Briefings in Bioinformatics, 2025, **26(5)**, bbaf507
https://doi.org/10.1093/bib/bbaf507
Problem Solving Protocol

OligoY pipeline for full Y chromosome painting

Isabela Almeida [1,2,3,*], Henry A.B. Bruno [1,2], Eduardo Guimarães Dupim [1], Mara Maria L. Santana Pinheiro [1], Antonio Bernardo Carvalho [4], Maria D. Vibranovski [1,2,*]

[1]Department of Genetics and Evolutionary Biology, Institute of Biosciences, University of São Paulo, São Paulo, 05508-090 SP, Brazil
[2]Interunit Graduate Program in Bioinformatics, University of São Paulo, São Paulo, 05508-000 SP, Brazil
[3]Cancer Program, QIMR Berghofer, Brisbane, QLD 4006, Australia
[4]Department of Genetics, Institute of Biology, Federal University of Rio de Janeiro, Rio de Janeiro, 21941-902 RJ, Brazil

*Corresponding authors. Isabela Almeida, E-mail: mb.isabela42@gmail.com; Maria D. Vibranovski, E-mail: mdv@ib.usp.br

Abstract

The Y chromosome's unique structure poses challenges for cytogenetic studies, especially in designing probes for FISH Oligopaint labeling experiments (Fluorescence *in situ* hybridization). The standard protocol for designing probes for these experiments discards repetitive sequences to avoid off-target hybridization. Given the highly repetitive nature of the Y chromosome, assemblies often remain fragmented, leaving significant regions incompletely sequenced, unprobed, and poorly understood. Among these, the remaining nonrepetitive sequences are usually insufficient to design probes and efficiently perform FISH Oligopaint assays, since they do not cover most regions of the chromosome. This limitation hinders comprehensive cytogenetic studies, which are crucial not only for understanding the Y chromosome's role in genetics but also for broader applications in evolutionary biology, medicine, and conservation. Here, we introduce a new computational pipeline to design full chromosome fluorescent labeling probes for the Y chromosome of any species of interest. Based on open-source tools, the OligoY pipeline increases the amount of contigs assigned to the Y chromosome from the reference genome assembly, and effectively uses repetitive sequences unique to the target chromosome to design probes. Throughout its steps, the pipeline gives the user the autonomy to choose parameters, maximizing the overall efficiency of cytogenetic experiments. After extensive *in silico* and *in situ* testing and validation with the human and *Drosophila melanogaster* genomes, we show for the first time a pipeline for FISH Oligopaint probe design that significantly increases previous Y chromosome staining with no off-target signal.

Keywords: cytogenetics; heterochromatin; FISH Oligopaint; Y chromosome

Introduction

The Y chromosome's highly heterochromatic nature makes it prone to the accumulation of repetitive DNA [1, 2], with assembly algorithms tending to combine repetitive reads into a single position. This leads to a loss of information on how these sequences are truly distributed in the chromosome [3] and its poor assembly in genome projects [3–5]. Frequently, a large proportion of the Y chromosome's sequenced regions remain as a collection of small, unmapped scaffolds, sometimes referred to as "chromosome U" (Fig. 1A) [6, 7]. There were specific efforts towards improving the assembly of Y chromosome heterochromatic regions [6, 8]. More recently, long-read technologies facilitated the sequencing and assembly of repeat-rich areas [9], enabling telomere-to-telomere assemblies for all chromosomes, including the Y, for humans and a few other organisms [10–12].

The Y chromosome contains several conserved satellite-specific sequence motifs [13], leading to their use as probes in cytological Y chromosome studies [14]. To optimize microscopy assays, these probes are designed to include motifs with a minimal repetition (e.g. tandem satellite regions), while avoiding repeats dispersed across the target chromosome [15]. However, satellite probes are limited to a few regions and are often shared with other chromosomes, making them unsuitable for comprehensive or specific targeting of the Y chromosome [14]. Chromosome painting techniques were developed to overcome these limitations: initially, individual chromosomes were isolated and used to create fluorescent DNA libraries [16, 17]. However, fluorescence labeling from microdissection or chromosome sorting is prone to fail to cover repetitive sequences of Y chromosomes, as suppression hybridization clocks repeats, leaving large regions unlabeled.

Nowadays, many tools can use whole-genome sequences to create chromosome-specific single-copy probes, i.e. FISH Oligopaint probes, which combined hybridize along the entire target chromosome [18]. These approaches work elegantly for all chromosomes, except for the Y [19–21], because they rely on its few remaining sparsely mapped nonrepetitive sequences to avoid off-target hybridization. Therefore, using such tools to design probes for the Y chromosome in a conventional manner typically results in low probe coverage and, consequently, insufficient fluorescence signals (Fig. 1B). Complete visualization of the Y chromosome would enable cytogeneticists to explore its interactions with other chromosomes and proteins across various cellular stages, enhancing our understanding of genomic

Received: March 5, 2025. **Revised:** August 21, 2025. **Accepted:** August 28, 2025

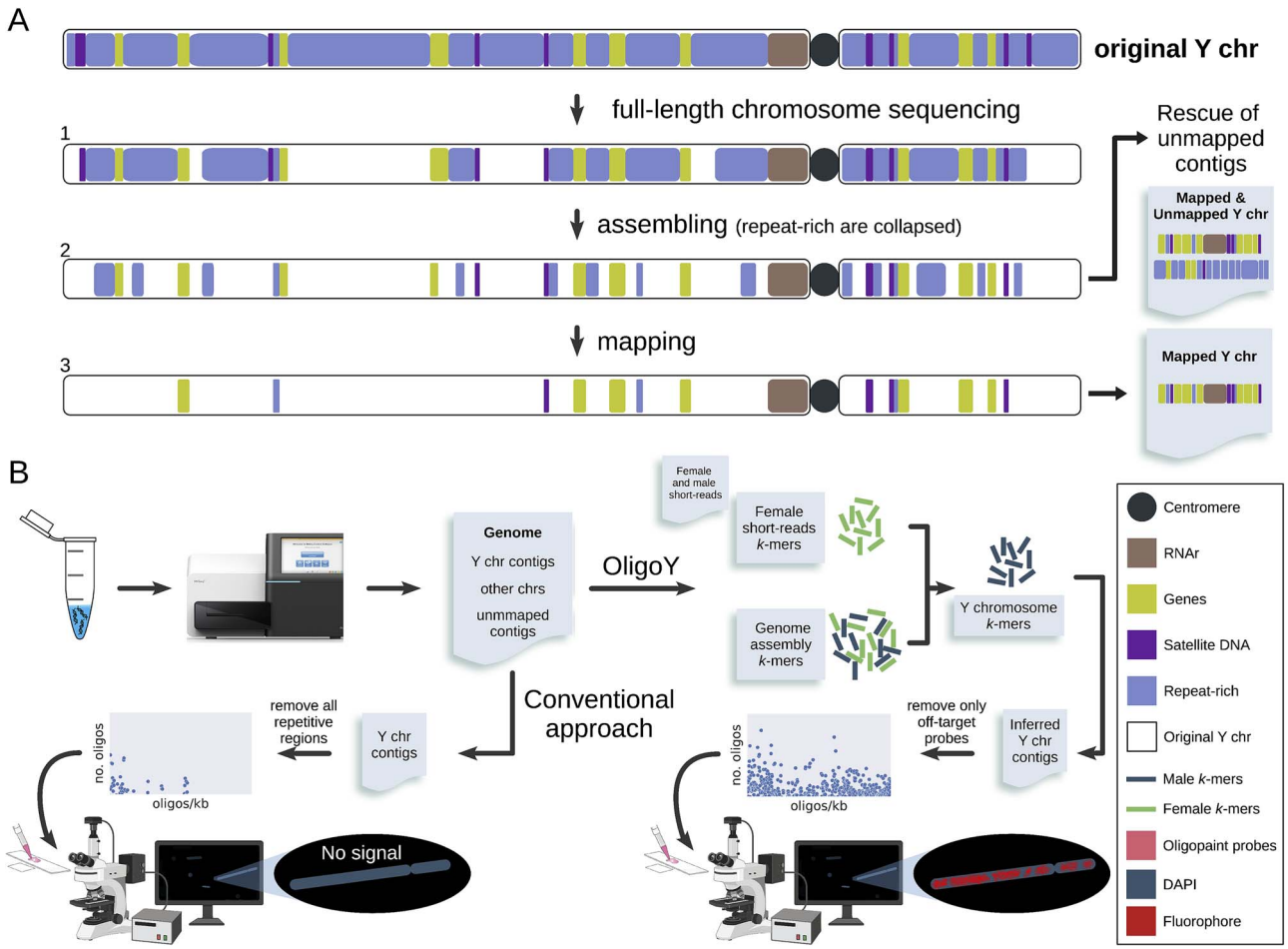

Figure 1. From the Y chromosome sequencing to cytogenetic assays. (A) Y chromosome sequencing, assembly, and mapping steps. Shows loss of sequencing information due to its haploid and repetitive nature at the following stages: (i) sequencing, depending on the technique used, (ii) assembly, as repeat-rich reads tend to be collapsed at a single position by assembly algorithms, and (iii) mapping, where only a few sequences, such as known genes, are correctly assigned to the Y chromosome. (B) Stepwise comparison of conventional FISH Oligopaint probe design and assay with OligoY. In conventional strategies, repetitive sequences are masked to avoid off-target signals, resulting in less probes and consequently less targeted areas. In contrast, OligoY produces higher probe densities across both repetitive and nonrepetitive regions, leading to stronger and more consistent FISH signals.

architecture and gene regulation. This comprehensive mapping would facilitate more in-depth insights into the Y chromosome's potential regulatory roles on other chromosomes' expression, a subject of broad scientific interest given its implications in developmental biology, genetic disorders, and evolutionary adaptations [22, 23].

To maximize the coverage of the Y chromosome in FISH Oligopaint assays, we developed a novel computational pipeline, named OligoY (Fig. 1B). This pipeline integrates open-source tools and represents a conceptual shift in the need for discarding repetitive sequences. First, OligoY uses the YGS method (Y chromosome Genome Scan) [7] to recover all male-specific sequences, enriching the amount of known Y chromosome contigs by including complex regions that are typically unmapped in standard assemblies. Then, OligoY uses all the retrieved sequences to design candidate FISH Oligopaint probes with OligoMiner [18]. In an unprecedented way, we present a new approach for one of OligoMiner's probe filtering options. This new strategy allows OligoY to remove only probes that effectively target non-Y chromosome sequences, revoking the need to dismiss repeat-rich regions completely while eliminating the risk of off-target hybridization. This innovative approach substantially increases probe density and hybridization specificity across the Y chromosome, as demonstrated in our tests with both human and *Drosophila melanogaster*. Ultimately, OligoY has led to the first-ever successful FISH Oligopaint assays with the Y chromosome.

Results and discussion

Overview of the OligoY pipeline

The OligoY pipeline has two major steps: (i) the inference of Y chromosome contigs and (ii) the FISH Oligopaint probe design (Fig. 2A). First, OligoY uses the YGS method [7] to identify Y chromosome sequences in the provided genome assembly, assigning more contigs to it and enhancing its nucleotide content (Fig. 2B). These sequences are then used to design and comprehensively filter candidate oligopaint probes with our novel approach of OligoMiner [18]: after the initial design, candidate probes are aligned against a YGS-Y-excluded genome assembly and subjected to multiple layers of filtering, ultimately eliminating any off-target sequences (Fig. 2C and D). Users can run the entire pipeline using three BASH scripts (i.e. target-inference:YGS.sh, contaminant-removal_BlobTools.sh, and probe-design_OligoMiner.sh; Supplementary File S3 (Figs S1 and S2)).

To infer Y chromosome sequences, OligoY requires three input files: a male FASTA genome assembly of the target species and

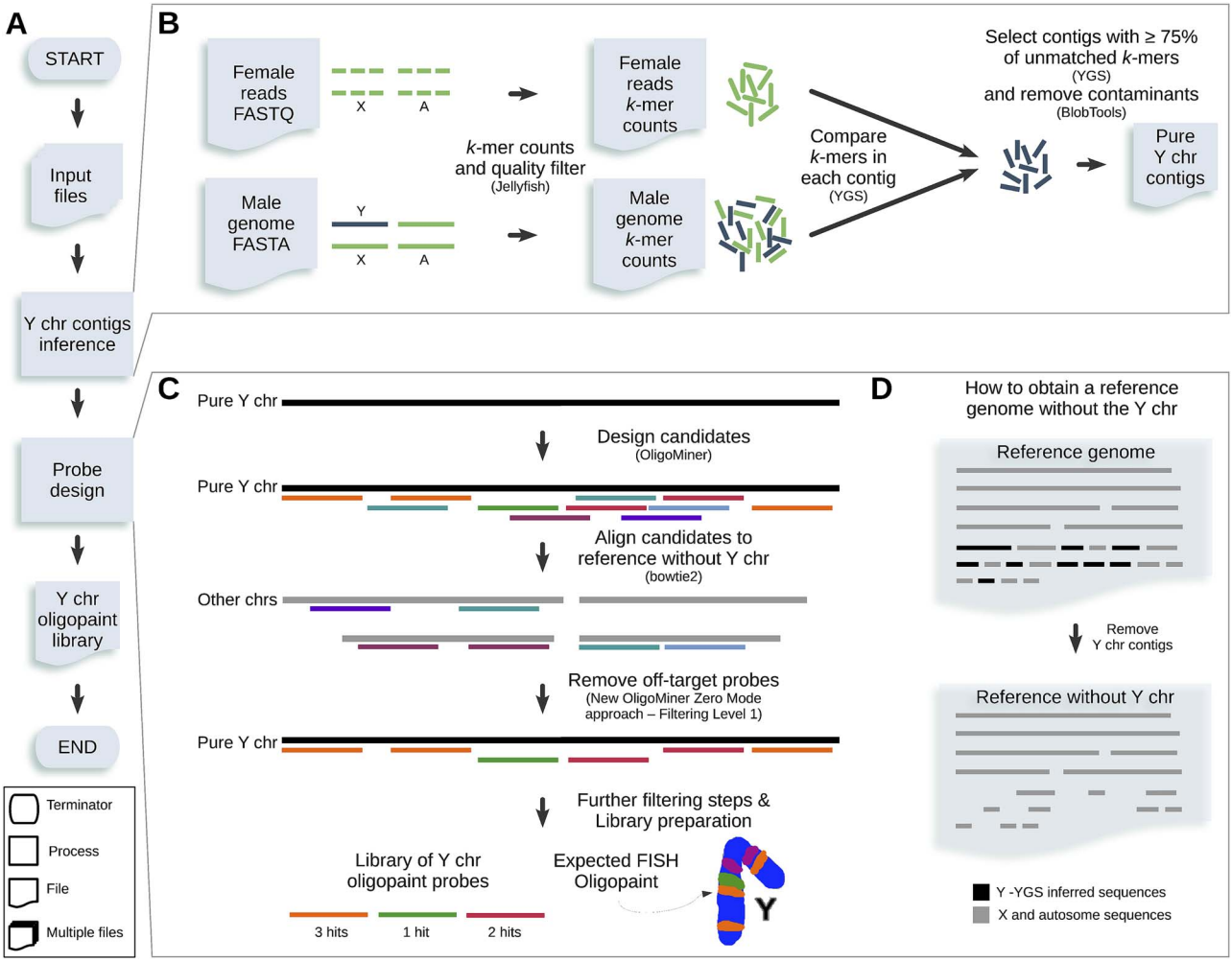

Figure 2. The OligoY pipeline. (A) Major steps and (B) Y chromosome contigs inference step: YGS compares *k*-mers from the male genome assembly and from the female sequencing reads. User-defined thresholds of valid single-copy *k*-mers (VSCK) unmatched by female reads (PVSCUK) are used to infer Y chromosome sequences (we used 75% PVSCUK and 20 VSCK). (C) Probe design step: candidate probes are designed and filtered against a reference genome without the Y chromosome sequences; with our novel OligoMiner's ZM approach, OligoY pipeline removes only the probes that effectively target other chromosomes, producing a library with both single and multiple hits, all exclusively Y probes. These are further filtered to produce the library (Supplementary File S3 (Figs S1 and S2)). (D) To obtain a reference genome file without the YGS Y chromosome inferred sequences, we remove these from the original FASTA file.

its female and male FASTQ reads. These files are preprocessed independently using Jellyfish [24] to create high-quality trimmed hash tables (Fig. 2B). Next, the pipeline runs the different subroutines of YGS to filter rare *k*-mers from the input hash tables, keeping a vector with the repetitive genome *k*-mers (lower-count >1) and three others with all k-mers from the genome and the female and male reads hash tables. All of this information is then compared to create internal vectors for each possible combination. At the end of its analysis, YGS generates a report with a series of metrics for each genome sequence, including the proportion of valid single-copy *k*-mers (VSCK) and the proportion of valid *k*-mers unmatched by female reads (PVSCUK). In this context, **valid** *k*-mers are those confirmed to be present in both the genome and male reads, which helps in eliminating some potential contaminants and assembly artifacts from the genome (Supplementary File S1). The VSCK and PVSCUK are important metrics used by OligoY to ascertain which sequences belong to the Y chromosome when they fall within a user-defined range (Fig. 2B). These steps are comprised within the script target-inference:YGS.sh with the key output being the Y chromosome FASTA file. The YGS method implementation is described in the Supplementary File S1. Following that, the OligoY pipeline includes a contaminant analysis step (script contaminant-removal_BlobTools.sh) using BlobTools [25, 26], which avoids inferring any male-specific contaminated sequence to the Y chromosome and designing oligopaint probes to it. To run this script, the user needs to set a path to species-specific databases and input the reference genome and the Y chromosome FASTA files, with the key output being the pure Y chromosome sequences.

Then, OligoY uses OligoMiner to design overlapping candidate probes from the input concatenated pure Y chromosome inferred sequences (Fig. 2C; script probe-design_OligoMiner). Employing our new OligoMiner approach, OligoY uses bowtie2 [27] to align these candidate probes against the input genome assembly without the YGS-Y chromosome inferred sequences (Fig. 2D). Next, OligoMiner's filter of predicted target number of alignments is used with the zero mode (ZM) option, removing all the probes that effectively align outside the Y (Filtering level 1). This is achieved by retaining only probes that aligned zero times in the

provided genome—which is a genome assembly without the YGS inferred Y chromosome contigs. Therefore, while only specific Y chromosome probes are kept, this OligoY strategy captures probes whether they appear once on the chromosome or repeatedly. Due to the complexity of this chromosome, additional OligoMiner filtering steps are also used, i.e. the elimination of probes that have an abundance of *k*-mers in other chromosomes (Filtering level 2) and of those that are likely to form secondary structures (Filtering level 3). In addition to that, OligoY has three other filtering routines to guarantee the efficiency of FISH Oligopaint assays. We design overlapping probes so that all potential probe regions can be identified; however, the overlapping of synthesized probes is detrimental to *in situ* experiments, particularly to the hybridization procedure. To address this, the OligoY pipeline includes an in-house Python program to remove remaining overlapping probes (noOverlap.py, Filtering level 4). Probes with multiple hits only need to be synthesized once, due to the library amplification step. Therefore, we implemented an in-house Python program to retain only the first occurrence of a probe (nonRedundant.py, Filtering level 5). The multiple copies of each probe are only relevant for probe density estimations, which can be calculated with the Filtering level 4 output file. Finally, OligoY selects probes with no alignments to female reads (FASTQ converted to FASTA with seqtk by Li—available at https://github.com/lh3/seqtk). This helps to eliminate any possibility of probes in the final library aligning outside the Y chromosome (Filtering level 6).

Pipeline development and validation

We performed a literature review of state-of-the-art tools for FISH Oligopaint probe designing methods. We found OligoMiner [18], iFISH [28], ProbeDealer [29], Chorus2 [30], and PaintSHOP [31]. From these, OligoMiner is a pioneer in facilitating the design of genome-wide oligopaint probes, allowing users to control a broad range of parameters, from probe design to extensive filtering options. The subsequent tools enhanced user-accessibility and provided new features. However, they lost some of the flexibility and power in OligoMiner, which was key for us to develop a new approach and maximize probe design for such a complex chromosome as the Y. Additionally, these tools generally do not outperform OligoMiner in probe density, with many having incorporated OligoMiner's parameters thresholds (stringent, balance, and coverage) in their interactive web applications [18, 28, 29, 31]. Altogether, none of these tools have shown to efficiently design probes for the Y chromosome.

To develop and evaluate the efficiency of our pipeline, we selected the model organism *D. melanogaster*, for its highly repetitive Y chromosome and incomplete genome assemblies that are likely to mimic the situations faced by future pipeline users. Namely, we grouped the draft genome assemblies used here as the most fragmented (Illumina [32], Sanger [6], and Release 6 [33]) and the more contiguous ones [Chang and Larracuente (CL) and the Falcon genomes [23, 34]; Supplementary File S4 (Table S1)]. We also assess OligoY's performance by comparing the number of probes it generated for the complete T2T-CHM13 (herein T2T) Y chromosome [11, 12] with the number of probes we could design with the standard use of OligoMiner. For a complete description of the materials and methods used to develop and validate the pipeline, see the Supplementary File S2.

Our OligoY technique significantly outperformed the default OligoMiner analysis for the Y chromosome, generating a markedly higher number of probes. Using OligoMiner alone, for *D. melanogaster* we designed from 587 to 1757 probes across varying stringency levels, corresponding to lengths and experimental conditions that range from more rigorous to more permissive settings [18] (stringent, balance, and coverage). However, with OligoY we were able to design from 6428 to 19 007 probes under comparable conditions using the YGS-derived sequences from the same input file as [18], i.e. Release 6 (Fig. 3A; Supplementary File S4 (Tables S2 and S3)). For the human Y chromosome, OligoY generated over double the number of probes designed by OligoMiner alone. Here, to analyze the components of our pipeline that contributed to the enhancement of probe yield, our *in silico* tests were structured to evaluate the outcomes of each major step within the OligoY approach: Y chromosome contig inference and probe design.

In silico tests show an enhancement of nucleotide content in the inferred Y chromosome

In genome analysis, the choice of *k*-mer length is a balance between capturing unique sequence information and avoiding repetitive elements. Our tests on *D. melanogaster*'s Y chromosome contig inference demonstrated a consistent pattern across the most and less fragmented input genomes for both *k*-mer values employed here (15 or 18). We consistently observed more contigs being inferred for each individual Y chromosome group of sequences for Illumina (I2), Sanger (W3), and Release 6 (R6) assemblies, which had the smallest nucleotide content among the tested ones. In fact, Illumina's estimated Y chromosomal sequences were found to be as low as ≈1.5 megabases (Mb). Conversely, when using Falcon (F1) and CL input genomes, the least fragmented genomes, we identified up to 13.8 Mb of Y chromosome sequences (Fig. 3A; Supplementary File S4 (Tables S4 and S5)). Notably, for all tests conducted with *D. melanogaster*, *k*-mer value 18 did not show significant advantages over *k*-mer value 15. Hence, we support Carvalho and Clark's recommendation to use *k*-mer 15 for smaller genomes and 18 for larger ones [7]. However, OligoY users should consider that this choice is influenced by several factors, including genome size, genome complexity, repeat content, and sequencing technology. We also removed the few contaminant sequences identified with BlobTools from the YGS-Y chromosome inferred data (Supplementary File S4 (Tables S6-S8)). Only sequences classified under Arthropoda or Eukaryota taxa were retained, ensuring that oligopaint probes would not be unnecessarily designed.

Over half of the sequences in Release 6 are unmapped contigs (Supplementary File S4 (Table S9)), accounting for 3.05 Mb of genetic material. The Y chromosome in this assembly was made up of a single scaffold and 199 contig sequences, amassing a total of 4.53 Mb [33]. While using this assembly as in input, YGS implementation in our pipeline facilitated the inference of numerous unmapped contigs as part of the Y chromosome, thereby increasing its known sequences to a minimum of 769 contigs (6.57 Mb, using a *k*-mer of 15) and a maximum of 821 contigs (6.65 Mb, using a *k*-mer of 18; Supplementary File S4 (Table S5)).

Importantly, when running YGS with CL's assembly [23], 1.15 Mb across 16 contigs of their Y chromosome mapped sequences (14.6 Mb across 106 contigs in total) were not inferred as such with either *k*-mer value 15 or 18. One of these was identified as a contamination, and the remaining did not meet our cut-off thresholds (VSCK ≥ 20 and/or PVSCUK ≥ 75%). This indicates that by enforcing a high VSCK threshold, some of the repeat-rich Y chromosome contigs are lost, even though this may be rectified by using a larger *k*-mer value. Additionally, our YGS analysis selected two sequences, each with at least 97.3% of valid *k*-mers (i.e. PVSCUK ≥ 97.3%), that were not present among the female reads. While CL relied on an extensively detailed

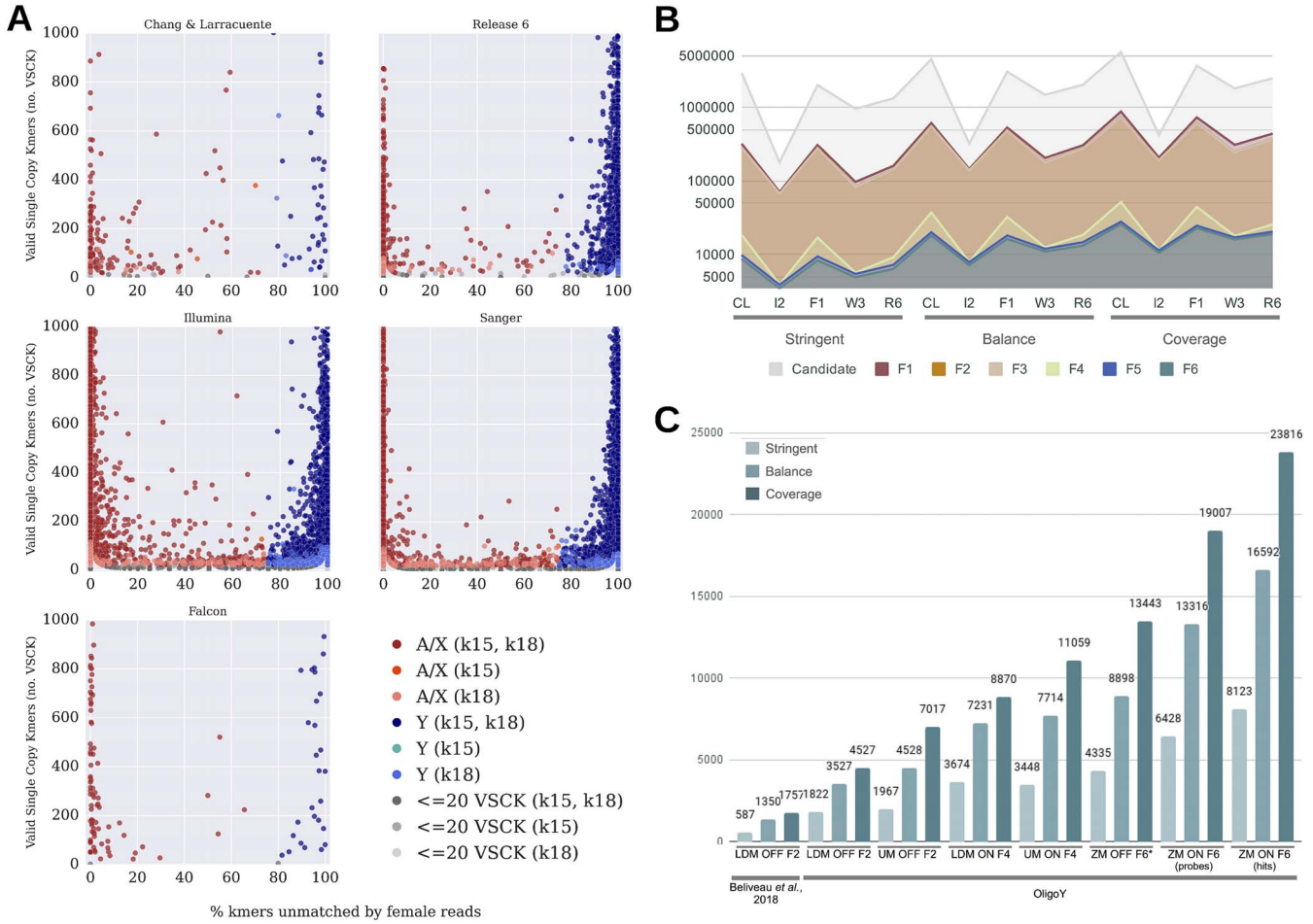

Figure 3. Parameter evaluation. (A) YGS chromosome sequence inference with *k*-mer values 15 and 18; (B) loss of probes throughout OligoY filtering steps (log scale, F1–F6 represent Filtering Levels 1–6); and (C) number of probes designed for the R6 Y chromosome in OligoMiner publication [18] (first three bars to the left) and by OligoY (with YGS-derived sequences). *For ZM, when overlapping probes were disabled (OFF), the resulting F6 probes were not subjected to F4, which is the filtering level designed to remove overlapping probes.

approach with a series of manual curations to enrich the known sequences for the Y chromosome of *D. melanogaster* [23], OligoY users can achieve similar results with significantly less effort and meticulousness because we have incorporated the YGS method into it.

Designing probes with a density of over 1 probe/kb

When designing Y chromosome libraries, a higher probe per chromosome kb density increases the library efficiency. According to Beliveau *et al.* [18], the smaller the target region, the higher the probe per kilobase (kb) density. For the Y chromosome, its smaller size combined with incomplete sequencing and assembly are playing factors that require the highest probe density available. Considering this, the success of OligoY in achieving a much denser coverage of the target chromosome is intricately connected to four key features: (i) to design probes from a YGS-derived inputs, i.e. more extensive nucleotide sequence compared with the initial inputs; (ii) extensive filtering steps to guarantee no contaminants and off-target probes are kept; (iii) designing candidate probes that meet the desired criteria with the overlap option, increasing the chances of retaining signals from regions that might otherwise be discarded; and (iv) allowing multiple hits in the target chromosome to be kept.

Previously, Beliveau *et al.* [18] designed Y chromosome probes for the Release 6 genome targeting different settings from strict to permissive, resulting in oligopools of 587 (stringent), 1350 (balance), and 1757 (coverage) probes. They used OligoMiner's linear discriminant analysis (LDA) mode and kmerFilter.py to remove probes with abundant *k*-mers. When solely changing the input target chromosomes to a YGS-derived one, we were able to design 1822, 3527, and 4527 for each of those corresponding parameter conditions (Supplementary File S4 (Tables S2, S3 and S8)). This shows that by implementing YGS in our pipeline and inputting a more extensive nucleotide sequence, we more than double the number of target probes that will compose the oligopaint library.

To enhance the likelihood of preserving signals from regions for which candidate probes were filtered out, we used the native overlapping option from OligoMiner. Comparing with Beliveau *et al.* [18] use of the LDA mode, we now designed from 3674 (stringent) to 8870 (coverage) probes with the YGS-derived Release 6 Y chromosome sequences (Supplementary File S4 (Tables S2 and S3)). Despite the LDA mode keeping an initial larger number of probes for simulating the number of hybridization in the reference genome, when applying further filtering steps most of the probes with more than one hit are removed. Using the same input and OligoMiner's unique mode (UM), our tests resulted in 3122, 7101, and 10 415 probes (stringent, balance, and coverage parameters), over 5× the number of probes designed by Beliveau *et al.* [18].

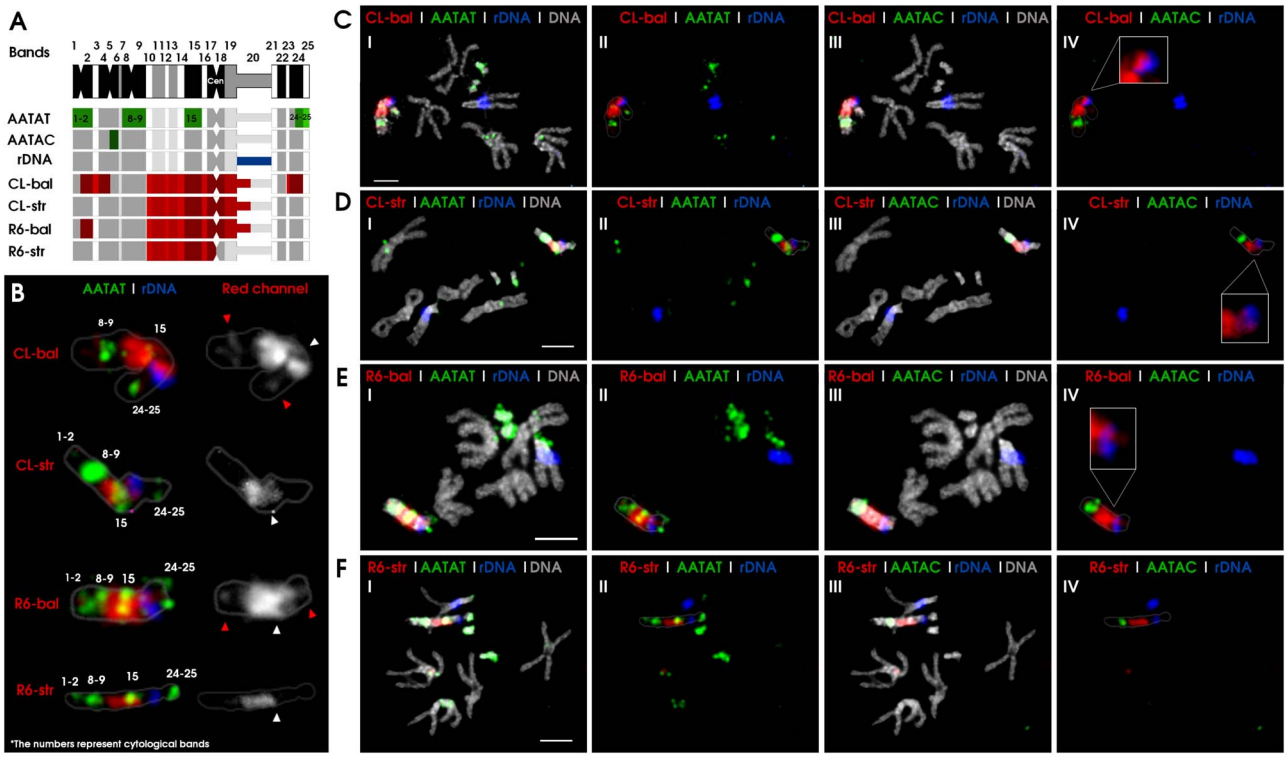

Figure 4. FISH Oligopaint on larval neuroblast chromosome spreads from *D. melanogaster* (ISO1 strain) for CL and R6 assemblies. (A) Schematic representation of the Y chromosome showing 25 cytological bands and the approximate cytological location of the satellite probes (AATAT and AATAC) [15], ribosomal DNA (rDNA), [23] and the oligopaint libraries (CL-balance, CL-stringent, R6-balance, and R6-stringent). Cen: centromere. (B) Comparison of the labeling patterns observed in all four oligopaint libraries on the Y chromosome. The AATAT and rDNA probe positions are reference points for contextualizing the oligopaint labeling. Oligopaint labels in the red channel. Red arrowheads show the extra labels of balance libraries (CL-baland R6-bal). The centromeres are indicated by white arrowheads. Images with FISH experiments for: (C) CL-balance library, (D) CL-stringent library, (E) R6-balance library, and (F) R6-stringent library. I and III show the labeling of three probes (oligopaint, satellite, and rDNA) combined with DAPI (DNA, 4′,6-diamidino-2-phenylindole fluorescent stain), while II and IV show only the three probes (colors specified on panel text). The zoom-in boxes in IV reveal the violet color formed by partial overlapping between oligopaints (red) and rDNA (blue). Note that oligopaints libraries (red) exclusively label the Y chromosome. CL-bal (Cy5: red), CL-str (Cy5: red), R6-bal (Cy5: red), R6-str (ATTO488: pseudocolor in red), AATAT (avidin-FITC: green), AATAC (avidin-FITC: green), rDNA (antiDIG-rhodamine: pseudocolor in blue), and DAPI (pseudocolor in gray). Bar, 2.5 mm.

With the feature in our pipeline that allows probes with multiple hits to be kept, we designed up to 26 323 probes with 48 497 hits on the Y chromosome (coverage and *k*-mer 18 parameters with YGS-derived CL's genome). The same parameters and input with the standard OligoMiner mode yielded only 11 593 (UM) and 9423 probes (LDA, Fig. 3C). For the YGS-derived Release 6 Y chromosome sequences, the same settings yielded 19 264 probes with 24 075 hits as compared to the UM 10 436 and LDA mode 8352 single-hit probes.

Regardless of the input genome and experimental conditions we tested, OligoY produces numerous candidate probes that undergo extensive filtering. Most of them are lost in the first three filtering steps, highlighting their effectiveness in ensuring that only non-off-target probes are kept in the final pool, regardless of having single or multiple hits on the Y chromosome (Fig. 3B; Supplementary File S4 (Table S2)). Moreover, while the final filtering steps remove just a few probes (Fig. 3B), they assure that (i) final probes are not overlapping sequences, which would negatively affect FISH assays, (ii) only one copy of each probe is kept, since candidates with multiple hits are designed from multiple matching target regions, and (iii) none of these probes align against the female reads, which could have been not assembled and yet abundant fragments.

Designing probes for the Y chromosome is particularly challenging due to its high proportion of extensive repetitive structures, such as palindromes and tandem repeats, which complicate both assembly and probe design. This difficulty is even more pronounced in species such as *D. melanogaster*, where strong sequencing biases inherent to current long-read technologies (e.g. Nanopore and PacBio) hinder the complete assembly of Y-linked genes [35]. In the human genome, the T2T Y chromosome represents the first complete assembly with no unmapped sequences [11, 12]. As a proof of concept, we selected this sequence to test whether our pipeline can integrate tools to optimize Oligopaint probe design, even when assemblies are complete. Considering the human Y chromosome is fully assembled, we did not need to run OligoY's Y chromosome inference step. Using this Y chromosome and the balance set of parameters, we compared OligoMiner's LDA mode performance with our OligoY approach (Supplementary File S4 (Table S2)). After filtering candidate probes up to OligoY's Filtering level 3, OligoMiner alone produced a total of 68 879 probes. With OligoY's probe design strategy, we achieved a total of 140 253 probes and 263 817 hits. Although the probe density with OligoMiner alone is presumably enough for observing fluorescent signals (1.1 probes/kb) [18, 36], this analysis highlights that OligoMiner on its own, without our novel ZM approach, is not enough to design higher probe coverage, with which we achieved double probe density (2.25 probes/kb). Additionally, by adding the YGS method to our pipeline, users do not need to rely on complete genome assemblies.

Drosophila melanogaster in situ tests

The performance of YGS-derived CL's and Release 6 oligopaint probe libraries' produced from the stringent and balanced settings was assessed using FISH Oligopaint assays (CL-stringent, CL-balance, R6-stringent, andR6-balance). More specifically, these libraries were designed from YGS-derived sequences commonly inferred with both tested *k*-mer values. Both inputs assemblies showed a probe density of at least 1 probe/kb across the tested settings (ranging from 1.2 to 3.6 probes/kb from stringent to coverage parameter settings; Supplementary File S4 (Tables S2 and S10)). Importantly, CL's YGS-Y chromosome inferred sequences were roughly double of Release 6 YGS-Y ones. However, the sets of probes designed from CL's YGS-Y achieved approximately twice the number of hits and were 30% larger than those of Release 6.

Despite the highly repetitive nature of the Y chromosome, all four oligopaint libraries demonstrated remarkable specificity, with no evidence of off-target hybridization in other chromosomes (Fig. 4C–F; Supplementary File S3 (Figs S3-S6) and Supplementary File S4 (Tables S11-S14)). Interestingly, adopting more stringent length and experimental conditions to design Y oligopaint probes with OligoY is not a requirement. In fact, the more relaxed parameters we examined maintained target specificity (Fig. 4C and E; Supplementary File S3 (Figs S3 and S5)). Our microscopy image analysis suggests a correlation between the number of oligos per library and fluorescence intensity (Supplementary File S4 (Tables S2 and S10)): CL libraries showed stronger fluorescence than the R6 libraries, while the balance libraries fluoresced more than the stringent ones. Moreover, the CL-balance library displayed two additional labels near the ends of each arm—between chromosome bands ∼2–4 and bands ∼24–25 on the long and short arms, respectively (Fig. 4A and B; Supplementary File S3 (Fig. S3)). Bands 2–4 labels were also present in the R6-balance library at a lower fluorescence intensity (Supplementary File S3 (Fig. S5)), and these findings support our suggestion that more permissive sets of length and experimental conditions can be used to design probes for the Y chromosome. Particularly to *D. melanogaster*, our findings indicate that the CL-balance library represents the optimal choice for Y chromosome labeling, offering a compelling balance between specificity and broader coverage.

Collectively, the libraries CL-balance, CL-stringent, R6-balance consistently tagged two areas on both sides of the Y chromosomal centromere (Fig. 4A–E) while R6-stringent exclusively hits a portion of the long arm (Fig. 4B and F; Supplementary File S3 (Fig. S6)). We used two known satellite probes positions as reference (AATAC and AATAT [15]) to demonstrate that OligoY probes target Y chromosome cytological bands 10–20, including the pericentromere and the rDNA region [23] (Fig. 4A). Our oligopaint libraries outperformed the tested satellite probes in both coverage and specificity: while the Y chromosome has been extensively and exclusively stained with our OligoY probes, AATAT satellite probe targets four much smaller chromosome sites, predominantly mapping to chromosome 4 and to a lesser extent targeting chromosomes 3 and X (Fig. 4C–F, I and II). Moreover, although AATAC specifically labels the Y chromosome, this is limited to the small region of band 6 on the long arm (Fig. 4C–F, III and IV).

Our OligoY Oligopaint libraries selectively target Y-linked rDNA, despite it occurring on both sex chromosomes in *D. melanogaster* [14]. This showcases our method's ability to retrieve Y-linked probes despite its sequences' repetitiveness. The target rDNA region was effectively observed from the overlap between the signals from a rDNA probe (blue channel) and the rDNA region from CL libraries (red channel), resulting in the emergence of a violet staining in the Y chromosome band 20 area (Fig. 4C-IV and D-IV).

Conclusion

We developed the OligoY pipeline to optimize FISH Oligopaint probe coverage on the Y chromosome, significantly hitting the target across all genome versions and modes tested. By integrating a new approach of the OligoMiner ZM method with the YGS method, OligoY enhances known nucleotide content and designs probes for both nonrepetitive and repetitive sequences, increasing target areas without any off-target signals. When validating on *D. melanogaster*, OligoY achieved precise targeting, significantly expanding coverage to previously sparsely characterized heterochromatic regions: it resolved cytogenetic gaps by staining additional bands (2–4 and 24–25) alongside specific regions like h6 (AATAC) and h1-2, 8-9, 15, 24 (AATAT). While testing our novel ZM approach alone on the complete T2T human Y chromosome, OligoY still produced twice the number of probes compared to the original OligoMiner approach.

OligoY's flexibility makes it a powerful tool for cytogenetic studies, extending its use to other challenging genomic regions, such as W chromosomes, neo-sex chromosomes, and heterochromatin, as well as poorly characterized sequences across diverse species. For chromosomes that typically lack the accumulation of repetitive sequences, e.g. homomorphic sex chromosomes [37], OligoY pipeline would likely not be necessary as OligoMiner itself is comprehensive enough to design probes for these regions. On the other hand, its novel ZM approach enables effective probe design for repetitive regions regardless of chromosomal type, allowing researchers to study condensed or less-characterized chromosomes. While our *in situ* validations were limited to larval metaphases due to the absence of Y polytene chromosomes in our experimental model, the potential for greater extension of our oligos coverage in less condensed chromosomal states should not be overlooked. Despite this, OligoY has established a robust foundation for FISH Oligopaint assays, providing a versatile, scalable solution for chromosome analysis.

Key Points

- OligoY is a computational pipeline that uses open-source tools to design FISH Oligopaint probes specifically for the Y chromosome.
- *In silico* and *in situ* analysis confirmed that OligoY probes effectively cover Y chromosome sequences and improve FISH assays.
- Unlike other probe design methods, OligoY includes Y chromosome-specific repetitive sequences without any off-targets.
- The method was validated across multiple datasets, demonstrating its applicability to a wide range of species with complex Y chromosomes.

Author contributions

I.A., A.B.C., and M.D.V. conceived the project. I.A. designed and directed the study, performed all dry-lab analyses, developed the

pipeline, critically reviewed the computational code, created and maintains the GitHub repository. H.A.B.B. implemented the FISH Oligopaint protocol in the laboratory, performed the majority of wet-lab experiments, and reviewed the computational code. E.G.D. developed the YGS script and I.A. compared its credibility to the original tool. M.M.L.S.P. produced male slides. A.B.C. provided genome assemblies. I.A. and M.D.V. interpreted the data and co-wrote the manuscript with input from all authors.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by: Fundação de Amparo à Pesquisa do Estado de São Paulo (M.D.V.: 2015/20844-4, 2023/14607-6, and 2013/08028-1; I.A.: 2019/14878-4; H.A.B.B.: 2019/10559-1 and 2023/13517-3), Wellcome Trust (A.B.C.: 207486/Z/17/Z), and Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (A.B.C.: CNE2018).

Data availability

The datasets that were used in this study are all publicly available, and details of how to access them can be found in the main text at the first mention of the dataset. An open-source BASH implementation of the OligoY is accessible at https://github.com/isabela42/OligoY.

References

1. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 2003;**423**:825–37. https://doi.org/10.1038/nature01722
2. Bachtrog D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet* 2013;**14**:113–24. https://doi.org/10.1038/nrg3366
3. Cechova M. Probably correct: rescuing repeats with short and long reads. *Genes* 2020;**12**:48. https://doi.org/10.3390/genes12010048
4. Nishibuchi G, Déjardin J. The molecular basis of the organization of repetitive DNA-containing constitutive heterochromatin in mammals. *Chromosome Res* 2017;**25**:77–87. https://doi.org/10.1007/s10577-016-9547-3
5. Tomaszkiewicz M, Medvedev P, Makova KD. Y and W chromosome assemblies: approaches and discoveries. *Trends Genet* 2017;**33**:266–82. https://doi.org/10.1016/j.tig.2017.01.008
6. Hoskins RA, Smith CD, Carlson JW. *et al.* Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. *Genome Biol* 2002;**3**:research0085.1. https://doi.org/10.1186/gb-2002-3-12-research0085
7. Carvalho AB, Clark AG. Efficient identification of Y chromosome sequences in the human and *Drosophila* genomes. *Genome Res* 2013;**23**:1894–907. https://doi.org/10.1101/gr.156034.113
8. Hoskins RA, Carlson JW, Kennedy C. *et al.* Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* 2007;**316**:1625–8. https://doi.org/10.1126/science.1139816
9. Debladis E, Llauro C, Carpentier M-C. *et al.* Detection of active transposable elements in *Arabidopsis thaliana* using oxford nanopore sequencing technology. *BMC Genomics* 2017;**18**:537. https://doi.org/10.1186/s12864-017-3753-z
10. Wang B, Yang X, Jia Y. *et al.* High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. *GPB* 2022;**20**:4–13. https://doi.org/10.1016/j.gpb.2021.08.003
11. Nurk S, Koren S, Rhie A. *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53. https://doi.org/10.1126/science.abj6987
12. Rhie A, Nurk S, Cechova M. *et al.* The complete sequence of a human Y chromosome. *Nature* 2023;**621**:344–54. https://doi.org/10.1038/s41586-023-06457-y
13. Thakur J, Packiaraj J, Henikoff S. Sequence, chromatin and evolution of satellite DNA. *Int J Mol Sci* 2021;**22**:4309. https://doi.org/10.3390/ijms22094309
14. Jagannathan M, Warsinger-Pepe N, Watase GJ. *et al.* Comparative analysis of satellite DNA in the *Drosophila melanogaster* species complex. *G3* 2017;**7**:693–704. https://doi.org/10.1534/g3.116.035352
15. Bonaccorsi S, Lohe A. Fine mapping of satellite DNA sequences along the Y chromosome of *Drosophila melanogaster*: relationships between satellite sequences and fertility factors. *Genetics* 1991;**129**:177–89. https://doi.org/10.1093/genetics/129.1.177
16. Pinkel D, Landegent J, Collins C. *et al.* Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proc Natl Acad Sci U S A* 1988;**85**:9138–42. https://doi.org/10.1073/pnas.85.23.9138
17. Lichter P, Cremer T, Borden J. *et al.* Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Hum Genet* 1988;**80**:224–34. https://doi.org/10.1007/BF01790090
18. Beliveau BJ, Kishi JY, Nir G. *et al.* Oligominer provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. *Proc Natl Acad Sci U S A* 2018;**115**:E2183–92. https://doi.org/10.1073/pnas.1714530115
19. Rosin LF, Nguyen SC, Joyce EF. Condensin II drives large-scale folding and spatial partitioning of interphase chromosomes in *Drosophila* nuclei. *PLoS Genet* 2018;**14**:e1007393. https://doi.org/10.1371/journal.pgen.1007393
20. Nguyen HQ, Chattoraj S, Castillo D. *et al.* 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. *Nat Methods* 2020;**17**:822–32. https://doi.org/10.1038/s41592-020-0890-0
21. Rosin LF, Gil J, Drinnenberg IA. *et al.* Oligopaint DNA fish reveals telomere-based meiotic pairing dynamics in the silkworm, *Bombyx mori*. *PLoS Genet* 2021;**17**:e1009700. https://doi.org/10.1371/journal.pgen.1009700
22. Fingerhut JM, Moran JV, Yamashita YM. Satellite DNA-containing gigantic introns in a unique gene expression program during *Drosophila* spermatogenesis. *PLoS Genet* 2019;**15**:e1008028. https://doi.org/10.1371/journal.pgen.1008028
23. Chang C-H, Larracuente AM. Heterochromatin-enriched assemblies reveal the sequence and organization of the *Drosophila melanogaster* Y chromosome. *Genetics* 2018;**211**:333–48. https://doi.org/10.1534/genetics.118.301765
24. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;**27**:764–70. https://doi.org/10.1093/bioinformatics/btr011

25. Laetsch DR, Blaxter ML. BlobTools: interrogation of genome assemblies. *F1000Research* 2017;**6**:1287. https://doi.org/10.12688/f1000research.12232.1
26. Vanderlinde T, Dupim EG, Nazario-Yepiz NO. *et al*. An improved genome assembly for *Drosophila navojoa*, the basal species in the *mojavensis* cluster. *J Hered* 2018;**110**:118–23. https://doi.org/10.1093/jhered/esy059
27. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9. https://doi.org/10.1038/nmeth.1923
28. Gelali E, Girelli G, Matsumoto M. *et al*. iFISH is a publically available resource enabling versatile DNA FISH to study genome architecture. *Nat Comm* 2019;**10**:1636. https://doi.org/10.1038/s41467-019-09616-w
29. Mengwei H, Yang B. *et al*. Probedealer is a convenient tool for designing probes for highly multiplexed fluorescence in situ hybridization. *Sci Rep* 2020;**10**:22031. https://doi.org/10.1038/s41598-020-76439-x
30. Zhang T, Liu G, Zhao H. *et al*. Chorus2: design of genome-scale oligonucleotide-based probes for fluorescence *in situ* hybridization. *Plant Biotechnol J* 2021;**19**:1967–78. https://doi.org/10.1111/pbi.13610
31. Hershberg EA, Camplisson CK, Close JL. *et al*. PaintSHOP enables the interactive design of transcriptome- and genome-scale oligonucleotide FISH experiments. *Nat Methods* 2021;**18**:937–44. https://doi.org/10.1038/s41592-021-01187-3
32. Gutzwiller F, Carmo CR, Miller DE. *et al*. Dynamics of *Wolbachia pipientis* gene expression across the *Drosophila melanogaster* life cycle. *G3* 2015;**5**:2843–56. https://doi.org/10.1534/g3.115.021931
33. Hoskins RA, Carlson JW, Wan KH. *et al*. The release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res* 2015;**25**:445–58. https://doi.org/10.1101/gr.185579.114
34. Kim KE, Peluso P, Babayan P. *et al*. Long-read, whole-genome shotgun sequence data for five model organisms. *Sci Data* 2014;**1**:140045. https://doi.org/10.1038/sdata.2014.45
35. Bernardo Carvalho A, Kim BY, Uno F. Strong sequencing bias in Nanopore and PacBio prevents assembly of *Drosophila melanogaster*Y-linked genes. *BioRxiv* 2025;639762. https://doi.org/10.1101/2025.02.23.639762 preprint: not peer reviewed.
36. Albert PS, Zhang T, Semrau K. *et al*. Whole-chromosome paints in maize reveal rearrangements, nuclear domains, and chromosomal relationships. *Proc Natl Acad Sci U S A* 2019;**116**:1679–85. https://doi.org/10.1073/pnas.1813957116
37. Furman BLS, Metzger DCH, Darolti I. *et al*. Sex chromosome evolution: so many exceptions to the rules. *Genome Biol Evol* 2020;**12**:750–63. https://doi.org/10.1093/gbe/evaa081

Briefings in Bioinformatics, 2025, **26(5)**, bbaf507
https://doi.org/10.1093/bib/bbaf507
Problem Solving Protocol