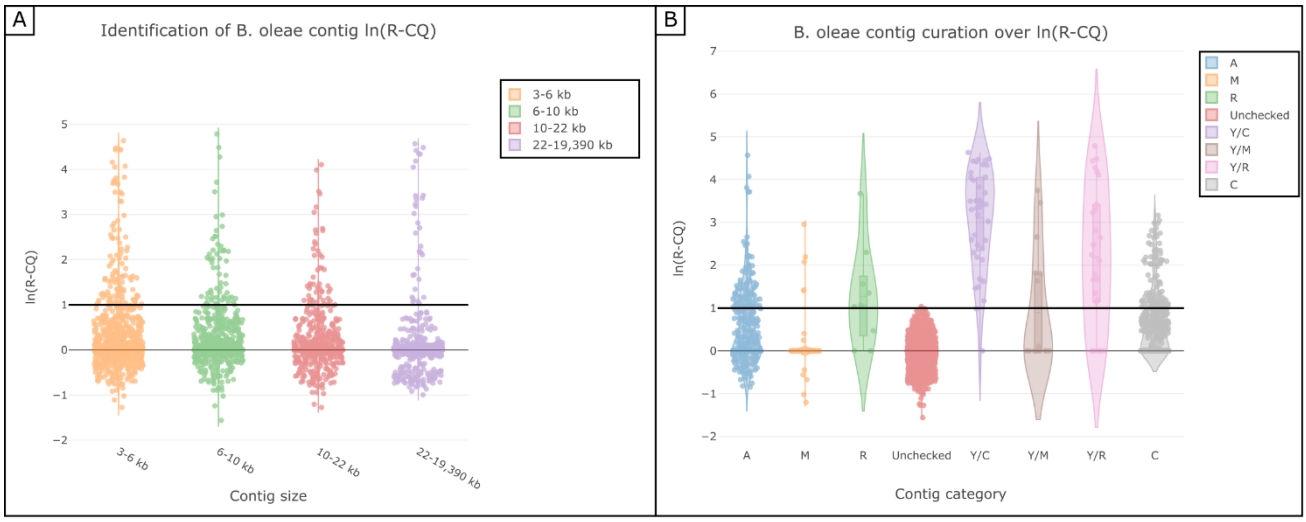

Figure 9 Violin plots of ln(R-CQ) values for B. oleae contigs, horizontal line indicates the ln=1 cutoff used. A) Contigs are split into four quantiles based on size. B) Distribution of CQ values on the different curation categories.

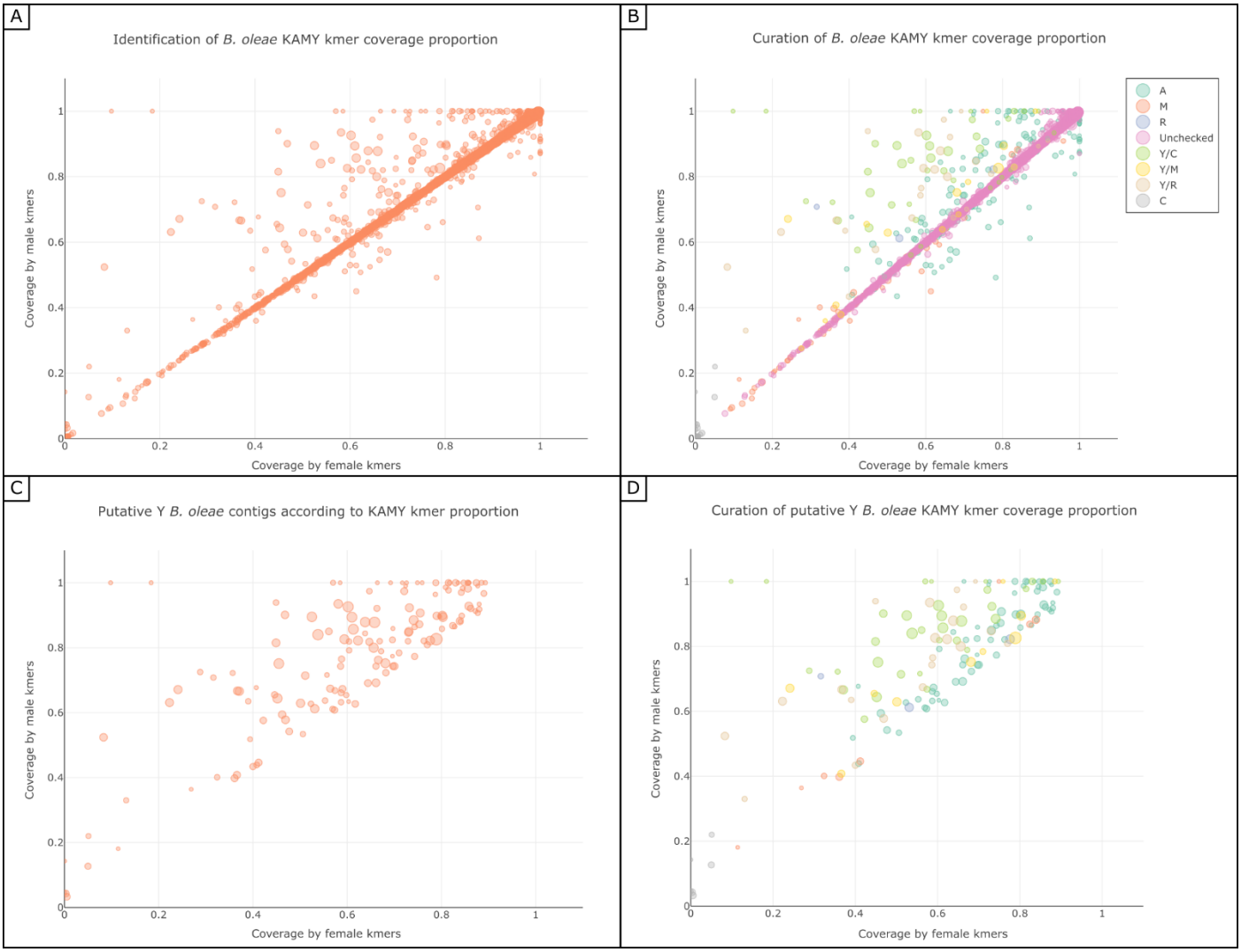

Figure 10 Scatterplot of male VS female kmer coverage of B. oleae contigs as identified by KAMY. A) KAMY kmer coverage of total B. oleae contigs. B) Curation of total B. oleae contigs over KAMY kmer coverage. C) Kmer coverage of putative Y B. oleae contigs. Identified by KAMY. D) Curation of putative Y B. oleae contigs identified by KAMY.

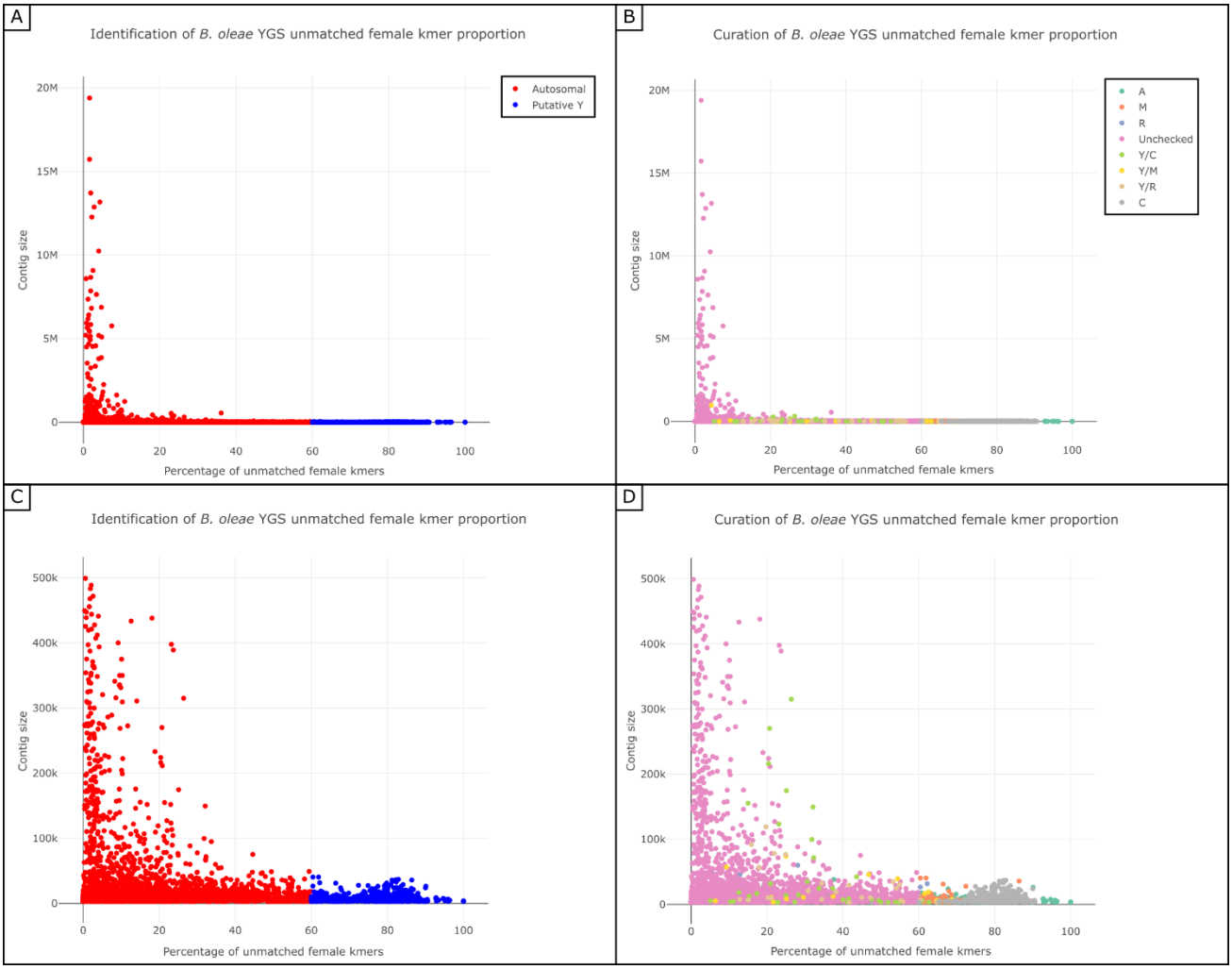

Figure 11 Scatterplot of the percentage of unmatched female kmers as identified by YGS over the size of B. oleae contigs. A) Percentage of unmatched female kmers of total B. oleae contigs. B) Curation of total B. oleae contigs over percentage of unmatched female kmers. C) Percentage of unmatched female kmers of small (< 500 kb) B. oleae contigs. D) Curation of small (< 500 kb) B. oleae contigs over percentage of unmatched female kmers.

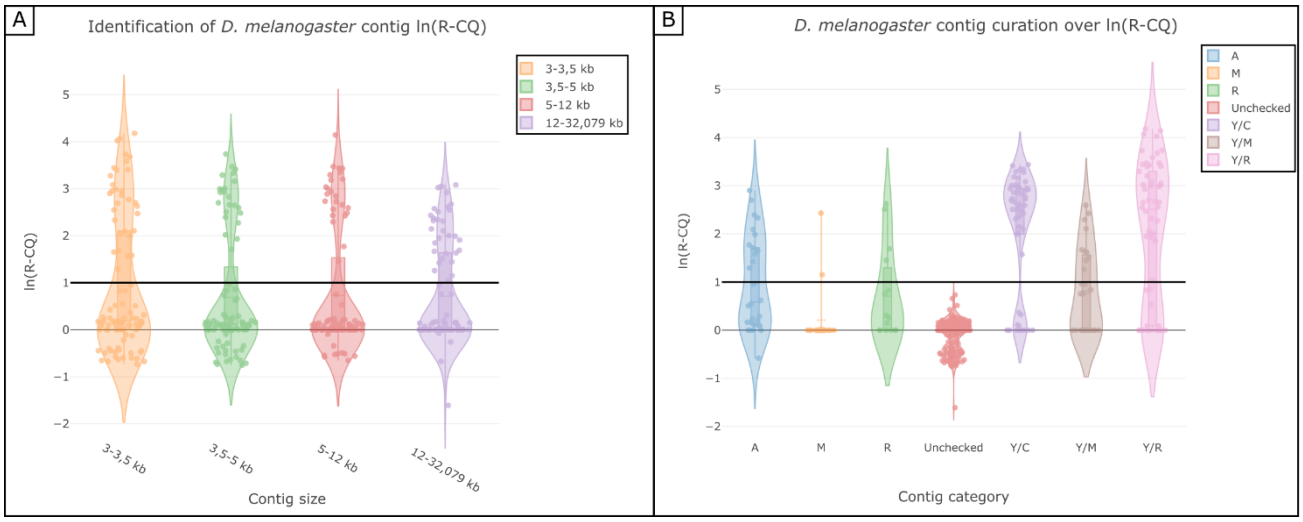

Figure 12 Violin plots of ln(R-CQ) values for D. melanogaster scaffolds, horizontal line indicates the ln=1 cutoff used. A) Scaffolds are split into four quantiles based on size. B) Distribution of CQ values on the different curation categories.

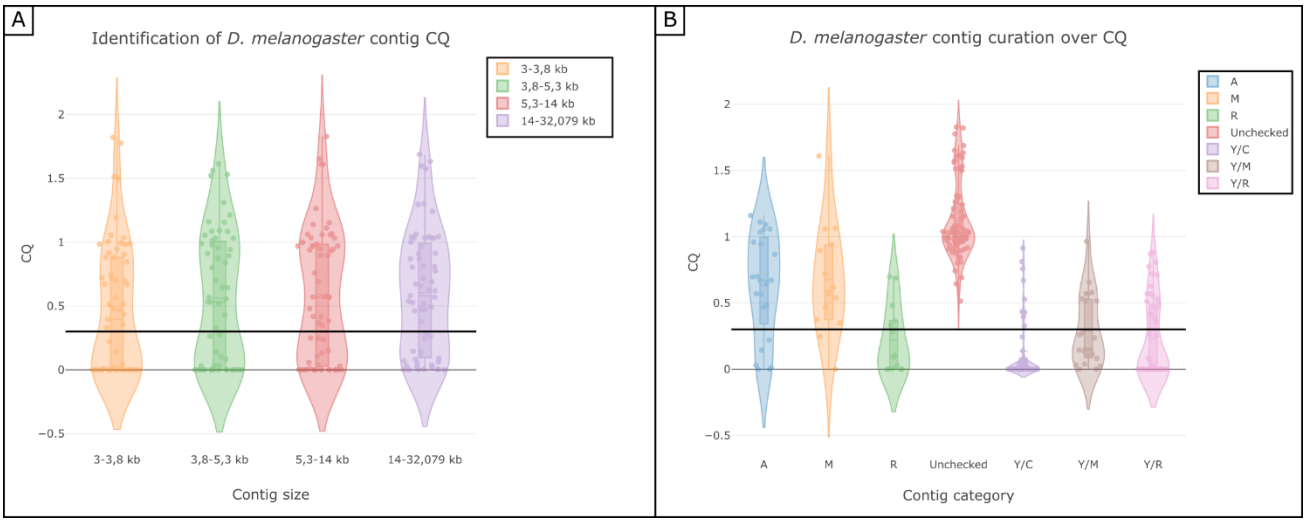

Figure 13 Violin plots of CQ values for D. melanogaster scaffolded genome, horizontal line indicates the 0.3 cutoff used. A) Scaffolds are split into four quantiles based on size. B) Distribution of CQ values on the different curation categories.

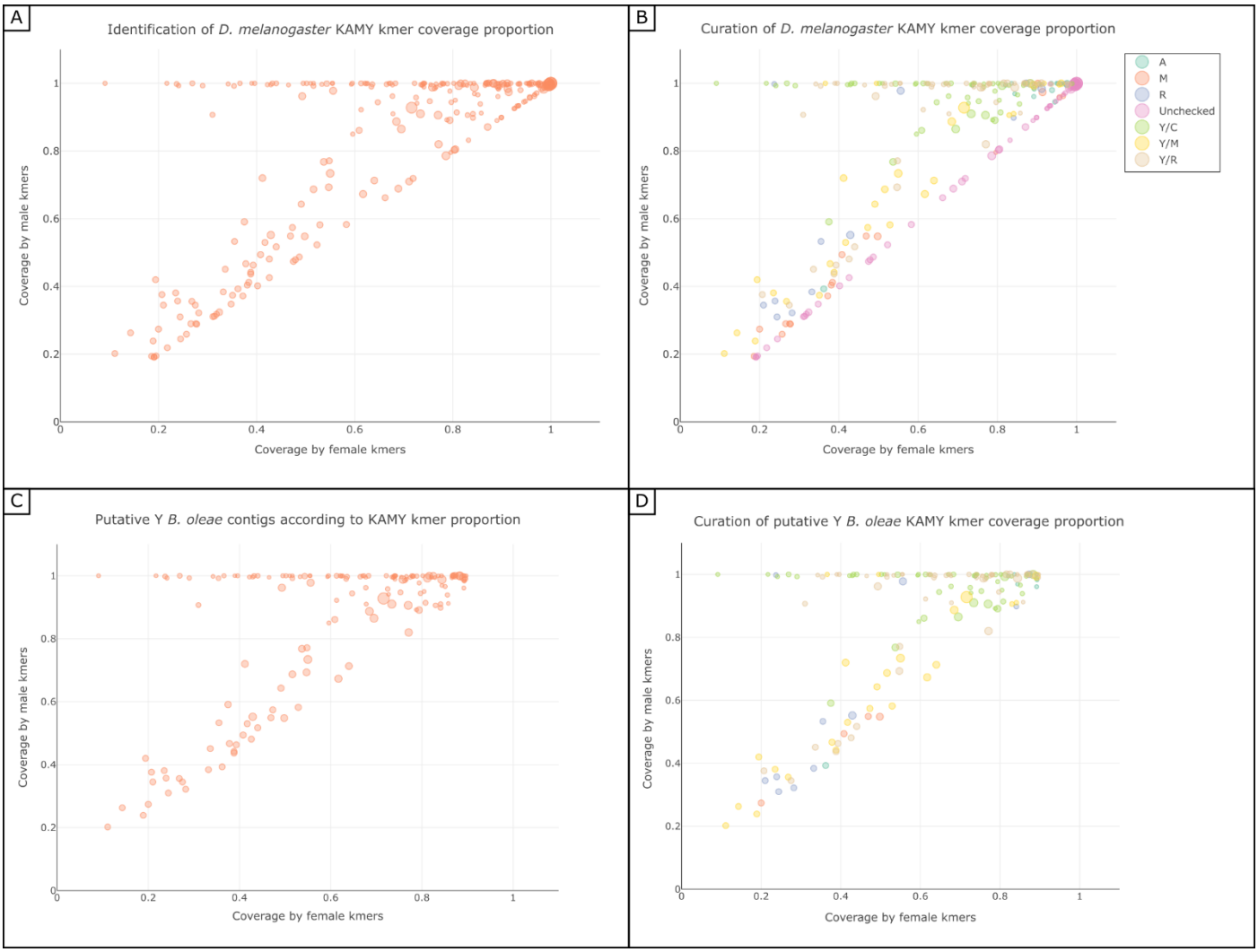

Figure 14 Scatterplot of male VS female kmer coverage of D. melanogaster scaffoldss as identified by KAMY. A) KAMY kmer coverage of total D. melanogaster scaffolds. B) Curation of total D. melanogaster scaffolds over KAMY kmer coverage. C) Kmer coverage of putative Y D. melanogaster scaffolds.

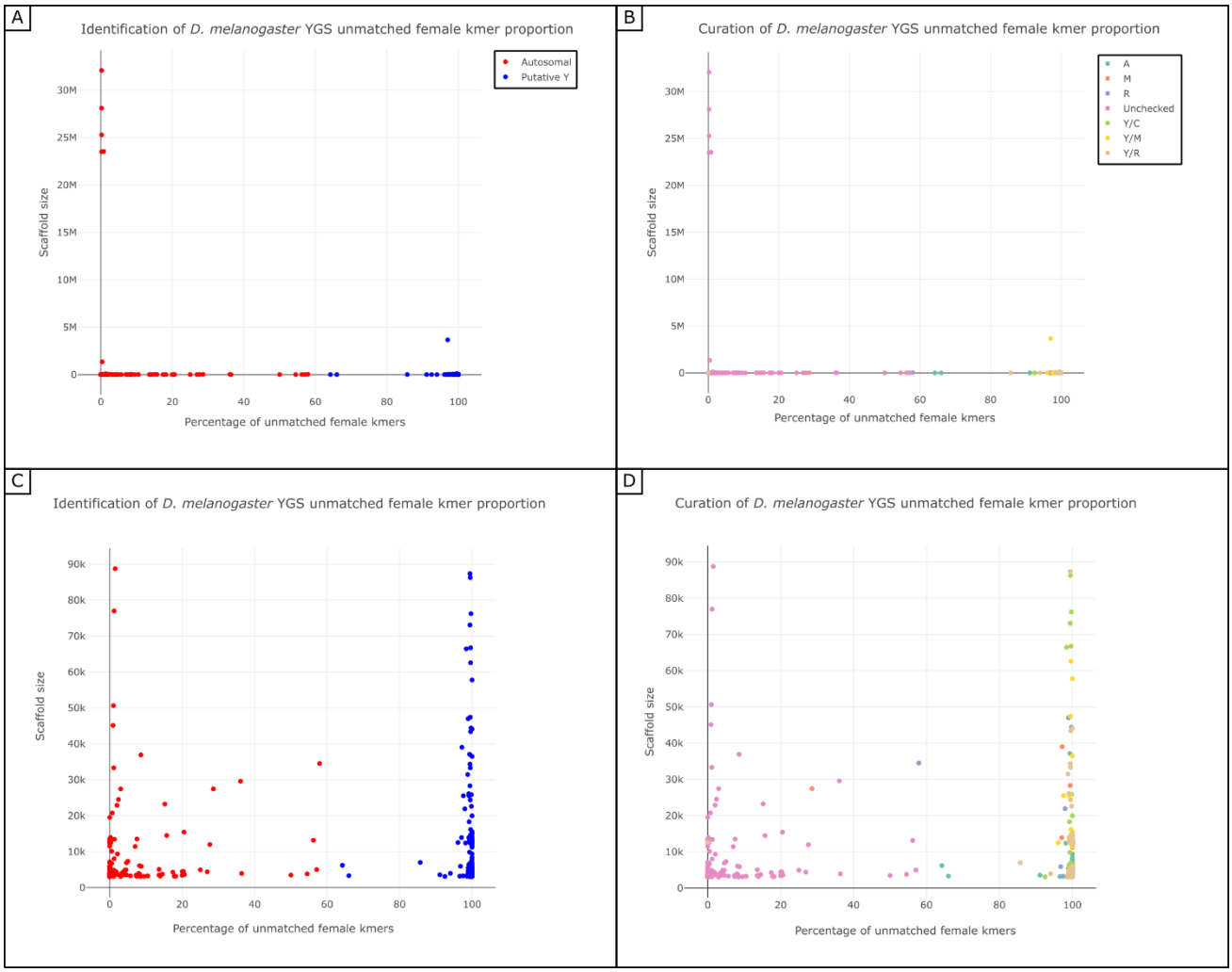

Figure 15 Scatterplot of the percentage of unmatched female kmers as identified by YGS over the size of D. melanogaster scaffolds. A) Percentage of unmatched female kmers of total D. melanogaster scaffolds. B) Curation of total D. melanogaster scaffolds over percentage of unmatched female kmers. C) Percentage 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.

Table 1 Manual curation of method outputs over C. capitata Y chromosome.

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	60	60	60	41	60
Y/R	35	37	36	13	37
Y/M	7	14	10	6	14
A	0	132	0	1	132
M	1	5	3	1	7
R	1	0	0	0	1
Total Y	102	111	106	60	111
Total False Positives	2	137	3	2	140
Total Predictions	104	248	109	62	251
Y-Specificity	0.98	0.45	0.97	0.97	
Y-Sensitivity	0.92	1.00	0.95	0.54	

Table 2 Percentage of C. capitata manually curated contigs identified by every method over the total number of curations in the respective category.

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	100.00%	100.00%	100.00%	68.33%	60
Y/R	94.59%	100.00%	97.30%	35.14%	37
Y/M	50.00%	100.00%	71.43%	42.86%	14
A	0.00%	100.00%	0.00%	0.76%	132
M	14.29%	71.43%	42.86%	14.29%	7
R	100.00%	0.00%	0.00%	0.00%	1
Total Y	91.89%	100.00%	95.50%	54.05%	111
Total False Positives	1.43%	97.86%	2.14%	1.43%	140
Total Predictions	41.43%	98.80%	43.43%	24.70%	251
Y-Specificity	98.08%	44.76%	97.25%	96.77%	
Y-Sensitivity	91.89%	100.00%	95.50%	54.05%	

Table 3 Size (bp) of curated C. capitata Y contigs identified by every method and their respective Y category size.

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	15,257,287	15,257,287	15,257,287	12,066,473	15,257,287
Y/R	8,270,098	9,240,983	8,355,291	4,245,375	9,240,983
Y/M	819,727	2,318,182	1,107,613	707,640	2,318,182
Total Y size	24,347,112	26,816,452	24,720,191	17,019,488	26,816,452

Table 4 Number of PCR validated C. capitata Y contigs identified by every method, the respective Y category and total size (bp) of PCR validated Y.

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	6	6	6	6	6
Y/R	4	4	4	3	4
Y/M	1	2	1	0	2
Total PCR validated Y	11	12	11	9	12
PCR Validated Size	6,175,675	7,380,426	6,175,675	4,454,309	7,380,426

Table 5 Manual curation of method outputs over B. oleae Y chromosome.

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	38	33	4	16	40
Y/R	20	21	3	17	26
Y/M	6	10	6	2	17
C	87	6	344	0	344
A	84	67	142	8	268
M	5	8	77	4	88
R	6	3	3	1	9
Total Y	64	64	13	35	83
Total False Positives	182	84	566	13	709
Total Predictions	246	148	579	48	792
Y-Specificity	0.26	0.43	0.02	0.73	
Y-Sensitivity	0.77	0.77	0.16	0.42	

Table 6 Percentage of B. oleae manually curated contigs identified by every method over the total number of curations in the respective category.

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	95.00%	82.50%	10.00%	40.00%	40
Y/R	76.92%	80.77%	11.54%	65.38%	26
Y/M	35.29%	58.82%	35.29%	11.76%	17
C	25.29%	1.74%	100.00%	0.00%	344
A	31.34%	25.00%	52.99%	2.99%	268
M	5.68%	9.09%	87.50%	4.55%	88
R	66.67%	33.33%	33.33%	11.11%	9
Total Y	77.11%	77.11%	15.66%	42.17%	83
Total False Positives	25.67%	11.85%	79.83%	1.83%	709
Total Predictions	31.06%	18.69%	73.11%	6.06%	792
Y-Specificity	26.02%	43.24%	2.25%	72.92%	
Y-Sensitivity	77.11%	77.11%	15.66%	42.17%	

Table 7 Size of curated B. oleae Y contigs identified by every method and their respective Y category size (bp).

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	1,892,498	1,886,310	17,241	950,723	1,911,708
Y/R	603,111	678,062	13329	569917	716,311
Y/M	82,419	1,242,532	61,557	42,835	1,308,105
Total Y size	2,578,028	3,806,904	92,127	1,563,475	3,936,124

Table 8 Number of PCR validated B. oleae Y contigs identified by every method, the respective Y category and total size (bp) of PCR validated Y.

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	21	21	0	10	21
Y/R	13	14	0	12	14
Y/M	0	1	0	0	1
Total PCR validated Y	34	36	0	22	36
PCR validated Y size	2,301,505	2,377,015	0	1,392,402	2,377,015

Table 9 Number of PCR validate B. oleae false positive contigs identified by every method, the respective Y category and total size (bp) of PCR validated false positives.

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	7	7	0	3	7
Y/R	1	0	0	0	1
Y/M	2	2	0	1	2
Total PCR validated false positives	10	9	0	4	10
PCR validated false positive size	90,979	70,295	0	17,213	90,979

Table 10 Manual curation of method outputs over D. melanogaster Y chromosome.

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	42	52	52	36	54
Y/R	40	53	54	32	59
Y/M	10	24	27	15	28
A	12	6	28	6	29
M	2	4	16	2	17
R	4	12	13	9	15
Total Y	92	129	133	83	141
Total False Positives	18	22	57	17	61
Total Predictions	110	151	190	100	202
Y-Specificity	0.84	0.85	0.70	0.83	
Y-Sensitivity	0.65	0.91	0.94	0.59	

Table 11 Percentage of D. melanogaster manually curated contigs identified by every method over the total number of curations in the respective category.

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	77.78%	96.30%	96.30%	66.67%	54
Y/R	67.80%	89.83%	91.53%	54.24%	59
Y/M	35.71%	85.71%	96.43%	53.57%	28
A	41.38%	20.69%	96.55%	20.69%	29
M	11.76%	23.53%	94.12%	11.76%	17
R	26.67%	80.00%	86.67%	60.00%	15
Total Y	65.25%	91.49%	94.33%	58.87%	141
Total False Positives	29.51%	36.07%	93.44%	27.87%	61
Total Predictions	54.46%	74.75%	94.06%	49.50%	202
Y-Specificity	83.64%	85.43%	70.00%	83.00%	
Y-Sensitivity	65.25%	91.49%	94.33%	58.87%	

Table 12 Size of curated D. melanogaster Y contigs identified by every method and their respective Y category size (bp).

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	596,786	635,187	634,482	276,562	641,918
Y/R	449,913	587,777	609,288	255,251	648,024
Y/M	3,871,238	4,125,863	4,168,806	247,857	4,174,087
Total Y size	4,917,937	5,348,827	5,412,576	779,670	5,464,029

Table 13 Number of Y-annotated D. melanogaster scaffolds identified by every method, their respective curation and total size (bp).

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	4	6	6	1	6
Y/R	15	18	18	10	19
Y/M	2	7	7	3	7
M	1	4	3	3	4
R	1	0	1	0	1
Total Y	23	35	35	17	37
Total Y Size	4,128,649	4,224,498	4,256,281	257,078	4,294,999

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

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y/C	0	0	0	0	0
Y/R	1	0	1	1	1
Y/M	1	1	1	0	1
R	0	1	1	1	1
A	1	1	1	1	1
Total X;Y	3	3	4	3	4
Total X;Y Size	55,093	56,003	60,556	13,145	60,556

Table 15 Size (bp) of annotated D. melanogaster contigs identified by every method as putative Y and their respective chromosomal annotation.

Curation	R-CQ	KAMY	YGS	CQ	TOTAL
Y	4,128,649	4,224,498	4,256,281	257,078	4,294,999
X;Y	55,093	56,003	60,556	13,145	60,556
3CEN	235,313	247,956	260,226	59,691	268,716
2CEN	45,796	102,355	129,361	90,207	129,361
Unknown	675,712	1,005,030	1,279,821	590,167	1,361,347
Total size	5,140,563	5,635,842	5,986,245	1,010,288	6,114,979