TITLE: Chromosome-scale scaffolding of the fungus gnat genome (Diptera: *Bradysia coprophila*).

AUTHORS:
John M. Urban[1], Susan A. Gerbi[2#], Allan C. Spradling[1#]

[1] Carnegie Institution for Science, Department of Embryology, Howard Hughes Medical Institute Research Laboratories, 3520 San Martin Drive, Baltimore, MD, 21218, USA

[2] Brown University Division of Biology and Medicine, Department of Molecular Biology, Cell Biology and Biochemistry, Providence, RI 02912 USA

[#] Co-senior authors contributed equally to this work.

EMAILS:
John M. Urban : jurban@carnegiescience.edu
Susan A. Gerbi : susan_gerbi@brown.edu
Allan C. Spradling : spradling@carnegiescience.edu

CORRESPONDING AUTHOR:
John M. Urban

ABSTRACT:

Background: The fungus gnat, *Bradysia* (*Sciara*) *coprophila*, has compelling chromosome biology. Paternal chromosomes are eliminated during spermatogenesis whereas both maternal X sister chromatids are retained. Embryos start with three copies of the X chromosome, but 1-2 copies are eliminated from somatic cells as part of sex determination, and one is eliminated in the germline to restore diploidy. These developmentally normal events present opportunities to study chromosome movements that are unusual in other systems. To support such studies, we previously generated a highly contiguous optical-map-scaffolded long-read assembly (Bcop_v1) of the male somatic genome. However, the scaffolds were not chromosome-scale, the majority of the assembly lacked chromosome assignments, and the order and orientation of the contigs along chromosomes remained unknown.

Findings: Male pupae Hi-C data was used to correct, order, and orient the contigs from Bcop_v1 into chromosome-scale scaffolds, producing the updated assembly, Bcop_v2. Several orthogonal analyses allowed us to (i) identify the corresponding chromosome for each scaffold, (ii) orient them with respect to polytene maps, and (iii) determine that they were highly concordant with the chromosomes they represent. Gene annotations produced for Bcop_v1 were lifted over to Bcop_v2. Chromosomal repeat distributions highlight a potential telomeric sequence. Finally, the Hi-C data shed new light on three "fold-back regions" seen to physically interact in images of polytene X chromosomes.

Conclusions: Studies of the unusual chromosome movements in *Bradysia coprophila* will benefit from the updated assembly (Bcop_v2) where each somatic chromosome is represented by a single scaffold.

KEYWORDS: *Bradysia coprophila*, *Sciara coprophila*, fungus gnats, Diptera, insects, genome assembly, chromosomes, scaffolding, Hi-C

Introduction

The lower Dipteran dark-winged fungus gnat, *Bradysia* (*Sciara*) *coprophila*, is an important model system for studying chromosome biology due to its dynamic genome. Most notable is its chromosome cycle. Rather than having either a diploid or haploid set of the same chromosomes (X, II, III, IV, and L) in every cell, the chromosome constitution of a cell varies based on whether it is somatic or germline, male or female, and early embryo or late embryo [1,2]. Regarding somatic or germline, there are chromosomes referred to as the L chromosomes that are only found in the germline [1,2]. Regarding male or female, half of all females, but no males, contain a variant of the X chromosome called the X' (X prime) [1,2]. Regarding embryo age, embryos start out with three copies of the X chromosome, two of which are paternally derived, and with a variable number of the germline-limited L chromosomes, but later embryos have only 1-2 X chromosomes and no L chromosomes in somatic cells. The L chromosomes are eliminated from all somatic nuclei not set aside for the germline in the 5th-6th nuclear division [1,2]. The difference in X chromosome copy number arises because embryos from XX mothers are destined to be male and eliminate two paternal X chromosomes from somatic nuclei in the 7th-9th nuclear division whereas embryos produced by X'X mothers are fated to be female and eliminate only one paternal X [1,2]. Accordingly, sex is determined by the chromosome constitution of the mother in this species, which lacks a Y chromosome. Later in development, both the male and female germline restore diploidy for the X by eliminating one paternal copy, and also eliminate one or more L chromosomes to prevent their accumulation [1,2]. Thereafter, the meiotic events in oogenesis appear to proceed normally whereas those in spermatogenesis do not. In meiosis I of spermatogenesis, all paternal somatic chromosomes are eliminated in a bud of cytoplasm while the L chromosomes appear to escape this imprinting effect and are retained with the maternal set of somatic chromosomes at the single pole of a naturally occurring monopolar spindle [1,2]. In meiosis II, both maternally-derived sister chromatids of the X chromosome undergo developmentally programmed nondisjunction and are retained at one pole of a bipolar spindle [1,2]. Rather than four products, these unusual events lead to one product of male meiosis: a sperm that contains only maternally derived somatic chromosomes (X, II, III, and IV), that is diploid for X, haploid for the autosomes (II, III, and IV), and variable for L chromosomes. In addition to the unusual chromosome cycle, *Bradysia coprophila* larval salivary glands have highly polytene chromosomes with over 8000 copies in each nucleus [3], and developmentally-programmed gene amplification [4,5]. Finally, there is mounting evidence that this genome may be a great model for studying horizontal gene transfer (HGT) in insects [6–8]. Overall, fungus gnats present many opportunities to study chromosome biology across development and evolution.

Previously, we published a highly contiguous assembly of the somatic chromosomes (X, II, III, and IV) named Bcop_v1 [9]. We targeted the somatic genome in males as it presented the lowest complexity version of the genome, lacking both the L and X' chromosomes, which have since been the focus of other studies [6,10]. Specifically, we sequenced genomic DNA from male embryos using long-read (PacBio RS II SMRT and Oxford Nanopore Technologies MinION) and short-read (Illumina) technologies [9]. As part of that process, we assembled the sequencing data ~100 different ways, and used many evaluations to determine the best assembly and polishing pipelines for our datasets [9]. Two assemblies with the best evaluation scores were scaffolded with optical maps from BioNano Genomics, then subject to further evaluations[9]. The BioNano-scaffolded assembly with the best evaluation scores, Bcop_v1 (NCBI GenBank

GCA_014529535.1, WGS VSDI01), had a contig NG50 of ~2.4 Mb and scaffold NG50 of ~8.2 Mb (Table 1) [9]. Since, in *Bradysia coprophila*, late male embryos have a single copy of the X, but are diploid for the autosomes, it was possible to classify each contig as either X-linked or autosomal using sequencing depth [9]. Moreover, using the known chromosomal locations of a handful of sequences, we were able to anchor ~28-33% of the autosomal sequences into chromosomes II, III, and IV [9]. Overall, nearly 50% of the expected genome size was anchored into chromosomes based on read depth and known sequence locations [9]. Nevertheless, chromosome assignments are not available for the majority of that assembly (Bcop_v1). Moreover, Bcop_v1 lacks any information regarding the ordering and orientation of the contigs along chromosome sequences. To better facilitate studies into gene amplification and the unusual chromosome dynamics in *Bradysia coprophila*, our goal was to develop better models of the somatic chromosome sequences. We approached this by using chromosome conformation capture with deep sequencing (Hi-C) [11] to scaffold the chromosomes as has been successfully done by many others [12–14]. Specifically, Hi-C is powerful at putting chromosome sequences together from a collection of sub-chromosomal sequences because of three major attributes of *in vivo* chromosome interactions captured by Hi-C data: (i) the interactions arbitrarily yield Mb-scale and chromosome-scale information without the need for sequencing ultra-long DNA molecules, (ii) intra-chromosomal interactions are much more frequent than inter-chromosomal interactions, making it possible to use interaction frequencies to cluster contigs into 'chromosome groups', and (iii) interaction frequencies decay with distance, making it possible to use them to order and orient the contigs in a chromosome group in the order and orientations they most likely appear along a chromosome [11–14].

Overall, we present here the first chromosome-scale genome assembly for *Bradysia coprophila*, specifically reporting the first full-length scaffolds for the somatic chromosomes (X, II, III, IV). In addition, we lifted over the annotations from the former genome assembly, allowing studies in progress by multiple groups to seamlessly transition to the updated genome assembly. Through several approaches, we demonstrate the chromosome identity and high structural accuracy of the chromosome-scale scaffolds, including through inspecting Hi-C interaction frequencies, correlation with polytene maps, how repeats are distributed, read depth, evolutionary expectations, and special features of the X chromosome. The termini of the chromosome-scale scaffolds hint that telomeres may be composed of Long Complex Terminal Tandem Repeat (LCTTR) sequences as seen in the non-biting midge *Chironomus* and the mosquito *Anopheles gambiae* [15,16], but also arrays of retrotransposon-related sequences. In addition, the Hi-C maps show three loci, separated by >10 Mb each, with high frequency interactions *in vivo*, potentially illuminating the loci of the three "fold-back regions" seen to physically interact in images of the polytenized X chromosome. Finally, multiple genomic datasets demonstrate that recently-discovered "alien" genes [7] are genuine parts of the chromosome-scale scaffolds, not contamination or mis-assemblies. The Hi-C data, the chromosome-scale genome sequence, its gene annotations, and insights produced here will be immediately useful to the growing research community interested in the unique biology of *Bradysia coprophila* as well as the broader research community interested in Dipteran evolution and comparative genomics.

Results

Hi-C guided assembly correction and chromosome-scale scaffolding of the reference genome.

To update the previous reference genome, Bcop_v1 [9] (for background on Bcop_v1, see the Introduction, Table 1), we generated Hi-C data for chromosome-scale scaffolding. Approximately 115.2 million 101bp paired-end reads were produced from the chimeric DNA molecules obtained with a commercially-available chromosome conformation capture (Hi-C) procedure from Phase Genomics (Proximo Hi-C kit) performed on whole male pupae (see Methods). The paired-end Hi-C reads were first used to produce a corrected assembly (called "Bcop_v1_corrected") by aligning them to the 205 primary contigs of Bcop_v1 ("Bcop_v1_primary") [9], then manually inspecting the Hi-C interaction frequency signal within the Bcop_v1 contigs for disruptions in the expected pattern corresponding to putative mis-joined regions (Fig. 1A-C). Bcop_v1 had 22 contigs with 46 such disruptions in the Hi-C signal (examples in Fig. 1B-C). The 22 contigs were "corrected" by breaking them into "sub-contigs" at the putative mis-joined regions. Of the 46 "debris" regions where breaks were made in Bcop_v1 contigs, 32 were adjacent to BioNano optical map gaps, i.e. regions joined together previously by BioNano optical maps in Bcop_v1, with gap lengths ranging from ~4.4 to ~674 kb. However, of the 156 BioNano gaps in Bcop_v1, there were 124 ranging from 25 bp to ~400 kb that were conserved after the correction step, 70 of which were on the same 22 contigs that had "debris" regions and 54 that were elsewhere within Bcop_v1. Thus, the regions joined by BioNano optical maps mostly did not present as putative mis-joins. Lower mapping quality from flanking repeats around BioNano gaps may have simply given rise to spurious Hi-C signals more often than random. In summary, the majority of mis-join signals (~70%) flanked a minority of assembly gaps, and mis-join signals affected a minority of contigs (10.7%) that were broken up by Hi-C-guided correction. This step transformed the input assembly (Bcop_v1_primary) that had 205 contigs into the corrected assembly (Bcop_v1_corrected), which had 297.

The Phase Genomics' Proximo Hi-C genome scaffolding platform was then used to create chromosome-scale scaffolds from the 297 contigs in the corrected input assembly (Bcop_v1_corrected; Fig. 2A). After scaffolding and post-processing (see Methods), there were four primary chromosome-scale scaffolds, each spanning 58-71 Mb and another 52 input contigs that summed to ~1.8 Mb, which were not placed in the chromosome-scale scaffolds (Table 1; Fig. 2B; see Methods). Thus, the vast majority of the ~299 Mb of primary sequence from Bcop_v1 (Table 1) input into the Hi-C scaffolding process was integrated into the four chromosome-scale scaffolds that summed to ~297 Mb. In addition, there were 539 "associated contigs" from Bcop_v1 [9], largely comprised of haplotigs and repeat variants, that were not included during the Hi-C scaffolding process (see Methods), which summed to 10.8 Mb and were ~20 kb on average. As a semi-final step in producing the updated reference genome, Bcop_v2, these 539 contigs were also added back to the assembly. Moving forward, the four chromosome-scale scaffolds are referred to as the "primary assembly" for Bcop_v2 whereas the "associated assembly" for Bcop_v2 is composed of the 52 unplaced input contigs and the 539 "associated contigs" from Bcop_v1, totaling 591 "associated contigs" for Bcop_v2. Most of the associated assembly is comprised of haplotigs and repeat variants. Researchers interested mainly in the four chromosome-scale scaffolds can opt to use or ignore the associated assembly. The four primary chromosome scaffolds make up ~96% of the entire assembly with lengths of 58.3, 70.5, 71.0, and 97.1 Mb (Tables 1 and

2), and together with the 591 associated contigs have an expected size of ~81 Mb, and an NG50 length of ~71 Mb pertaining to an LG50 count of only 2 scaffolds (Table 1). In the updated assembly (Bcop_v2), gaps from the original BioNano scaffolding process (for Bcop_v1) ranged from 25 bp to ~662 kb with a gap N50 of ~101.5 kb. Hi-C scaffolding introduced 241 gaps, all arbitrarily set to 100 bp, 5 of which are on a short ~363 kb scaffold. In total, there are 367 Hi-C and/or optical map gaps across Bcop_v2, 360 of which are within the four chromosome-scale scaffolds. In summary, Hi-C scaffolding transformed the corrected assembly into a preliminary version of Bcop_v2, which had 4 primary scaffolds and 591 associated contigs. The scaffolds were further oriented using the positions of known sequences as described in a subsequent section.

Assessing the consistency of the Hi-C data with chromosome specificity and structural accuracy of the scaffolds.

With regard to scaffolding accuracy, we interrogated 'chromosome specificity', the degree to which contigs in a given scaffold are likely to come from the same chromosome, and 'structural accuracy', the degree to which contigs are joined together in the correct order and orientation with respect to the chromosome. Before pursuing orthogonal evaluations of both metrics (see subsequent sections), the Hi-C signal across Bcop_v2 was manually inspected to ensure that these metrics were optimized according to expected interaction frequency patterns.

First, intra-chromosomal interaction frequencies are expected to exceed inter-chromosomal interaction frequencies [11]. As such, for high 'chromosome specificity', the same should be true for interaction frequencies within scaffolds compared to between scaffolds. As can be seen in Fig. 2A, the regions of high and low interaction frequencies are all over the map for the input into scaffolding (Bcop_v1_corrected) where contigs are simply ordered from longest to shortest. However, after scaffolding, in the Hi-C map for Bcop_v2, the contigs are ordered such that those regions of high average interaction frequencies are confined within four chromosome-sized scaffold squares along the diagonal and the regions of lowest interaction frequency correspond to inter-scaffold contacts (Fig. 2B). This indicates that the contigs within a given scaffold correspond to the same chromosome and that different scaffolds correspond to different chromosomes.

Second, within a chromosome, interaction frequencies are expected to be highest between adjacent loci and to decay with distance along the linear chromosome sequence [11]. As such, for high 'structural accuracy' within a given scaffold, short distances should have the highest interaction frequencies both within contigs and also between adjacent ends of neighboring contigs within the scaffold. The former was ensured in the previous pre-scaffolding "correction" step and the latter will hold true if the contigs were ordered and oriented correctly. High structural accuracy of Bcop_v2 is indicated by the strong Hi-C interaction frequency signal along the diagonal within the chromosome scaffolds (Fig. 2C-F), with no disruptions between neighboring contigs, and by the weaker off-diagonal interaction frequencies that rapidly decay with distance. In addition, several contigs were selected for further analysis of the interaction frequencies connecting them to adjacent contigs. For the start (or end) of each contig inspected, there were orders of magnitude more interactions with the immediately adjacent terminal of the nearest neighboring contig than any other region of any other contig within the scaffold. It also had comparable interactions with

its opposite flank within the same contig (see example in Fig. 2G). Overall, these results indicate that the adjacent contigs within a given scaffold correspond to adjacent loci in the chromosome according to the logic of Hi-C.

In sum, the Hi-C signal is consistent with well put together chromosome scaffolds that have both high chromosome specificity of the contigs within each scaffold and high structural accuracy in how they are ordered. Nonetheless, since the scaffolds were put together using this same Hi-C logic and data, we sought other ways to test the chromosome specificity and structural accuracy of the chromosome scaffolds, including looking at the order of multiple known anchor sequences, centromere positions, how repeats are distributed across the scaffolds, known features expected for the X, and comparisons to the chromosomes of other Dipteran genomes, all detailed below.

Establishing the chromosome identity, polytene orientation, and structural accuracy of the chromosome scaffolds using the locations of multiple anchor sequences.

The chromosomal identities of the chromosome-scale scaffolds were obtained using "anchor" sequences for 10 known unique chromosomal locations [17–29] by mapping them to the scaffolds (Fig. 3A-D; Tables 2 and 3; see Methods). All anchor sequences from a given chromosome mapped to the same chromosome-scale scaffold, and there were no conflicts of chromosome identity among the multiple anchor sequences (Fig. 3A-D). These results further confirm that each chromosome-scale scaffold corresponds to one and only one chromosome and that each scaffold corresponds to a different chromosome than the three other scaffolds. Thus, the anchor sequences are concordant with the high 'chromosome specificity' of the contigs within scaffolds observed in the Hi-C maps.

The structural accuracy was further tested by the analyzing the order of the anchor sequences across the scaffolds in addition to the approximate positions of centromeres mapped onto each scaffold with a known centromeric sequence (Fig. 3A-D; see below Results and Methods). Since both the polytene map and the scaffold locations of the 10 anchor sequences and 4 centromeres were known, it was also possible to orient the chromosome scaffold sequences in the same direction as locus numbers have been assigned on polytene maps for all four chromosomes [30] (Fig. 3A-D), which was the final processing step for Bcop_v2. Importantly in terms of the structural integrity of how the scaffolds were ordered by Hi-C, when multiple anchors were present for a given chromosome, they all mapped in the expected order along the same scaffold (Fig. 3A-D). Thus, a sampling of 14 sites across the four chromosomes further confirms the high structural accuracy of the order of contigs along the chromosome-scale scaffolds.

Chromosome lengths and copy numbers further confirm chromosome identities.

The chromosome identities of each scaffold were further supported by comparing the scaffold lengths with prior expectations of the chromosome lengths. In very early studies of these chromosomes by H.V. Crouse [31], chromosome II was predicted to be shortest, chromosome IV was established as the longest, and chromosomes X and III were expected to be of equivalent lengths. Crouse's early expectations are confirmed by the chromosome-scale scaffolds (Table 2;

282 Fig. 3E), of which II is shortest (58.3 Mb), IV is the longest (97.1 Mb), and X and III are
283 approximately the same length (70.5 and 71.0 Mb, respectively). Moreover, the chromosome
284 lengths reflected by the scaffolds are concordant with the expected sizes of each chromosome
285 using the number of polytene sections assigned by Crouse [31] and the expected genome size from
286 DNA content measurements [32] (Table 2). The total length (226 Mb) of the autosomal scaffolds
287 (II, III, IV) was also concordant with these predictions (214-231 Mb) (Table 2). Note that the
288 chromosome length expectations and the number of zones in each from Crouse [31] were later
289 revised such that X and II were both shortest and of equivalent length [30], but this under-estimate
290 of the X was anticipated by Crouse [31] and is now refuted by the chromosome-scale scaffolds
291 that support Crouse's earlier length expectations (Table 2; Fig. 3E).
292
293 The chromosome identity of the X chromosome scaffold and of the autosomal nature of
294 the other scaffolds (II, III, and IV) were further supported by read depth analysis. Females are
295 diploid for both the autosomes and the X chromosome. Males are diploid for the autosomes, but
296 haploid for X. Thus, read depth for females along the X and autosomes should be equivalent
297 whereas read depth along the X from males should be half that of autosomes. These read depth
298 expectations are borne out by multiple genomic technologies (Fig. 3F), including paired-end
299 Illumina reads for both types of adult female and adult males [10], long reads from Nanopore and
300 PacBio for male embryos [9]; and BioNano Genomic optical maps from male pupae [9]. The
301 diploid-level read depth in both males and females across the scaffolds for chromosomes II, III,
302 and IV confirms that they are strictly composed of autosomal contig sequences (Fig. 3F). Similarly,
303 the haploid-level read depth in all male samples, but diploid level in female samples, across the
304 entirety of the X chromosome scaffold, confirms not only the chromosome identity of the scaffold
305 as X, but that all the contigs within it are indeed X-linked (Fig. 3F).
306
307
308 **Gene annotations for the chromosome-scale assembly (Bcop_v2).**
309
310 There are presently two high quality gene annotation sets produced using the previous
311 genome assembly (Bcop_v1). One was constructed by us [9] using Maker2 [33], which can be
312 found at USDA Ag Data Commons [34]. The other was made by the NCBI Eukaryotic Genome
313 Annotation Pipeline for RefSeq, is called NCBI *Bradysia coprophila* Annotation Release 100, and
314 can be found at NCBI [35]. These gene annotations have already been demonstrated to score highly
315 in 'completeness' as determined by the number of Dipteran BUSCOs (97%) found within them,
316 where the BUSCOs used were defined as genes present as single copy orthologs (SCOs) in 95%
317 of Dipterans [9,36]. While the FASTA files containing the transcript and protein sequences are in
318 no need of updating, the GFF files that detail the coordinates of the gene models, including their
319 exons, introns, CDS, and UTRs, on Bcop_v1 are not useful for the updated assembly, Bcop_v2.
320 Since Bcop_v2 was produced by stitching together the Bcop_v1 sequences, a simple "liftover"
321 procedure, where features from an assembly can be exactly mapped to an updated version, was
322 adequate for creating new GFF files detailing the coordinates of the genes on Bcop_v2.
323
324 Both gene annotation sets had similar lift-over success rates for all metrics (Table 4). Using
325 the Maker2 annotation set as an example, of the 23,117 genes and 28,870 transcripts, only 2 genes
326 with 4 transcripts were not mapped to Bcop_v2. Of the 28,866 transcript models successfully lifted
327 over to Bcop_v2, all but one (28,865) had perfect full-length alignments to the original transcript

328 sequences, although 52 (<0.2%) were a bit shorter than the original. In total, only 53 transcript
329 sequences had differences in the updated genome assembly (Table 4). Even fewer, just 49
330 transcripts, had differences at the level of protein sequence, which is to say that nearly all (99.83%)
331 had identical protein sequences between assemblies (Table 4). An even higher percent at the gene
332 level was identical between assemblies. For example, 99.87% of lifted-over genes had identical
333 protein sequences for all transcript isoforms, and 99.91% had identical protein sequences for all or
334 at least a subset of transcript isoforms (Table 4).
335
336 We also visually confirmed in a genome browser that gene models at the same loci look
337 the same between assemblies. Moreover, the lifted over gene models were highly concordant with
338 RNA-seq coverage as illustrated in Fig. 4A-F, which shows six example genes of interest from
339 recent *B. coprophila* papers [7,37] and that, for simplicity, shows read depth from only one RNA-
340 seq sample of irradiated larvae [37]. The gene models were highly consistent with RNA-seq
341 coverage from all datasets we checked, including from various life stages [9], radiation conditions
342 [37], and unpublished data. Overall, gene/protein sequence comparisons between assemblies and
343 gene model concordance with RNA-seq both demonstrate that the lift-over process successfully
344 transferred the annotation information from Bcop_v1 to Bcop_v2. The chromosomal location of
345 genes allowed us to analyze the number of genes per chromosome (Table 3) as well as
346 chromosome-specific log2 fold-changes of male versus female gene expression in multiple life
347 stages (Fig. 4G), which supported a dosage compensation mechanism on the single male X as we
348 and others found previously [9,38].
349
350
351 **Gene sets on chromosome scaffolds fit expectations of chromosome similarity with other**
352 **species as a function of evolutionary distance.**
353
354 Over evolutionary time, chromosomes tend to undergo rearrangements that include
355 translocations. Through these rearrangements, genes get shuffled around the genome, and more
356 gene shuffling is likely to occur with increasing evolutionary distance. When defining
357 chromosomes as sets of genes, two closely related species are likely to have very similar
358 chromosomes and therefore highly intersecting chromosomal gene sets. In contrast, distantly
359 related species are likely to have smaller intersections between chromosomal gene sets,
360 approaching "random" proportions of shared genes between arbitrarily selected chromosomes
361 from each species. While seeing this trend with respect to *B. coprophila* genes on Bcop_v2
362 chromosome-scale scaffolds would further support the accuracy of the Hi-C scaffolding process,
363 not seeing this trend would raise concerns.
364
365 To test the evolutionary prediction described above, the chromosomal groupings of *B.*
366 *coprophila* genes (chromosomal gene sets) was compared with that of four other fly (order
367 Diptera) species with known evolutionary relationships (Fig. 5A) and which all had chromosome-
368 scale genome assemblies [39–42]. The most distantly related species selected was *Drosophila*
369 *melanogaster* from the so-called "higher Diptera" (sub-order Brachycera). In contrast, *B.*
370 *coprophila* and the three remaining species are all part of the "lower Diptera" (sub-order
371 Nematocera), which diverged from the higher Diptera ~200 MYA [43]. The closest related species
372 to *B. coprophila* was *Pseudolycoriella hygida*, which is within the same family (Sciaridae)
373 commonly known as black fungus gnats or dark-winged fungus gnats. *P. hygida* had a

374 chromosome-scale genome release in 2023 [42] following the release of Bcop_v2, which was
375 opportune for this analysis. The remaining two Nematocerans are mosquitoes (Culicidae family):
376 the yellow fever mosquito (Aedes aegypti) and the African malaria mosquito (Anopheles
377 gambiae). The proteomes of all fives species were compared to determine orthologs as well as to
378 construct a proteome-wide rooted species tree, which recapitulated the known evolutionary
379 relationships (Fig. 5B). This indicates that the genes determined to be orthologs across species
380 should also be informative when comparing their chromosomal locations.
381
382 To test if the degree of gene shuffling around the chromosomes of these five species also
383 reflected their known phylogenetic relationships, the genomic locations of "Single copy orthologs"
384 (SCOs) were visually compared across all five species (Fig. 5C) with respect to their locations
385 within *B. coprophila* chromosomes (Bcop_v2). The five species had 4,297 SCOs. The
386 chromosomal addresses of SCOs were compared between each pair of species relative to *B.*
387 *coprophila* in order to (i) map *B. coprophila* chromosomes to their most likely counterparts in each
388 other species (Table 5) and (ii) visualize as dot plots (Fig. 5C). *B. coprophila* was plotted against
389 itself to demonstrate the base case of what the absence of gene shuffling looks like (Fig. 5C).
390 Visually, the amount of gene shuffling increased with evolutionary distance as expected. While
391 there is gene shuffling even between *B. coprophila* and the closest related species tested (*P.*
392 *hygida*), it is mostly intra-chromosomal and this pair demonstrates the highest amount of
393 conservation of SCOs from given *B. coprophila* chromosomes to their single corresponding
394 chromosomes in the other species (Fig. 5C; Table 5). In contrast, the most distantly related species
395 (*D. melanogaster*) visually has the most gene shuffling across chromosomes, and the mosquito
396 species have intermediate levels, as expected (Fig. 5C; Table 5). Thus, visual inspection of the
397 pairwise SCO dot plots makes a strong case that the *B. coprophila* gene sets defined by the
398 chromosome-scale scaffolds (Bcop_v2) follow evolutionary expectations.
399
400 The amount of "gene shuffling" was also quantified between all pairs of the five species
401 using the single summary metric of "entropy". Briefly, entropy can be thought of as a measure of
402 "mixed-up-ness" that can quantify systems as they evolve from a non-random to a random state
403 [44]. Here, it is used to measure the amount of inter-chromosomal gene shuffling between pairs of
404 species, and ultimately as a function of evolutionary distance from *B. coprophila*. A normalized
405 entropy metric that ranges between 0 and 1, called Min-Max Normalized Entropy (MMNE), was
406 computed for each pairwise comparison of two species. MMNE shows the observed entropy
407 relative to the minimum and maximum entropy values possible for that pairwise species
408 comparison given the number of shared SCOs, chromosomes, and associated probabilities (see
409 Methods; see example SCO probabilities in Table 6). The initial non-random lowest entropy state
410 (when MMNE is 0) is simply a comparison to self (Fig. 5D). Entropy is expected to increase
411 (MMNE approaches 1) with more distantly related species that had larger periods of time for inter-
412 chromosomal rearrangements to shuffle genes around, which is what is observed with increasing
413 evolutionary distance from *B. coprophila* (Fig. 5D). Furthermore, performing clustering on the
414 matrix of all pairwise inter-species entropy scores reproduces the species groupings expected by
415 the known evolutionary relationships (Fig. 5E), putting the fungus gnat family members (*B.*
416 *coprophila* and *P. hygida*) together within the larger Nematoceran cluster that also included the
417 pair of mosquitoes, all of which were on a separate branch from the single Brachyceran, *D.*
418 *melanogaster*. The results of these comparative genomic evaluations further bolster the confidence
419 we have in the Hi-C directed process that yielded Bcop_v2 chromosome-scale scaffolds.

420 **Investigating reported horizontal gene transfer using the chromosome scaffolds, Hi-C, and**
421 **other genomic data.**
422
423 Recently an analysis of the CYPome (P450 genes) of *B. coprophila* lead to the conclusion
424 that there has been horizontal gene transfer (HGT) from springtails, mites, and fungi [7]. There
425 were several lines of evidence that these so-called "alien" genes, which appear to have been
426 obtained by HGT, are truly integrated into the nuclear chromosomes as opposed to being present
427 from contamination sources [7]. For example, the "alien" genes had homology and synteny in the
428 genomes of closely related species from three different continents and had evidence of being
429 expressed and spliced [7], which is also shown for three such genes in Fig. 4. The locations of
430 these genes in the chromosome-scale scaffolds allowed a further investigation into the likeliness
431 that these 'alien' genes are truly integrated into the chromosomes of *B. coprophila* with a variety
432 of orthogonal genomic datasets, including Hi-C, long reads, and optical maps. First, the locations
433 of the P450 genes in the chromosome-scale scaffolds did not give rise to disrupted or otherwise
434 unexpected Hi-C signals. The contigs they reside on survived the Hi-C guided correction step and
435 were stitched into the chromosome-scale scaffolds as part of the Hi-C guided scaffolding step.
436 Moreover, the alien P450 genes have typical Hi-C interaction frequencies that decay with distance
437 that are comparable in all cases to nearby 'highly conserved fly genes' of similar length (Fig. 6).
438 Second, across a variety of genomic datasets from different technologies and samples, there were
439 no disruptions in sequencing depth seen over regions containing the putative "alien" P450 genes
440 (Fig. 6) whereas disruptions were seen over gaps in the assembly as expected. Finally, there exist
441 Hi-C interactions and single DNA molecules corresponding to long reads and optical maps that
442 physically connect the alien P450 genes to highly conserved Dipteran genes (data not shown). In
443 summary, while we make no claims here about HGT, we do confirm that the so-called "alien"
444 P450 gene sequences in the *B. coprophila* genome assembly are further supported as being part of
445 the nuclear chromosomes rather than from a separate contamination source by our analyses of a
446 diverse set of genomic datasets. The same conclusion can be made from similar analyses of five
447 terpene synthase genes within the *B. coprophila* chromosome-scale scaffolds (Bcop_v2), which
448 also have been proposed to have originated by HGT from mites [8].
449
450
451 **The locations of known centromeric repeats and peri-centromeric transposons further**
452 **support the chromosomal structures of the scaffolds.**
453
454 Centromere locations have previously been mapped on polytene maps [22,30,31,45].
455 Centromere positions in all chromosome scaffolds were approximated using the centromere-
456 associated repeat sequence, Sccr (*Sciara coprophila* centromeric repeat), that is known to
457 hybridize to the centromeres of all four somatic chromosomes [45], and an associated repeat family
458 that contains Sccr (see Methods). The Sccr-repeat-family-enriched positions were found at the
459 expected centromeric locations in all chromosome scaffolds (Fig. 3A-D, Fig. 7A). Specifically,
460 chromosomes X, II, and III are acrocentric with centromeres very close to the beginning of these
461 chromosomes (as oriented) at polytene zones X/1A, II/1C, and III/1C. In agreement, Sccr mapped
462 within the first few megabases of the corresponding 58-71 Mb scaffolds (Fig. 3A-D, Fig. 7A). In
463 contrast, chromosome IV is metacentric with a centrally-located centromere at polytene zone
464 IV/9B, where zones span from IV/1A to IV/20C. As expected, Sccr mapped to a medial region, at
465 positions spanning ~44.7-45.0 Mb, of the 97.1 Mb chromosome IV scaffold (Fig. 3A-D, Fig. 7A).

466 Importantly, the predicted centromeric positions in the chromosome scaffolds were concordant
467 with respect to their relative positions between other anchor sequences (see above Results) in the
468 order of anchor sequence loci along the polytene maps (Fig. 3A-D, Fig. 7A), again highlighting
469 the structural accuracy of the scaffolds.
470
471 A specific RTE-like retro-transposable element called ScRTE is increasingly enriched with
472 proximity to all centromeres, which was previously shown experimentally using a sub-sequence
473 from ScRTE, called "F4", as a probe [45]. If the scaffolds represent the chromosomes well and if
474 the centromeric positions within them defined above using Sccr are correct, then ScRTE should
475 also show increasing enrichment with proximity to those centromeric positions. To test this, we
476 found all matches specifically to the F4 probe sequence across the chromosome scaffolds as well
477 as the distribution of all RTE elements identified during repeat annotation (described below). Both
478 showed the expected trend of RTE transposons being increasingly enriched with proximity to the
479 centromere positions approximated by Sccr (Fig. 3A-D, Fig. 7A). This further supports the
480 centromere positions within the scaffolds. Moreover, it indicates that the contigs were ordered
481 within the scaffolds such that they are increasingly enriched for RTE with proximity to the
482 centromeres, and therefore again signifies high structural accuracy.
483
484
485 **Repeat locations show family-specific biases towards peri-centromeric and sub-telomeric**
486 **regions.**
487
488 In general, repeats and transposons are often enriched near centromeric, peri-centromeric,
489 telomeric- and sub-telomeric regions in Dipteran chromosomes [39,40,46]. To test this
490 expectation, we analyzed the distribution of repeats across the *B. coprophila* chromosome-scale
491 scaffolds. Centromeric positions are defined here by the positions mapped by the Sccr-associated
492 repeat family as above and peri-centromeric positions are defined as the much broader repeat-rich
493 regions flanking each side. The telomeric and sub-telomeric regions are defined here simply as the
494 regions near and including the scaffold ends, which have been shown to correspond to sub-
495 telomeres and telomeres in other Hi-C-based scaffolds of insect assemblies [47,48]. If present,
496 telomeric repeats are expected only in the terminal-most 10-50 kb, and sub-telomeres are the
497 immediately-adjacent, much broader repeat-rich regions at chromosome ends. For the sake of
498 clarity, and to avoid conflating telomeres and sub-telomeres, the beginning and end of each
499 chromosome scaffold, as oriented, will sometimes be referred to as its left and right termini,
500 respectively.
501
502 First, the entire genome assembly was mapped against itself (or individual chromosome
503 scaffolds against themselves) to create dotplots that naively show all pairwise regions that share
504 high sequence similarity with each other (Fig. 7B). The regions around and including centromeric
505 positions and scaffold termini were very obvious in these plots, with highly repetitive natures
506 locally within each chromosome scaffold and also with high similarity among the corresponding
507 regions across all chromosome scaffolds. The first 10 Mb of the left terminus of chromosome X is
508 composed of two adjacent repeat blocks. The first repeat block at the beginning of X corresponds
509 to the left telomeric/sub-telomeric zone, rDNA, and centromeric repeats (Sccr) among other
510 centromere-proximal repeats (Fig. 7B). To the right of the centromere but still on the left side of
511 the chromosome, the second repeat block contains two adjacent loci of interest that are discussed

512 in a subsequent section (centromere-proximal X' breakpoint and putative "FBR1"), and is more
513 similar to the repeat block in the final 5-8 Mb of the right terminus of the scaffold (distal telomeric
514 side) than it is to the left terminal block (Fig. 7B). The chromosome III scaffold has a similar
515 structure on its left terminus. It starts with a 10-12 Mb, centromere-containing, high density repeat
516 block followed on its right by a second smaller repeat block that is more closely related to the
517 terminal ~5 Mb repeat block at the distal telomeric end on the right terminus of the scaffold than
518 the centromeric end to the left terminus (Fig. 7B). Unlike X and III, the scaffold for chromosome
519 II has a single large (~8 Mb) repeat block at its centromeric end (left terminus) that has high
520 similarity with the terminal ~8 Mb at its distal telomeric end on its right side (Fig. 7B). The repeat
521 densities of the chromosome II repeat blocks seem lighter than the other chromosomes in this
522 method of analysis, but not necessarily so in other analyses (Fig. 7A-B). The metacentric
523 chromosome IV scaffold has a repeat block at both termini, each ~8 Mb, as well as a repeat block
524 in the middle of ~10 Mb surrounding the approximated centromere position in the scaffold (Fig.
525 7B). Finally, the sequences in the "associated assembly", which are not placed in the chromosome-
526 scale scaffolds, are enriched for matches within the centromere-proximal repeat blocks, but also
527 contain matches to rDNA and rDNA-proximal repeats, and matches to terminal repeat blocks to a
528 lesser extent (Fig. 7B). Overall, the repeat structures of the scaffolds help confirm their centromere
529 regions and follow expectations of repetitive structures associated with scaffold termini
530 corresponding to telomeric and/or sub-telomeric regions.
531
532 To further analyze the distribution of repeats in the chromosome-scale scaffolds, major
533 repeat families were modeled, annotated, classified when possible, and their relative densities
534 visualized along the linear scaffold sequences (Table 7). Tandem repeats were also mapped
535 independently. Both tandem repeats as well as the summation of all modeled repeat families are
536 indeed enriched around defined centromere positions and scaffold termini (Table 7; Fig. 7A; see
537 Methods). When inspecting different transposon families, their densities near centromeres and
538 scaffold termini (a proxy for telomeres) had different biases. Retrotransposable element families
539 (RTE, LTR, LINE) were generally more enriched around centromeric positions than near scaffold
540 termini (Fig. 7A). Their densities are higher at the centromeric (left) ends of the acrocentric
541 chromosomes (X, II, and III) compared to the right ends, and higher at the medial centromere
542 position of IV compared to its termini (Fig. 7A). These families have relatively little or no
543 enrichment in terminal regions, which have similar densities compared to the rest of the
544 chromosome lengths (Fig. 7A). LTR elements are particularly enriched near the centromere of
545 chromosome III (Fig. 7A). Sequence labeled as LINE was most abundant in general (Table 7; Fig.
546 7A). DNA transposon families have a different bias, seeming to have higher relative densities at
547 least in some telomeric and/or sub-telomeric regions than near centromeres (Fig. 7A). For
548 example, see the non-centromeric ends (right termini) of the scaffolds for chromosomes X and II
549 (Fig. 7A). A sub-family called hAT-Tag1 is a great example of an element whose highest densities
550 are in terminal regions compared to other regions of the chromosome scaffolds, including
551 centromeres (Fig. 7A). Nevertheless, although DNA elements may be more biased in termini of X
552 and II than their centromeres, DNA transposons are also found near all centromeric and peri-
553 centromeric regions, illustrated well by the medial centromere of chromosome IV (Fig. 7A).
554 Moreover, the bias of DNA transposons to these repeat-rich regions compared to gene-rich
555 intervening regions is not as high as for other transposon families (Fig. 7A). Finally, the "Rolling
556 Circle" Helitron family is distributed such that its highest densities appear similar in both terminal

557 and peri-centromeric regions, and the overall weight of Helitron elements is very concentrated in
558 these regions compared to gene-rich areas of the chromosome arms (Fig. 7A).
559
560 Repeats that could be recognized and classified as belonging to a particular repeat or
561 transposon family tend to be enriched with proximity to centromeres and termini (Fig. 7A). This
562 is also true for locations that match conserved classified repeats from across other arthropods (Fig.
563 7A). Thus, the regions associated with termini and around centromeres are enriched for repeats
564 that are conserved enough to classify into families and/or be most similar to repeats from other
565 species. However, the majority of repeat families in the repeat library trained *de novo* on the *B.*
566 *coprophila* genome could not be classified according to known repeat families in the same
567 automated fashion by the repeat modeling and annotation programs (Table 7), and are referred to
568 as 'unclassified repeats'. While the density of unclassified repeats is still moderately higher near
569 centromeres and telomeres, the bias is much lower, and the unclassified repeats are much more
570 spread out across the rest of the chromosome arms (Fig. 7A). It is possible some of these
571 unclassified repeat families actually correspond to gene families or repeat families, newly arisen
572 or otherwise, that are just not included as possible classification terms in the programs used.
573 Unclassified repeats may also consist of highly degraded copies of known repeat families that were
574 simply not recognized, which would indicate that repeat sequences are either better maintained
575 near centromeres and near chromosome ends, or younger in those regions.
576
577
578 **Chromosome-scale scaffold termini highlight arrays of retrotransposon-related and satellite**
579 **sequences as candidate telomeric repeats.**
580
581 The telomeres of most studied eukaryotes are composed of very short (5-10 bp) tandem
582 repeats that are maintained by telomerase [15,16,47,48], most often similar to the mammalian
583 hexamer (TTAGGG). For example, most insects have the pentamer, TTAGG [16]. However, most
584 or all Dipterans studied have lost both the short tandem repeats and the telomerase gene, TERT
585 [16]. The telomeres of *Drosophila*, a "higher" Dipteran, are composed of telomere-specific
586 retrotransposons (TSRTs) that point towards the centromere and are maintained by their reverse
587 transcriptase genes [15,16,47,48]. The telomeres of lower Dipterans are less well understood.
588 *Rhynchosciara americana* appears to have 14-22 bp long tandem repeats, which might be
589 maintained by a TSRT related to Drosophila TSRTs [16]. Non-biting midges (Chironomidae) have
590 "Long Complex Terminal Tandem Repeats" (LCTTRs) with a unit length of 176 bp in one species
591 and ~350 bp in another species, although the latter appears to have evolved from a simpler 175 bp
592 repeat [15,16]. The long repeats in midges might be maintained by homologous recombination or
593 gene conversion, although long complementary RNAs have been observed, suggesting an RNA
594 intermediate as with telomerase and TSRTs [15,16]. A mosquito has even longer repeat lengths of
595 ~820 bp in its telomeres, and at least one higher Dipteran, *Drosophila tristis*, has LCTTRs of
596 similar length to midges (181 bp) [15]. In at least some midges, and in some insects more generally,
597 telomeres can differ from each other, and some telomeric repeats may also be seen in centromeric
598 regions [15,16,47,48]. We analyzed the updated *Bradysia coprophila* genome to explore the
599 evidence for or against each telomeric system in this lower Dipteran. The chromosome-scale
600 scaffolds allowed us to closely inspect the sequence at each terminus to hunt for candidate telomere
601 repeat sequences, as has been done recently by others for many insects [47,48].
602

603 As expected for Dipterans, we found no evidence of the TERT gene in the *Bradysia*
604 *coprophila* genome or proteome, and no evidence for short (5-10 bp) tandem repeats at the scaffold
605 termini, including the canonical insect pentamer. We also did not find evidence for tandem repeats
606 in the 14-22 bp range near the scaffold termini. What we did find at the scaffold termini were (i)
607 arrays of retrotransposon-related sequences with unit lengths of 1.7-3.3 kb, (ii) arrays of LCTTRs
608 (or satellites) with unit lengths of 161-191 bp, and (iii) examples of shorter tandem repeats in the
609 51-56 bp range further into the sub-telomeres. These are briefly described below with the caveat
610 that one or more of the scaffold termini may not represent the true telomere sequences.
611
612 Retrotransposon-related sequences (RTRS) were enriched near the left termini of
613 autosomal scaffolds (Fig. 7C-F – Pink triangles) and both termini of the X scaffold (Fig. 7G-H –
614 Purple triangles). The left termini of scaffolds corresponding to chromosomes II, III, and IV are
615 composed of arrays of RTRS with unit lengths of 1.7-3.3 kb. Although the RTRS repeat units have
616 different lengths, they share ~1.2 kb of sequence, and all three were classified in the repeat
617 annotation as non-LTR LINE elements related to RTE-BovB, CRE, and/or CR1. This is
618 reminiscent of *Drosophila*, which have non-LTR elements for telomeres. However, the RTRS
619 arrays found at the very ends of the left termini of II and III point away from the centromere, which
620 differs from *Drosophila*. On the left terminus of II, there are ~8 copies of a 1.7 kb RTRS spanning
621 the first ~14 kb, and on III, there are ~3-4 copies of a 3.3 kb RTRS spanning the first ~12 kb (Fig.
622 7D-E). Each unit in these arrays has a single location with BLAST hits to reverse transcriptase
623 (RT) proteins. After the RTRS array on III, the repeat annotation shows a centromere-facing LINE
624 (RTE-ORTE). The left terminus of IV has a single left-facing LINE-related sequence and
625 corresponding BLAST hit to RT proteins ~10 kb into the scaffold. This is then followed by 13-14
626 copies of a centromere-facing 2.5 kb RTRS unit, although this unit does not have the region
627 corresponding to RT protein BLAST hits despite being related to the RTRS units on II and III that
628 do (Fig. 7F). In contrast to the left termini of the autosomal scaffolds, the right termini do not have
629 RTRS arrays, at least not so close to the termini (Fig. 7I). The right terminus of II has 1-2 partial
630 hits to the RTRS units in the sub-telomeres. The right terminus of III has two related hits: the very
631 final ~100 bp matches the RTRS units, and there is a centromere-facing RT protein hit to a region
632 labeled as LINE/CR1 ~10 kb from the end. Finally, ~90-115 kb from the right terminus of IV are
633 two sub-telomeric arrays of 2 and 6 centromere-facing RTRS units highly similar to those at the
634 left terminus of IV. Of note to *Sciara* researchers, the 1.7-3.3 kb RTRS units do not appear to be
635 related to the very abundant non-LTR ScRTE sequence and its corresponding F4 probe studied by
636 Escribá et al [45], which were not found within 10 kb of any terminus, typically having a first
637 instance 60-440 kb away from termini, much farther inward than the RTRS arrays described above.
638 The left and right termini of the scaffold corresponding to the X chromosome also had
639 retrotransposon-related sequence, but different from that on the autosomes. Both sides of the X
640 scaffold start with sequence classified as a centromere-facing LTR/Pao elements, and both are
641 followed immediately by terminus-facing DNA transposons (Fig. 7G-H). LTR elements would be
642 surprising candidates for telomere sequences compared to *Drosophila*, which depends on the
643 mechanism associated with non-LTR elements for telomere maintenance [16], and may indicate
644 that the X telomeres are not well-represented on this scaffold.
645
646 Long Complex Terminal Tandem Repeats (LCTTRs or satellite repeats) were also enriched
647 near the right termini of autosomal scaffolds (Fig. 7I; Table 8). Specifically, the right termini of
648 scaffolds corresponding to chromosomes II and III end with ~28 copies of a 175 bp satellite (Fig.

649 7I – Orange triangles; Table 8) and ~28 copies of a 161 bp satellite (Fig. 7I – Red triangles; Table
650 8), respectively. The most striking example is the right terminal ~40.3 kb of IV, which contains
651 229 copies of a 176 bp satellite (Fig. 7I – Blue triangles; Table 8). Each satellite species is very
652 close to the LCTTR unit length seen in midges of 176 bp. Although each right terminus ends with
653 a different satellite, each terminus or sub-telomeric region also highlights one or more of the others
654 and/or of two other prevalent satellites with unit lengths of 191 bp and 56 bp (Fig. 7I – Brown and
655 light purple triangles, respectively; Table 8). For example, the right terminus of II also has an array
656 of the 191 bp (brown) satellite, the right-terminal sub-telomeric regions of all chromosome
657 scaffolds have arrays of the 56 bp (purple) satellite, and arrays of the 161 bp (red) satellite are
658 found in the right-side sub-telomeric regions of scaffolds for X, II, and IV. The scaffold for
659 chromosome IV shows the richest examples of satellite arrays in its sub-telomeric regions, both
660 the left and right termini of which have arrays of four or more of these satellite species. The left
661 side of IV in particular is super-enriched for the 175 bp (orange) satellite that was seen at the right
662 terminus of II. Interestingly, all RTRS units above have 1-4 partial hits to the 176 bp (blue) satellite
663 sequence that was found in a long array concluding the right terminus of IV.
664
665 Centromeric and pericentromeric regions on the chromosome-scale scaffolds are also
666 enriched for the termini-associated satellites (Fig. 7J). For example, the 161 bp (red) satellite was
667 found in the centromeric regions of all three acrocentric chromosome scaffolds: X, II, and III. In
668 fact, the 161 bp (red) satellite was tightly associated with arrays of the 143-155 bp centromeric
669 sequence, Sccr (Fig. 7J, Yellow triangles), on the X scaffold, and also turned out to be part of the
670 Sccr-linked repeat family that helped identify centromeric positions in scaffolds, especially for
671 chromosome II (Fig. 7A; "Sccr family"). The centromeric region of II did not have strong
672 representation from Sccr (likely to be an assembly artifact), but had 47 copies of the Sccr-
673 associated 161 bp (red) satellite and multiple arrays of the 175 bp (orange) and 176 bp (blue)
674 satellites. The only metacentric scaffold, IV, had massive arrays of Sccr, but was depleted for
675 termini-associated satellites, although it had 131 tandem copies of the 191 bp (brown) satellite in
676 the pericentromeric region ~75 kb away from the closest Sccr repeats.
677
678
679 **Known biological features of the X chromosome further confirm chromosome identity and**
680 **structural accuracy.**
681
682 There are three regions along the X chromosome where it folds back on itself such that
683 these loci physically interact, leaving the X looking like a loop with its ends knotted up together
684 on polytene spreads [22,23,30,31,49]. For convenience, Figure 8 shows examples of the X
685 polytene chromosome doing this (Fig. 8A-B), but also see previous works [22,23,31,45,49]. These
686 regions have been referred to as "repeats on the X" [31] or "repeat regions R1, R2, and R3" [22,49].
687 The richest descriptions of them simply refer to their banding pattern when polytenized as "three
688 three-band non-adjacent pattern repeats" [22]. However, whether these regions truly are repeats of
689 each other at the level of DNA sequence or whether they even contain repetitive DNA sequences
690 is yet unknown. Therefore, we refer to the three "repeat" loci on the X here simply as the "Fold-
691 Back Regions" (FBRs 1-3). Each locus is reported to be composed of three polytene bands, which
692 indicates they correspond to DNA lengths of tens to hundreds of kb, assuming similar band lengths
693 as *Drosophila* [50]. The Fold-Back Regions have been mapped cytogenetically to the zones within
694 the X polytene map (Fig. 8C), and are all downstream from the centromere, with FBR1 being

closest and FBR3 being furthest away [22,23,30,49]. Specifically, slightly downstream of the centromere, which is part of locus X/1A, is FBR1 at locus X/1C. FBR1 is followed within a relatively short distance by FBR2 at locus X/4A. A relatively longer distance separates FBR2 from FBR3 at locus X/12C, which is closest to the distal (non-centromeric) end of the chromosome that ends at locus X/14C (Fig. 8C).

There is a variant of the X chromosome called the X' that is found only in female-producing females [1,2,10,22,49,51–54], which also folds back on itself via the fold-back region interactions. The X' has long been known to differ from the X by a long paracentric inversion that reverses the order of the fold-back regions such that FBR3 is centromere-proximal and FBR1 is most distal (Fig. 8C). An example X'X polytene folding back on itself via FBR interactions is specifically shown in Fig. 8B; note the lack of synapsis between the chromatids from different homologs (X and X') across most of chromosome length corresponding to the long inversion compared to the XX polytene in Fig. 8A. The breakpoints that correspond to the long paracentric inversion have been mapped cytogenetically on the polytene zones upstream of FBR1 and downstream of FBR3 (Fig. 8C) [22,49], and more recently have been explored bioinformatically using sequencing data from X'X females and the Bcop_v2 X chromosome scaffold reported here [10]. Using images of the X' breakpoints and FBRs [22] with polytene maps [30], we estimate that (i) the 5' breakpoint is downstream of both the centromere and the so-called dark "landmark" band [22,45] in an interband quite close to FBR1 (X/1C), likely at the end of X/1B or the beginning of X/1C, and (ii) the 3' breakpoint occurs downstream of FBR3 (X/12C), likely at the end of X/13C or beginning of X/14A (Fig. 8C).

If the Fold-Back Regions physically interact in the X chromosomes of male pupae (for which we collected Hi-C data), then they should appear on the Hi-C map of the X chromosome scaffold as three off-diagonal dots with high frequency long-range interactions. Indeed, the Hi-C map of the X chromosome shows dots corresponding to the frequent interactions of three distal loci in three-dimensional space (Fig. 8D). Moreover, Hi-C data from X'X adult females [10] shows the same three high frequency long-range interactions along the X scaffold, but also has two additional apparently-long-distance interactions corresponding to the breakpoints of the long inversion (Fig. 8E). Importantly, the locations of and nucleotide distances between the centromere, FBRs 1-3, and the X' breakpoints follow expectations based on the polytene maps (Fig. 8D-E, Table 9, see Methods). FBR1 is expected to be within ~3.4 Mb (~2 sub-zones) of the centromere and up to ~839 kb (~1/2 sub-zone) downstream of the centromere-proximal X' breakpoint. Indeed, putative FBR1 is ~3.49 Mb and ~620 kb downstream from the centromere and inversion breakpoint, respectively (Table 9). Moreover, FBR1 and FBR2 are expected to be separated by ~11.8 Mb (~7 sub-zones), and FBR2 and FBR3 are expected to be separated by 43.6 Mb (~26 sub-zones). The observed distances are 10.3 Mb and 42.7 Mb, respectively (Table 9). The 3' breakpoint is expected to be ~5 Mb downstream (~3 sub-zones) of FBR3, and was found to be 4.65 Mb downstream. Finally, FBR3 was predicted to be ~10.1 Mb upstream of the distal end of chromosome X, and the observed scaffold location is 12.3 Mb upstream (Table 9). In all cases, the observed distances and locations were exceptionally close to projections based on known polytene map distances and locations. Thus, it is extremely likely that the three distal loci that physically interact at high frequencies according to Hi-C data are indeed the three Fold-Back Regions, which have not been previously known at the sequence level. The Hi-C interactions in male pupae and female adults indicate that the physical interactions of the X chromosome seen in polytene spreads

741 are not *in vitro* artifacts, nor are the interactions limited to polytenes in larval salivary glands, as
742 they now have been observed *in vivo* across different tissues and life stages in both sexes (male
743 pupae and female adults). Overall, the FBR and X' breakpoint locations further confirm the
744 identity and structural accuracy of the X chromosome scaffold.
745
746 The three Fold-Back Regions on the X chromosome have been described as repeats in
747 cytogenetic studies [22,23,31,49]. However, the sequences were previously not known, and
748 whether they are composed of repeats has not been well-established. Given the likeliness that the
749 three loci on the X with long-range Hi-C interactions correspond to the three FBRs, their scaffold
750 locations and sequences were extracted and analyzed. FBR scaffold locations were discovered by
751 calling long-range Hi-C interaction peaks, after first filtering out short interactions less than 5 Mb
752 (Fig. 8F-G; see Methods). The length of each FBR was in the ~103 kb to ~1.1 Mb range with
753 interaction frequency summits in the 34-69 kb range, depending on how boundaries were defined
754 (see Methods). To maximize the chance of detecting long stretches of similarity among these
755 regions, we used the longest estimates. When directly comparing the extracted sequences of these
756 three loci by BLAST, there were stretches of similarity found in all three (Fig. 8H). Although some
757 matches reached the kb range, most stretches were short (< 200 bp), and most or all were associated
758 with repeats found elsewhere in the genome. The FBRs did not have shared repeat families specific
759 to these regions. Looking at the distribution of repeats across the X chromosome, it is plausible
760 that the first locus, putative FBR1, is embedded in a highly repetitive region, but less likely for the
761 two downstream loci (putative FBRs 2 and 3), which appear to be in relatively unique regions (Fig.
762 7A-B, Fig. 8G). When comparing the entire X chromosome to itself using dotplots, if these regions
763 had high similarity the length of up to three polytene bands, there would be off-diagonal "dots" on
764 the dotplots corresponding to the locus pairs analogous to that seen in the Hi-C interaction
765 frequency plots. However, such long stretches of similarity are not seen in those off-diagonal
766 locations (Fig. 7B). Overall, the evidence so far suggests that although the three FBR loci interact,
767 it is not necessarily due to vast stretches of homology that would span 1-3 polytene bands (tens to
768 hundreds of kb). Although these three interacting regions are not necessarily three repeated loci,
769 this does not exclude the possibility that one or a few of the shorter stretches shared between the
770 regions are important aspects of their physical interactions. In contrast to FBRs 2 and 3 being in
771 relatively repeat-poor regions of the X chromosome, the X' inversion breakpoints are both buried
772 in repetitive domains corresponding to peri-centromeric and sub-telomeric regions, which do share
773 more similarity with each other (Fig. 7A-B, Fig. 8G). The 5' breakpoint and nearby putative FBR1
774 are embedded in a repeat block seen on the X scaffold dotplot near the centromeric (left) end of
775 the chromosome that has high similarity with the right terminal repeat block (Fig. 7A-B, Fig. 8G).
776 The presence of the FBRs on both the X and X' suggest the FBRs and their interactions preceded
777 the inversion, and it is possible that the evolution of long paracentric inversion on the X' was
778 mediated by the systematic fold-back interactions of FBR1 and FBR3 bringing these peri-
779 centromeric and sub-telomeric repeat blocks close together.
780
781 In addition to repeats, there are also genes within the boundaries of all three putative FBR
782 loci (as well as for the X' inversion breakpoints). Thus, it is also possible that the long-range fold-
783 back interactions may be involved in regulation of a gene or genes in the fold-back regions, or that
784 these regions each have representatives from a shared gene family. To maximize the sensitivity of
785 finding a gene family with members in all three regions, we focused again on the broader region
786 boundaries, not the narrowest estimates. Depending on the gene annotation set and FBR

boundaries used, there are 85-203 genes across all three FBRs with 35-78 in FBR1, 29-80 in FBR2, and 21-45 in FBR3. Regarding whether there are homologous genes in these regions, some of the stretches of nucleotide similarity shared by all three FBRs do overlap gene boundaries, but almost exclusively in introns, with very few exon hits and no coding sequence (CDS) overlaps. In agreement, there appears to be no homologous protein family represented in all three regions when analyzing protein sequence similarity with BLASTP. The regions of similarity at the nucleotide level within introns are largely from transposons and other repeats. Nevertheless, we cannot exclude the possibility that some stretches of similarity in the introns or near genes in general are regulatory elements shared by the FBRs. It is also possible that FBR-specific sequences, not found elsewhere in the genome, are important. There are millions of FBR-specific kmers of various sizes (15-80 bp) not found elsewhere in the genome. These kmers correspond to on the order 0.5-1 Mb of sequence from hundreds to thousands of sub-regions of the FBRs up to hundreds of bp in length, which also lack BLAST hits elsewhere in the genome. Approximately 70% of this sequence is within genes and a partially-overlapping ~16-17% corresponds to highly diverged repeats, but ~23.5% is intergenic and not labeled as a repeat. Thus, there may be intergenic (and genic) FBR-specific nucleotide elements that are found in one or more of the FBRs, but not elsewhere in the genome. Overall, it will be exciting to further elucidate in future studies what sequences and proteins mediate the long-range fold-back interactions, what function the long-range interactions serve, whether these regions and interactions are involved in other aspects of X chromosome biology, and if these interactions play a role in facilitating structural rearrangements on the X and X' chromosomes.

Discussion

Bradysia coprophila, a dark-winged fungus gnat from the lower Diptera (Nematocera), has interesting chromosomal biology that includes developmentally-programmed examples of elimination, non-disjunction, polytenization, intrachromosomal amplification, and maternally-acting sex chromosomes [1,2]. It also presents opportunities for studying sex chromosome evolution [10,55] and horizontal gene transfer [6–8]. A crucial part of the toolbox for studying these phenomena is a genome sequence. The previous reference genome sequence for *Bradysia coprophila*, Bcop_v1, was a high quality, high contiguity assembly of the somatic genome (chromosomes X, II, III, IV) produced from PacBio RS II and Oxford Nanopore MinION long read data from male embryos, and subsequently scaffolded with optical maps of male pupae DNA molecules from BioNano Genomics [9]. Bcop_v1 was produced through an extensive evaluation process comparing nearly 100 assemblies [9]. Moreover, as *Bradysia coprophila* males have only one copy of the X chromosome, but two copies of autosomes, all contigs in Bcop_v1 were classified as X-linked or autosomal [9]. Finally, two gene annotation sets were constructed for Bcop_v1: we produced one using Maker2 [9,33,34] and NCBI produced one for RefSeq [9,35]. However, Bcop_v1 was not chromosome-scale and mostly not anchored into chromosomes, limiting its usefulness in studying the evolution and behaviors of *Bradysia coprophila* chromosomes.

Our goal in this project was to improve the reference genome assembly for *Bradysia coprophila* to better-facilitate studies of its interesting chromosome behaviors and evolution. To do so, we upgraded Bcop_v1 [9] to chromosome-scale scaffolds (Bcop_v2) using Hi-C technology. The updated assembly of the *Bradysia coprophila* somatic genome (Bcop_v2) represents each chromosome (X, II, III, IV) as its own scaffold for the first time, and is oriented in the same direction as polytene maps produced in the 1960s [30] to be consistent with historical research. The chromosomal identities of the scaffolds were established and supported several ways, including multiple anchor sequences, chromosome lengths, depth of coverage, centromere positions, and other idiosyncrasies of the X chromosome. The accuracy of the chromosomal structure of the scaffolds was also supported several ways, including the concordance of anchor sequence locations along the scaffolds with their locations along polytene maps, the relative positioning of centromeres and telomeres, the increasing enrichment of repeats with proximity to centromeric and telomeric regions, and other landmarks on the X chromosome (FBRs and X' breakpoints). As studies are ongoing across multiple research groups with the gene annotation sets produced for Bcop_v1, we lifted over these gene annotation sets directly onto the chromosome-scale scaffolds so researchers can seamlessly transition ongoing studies from Bcop_v1 to Bcop_v2. The chromosomal specificity of the genes within each chromosome scaffold was further supported by comparative genomics, which showed that the chromosomes of more closely related species were far more similar in terms of gene sets than more distantly related species, both visually and using entropy statistics, the latter of which also reproduces expected phylogenetic species groupings. The totality of evidence overwhelmingly suggests that Bcop_v2 contains high quality scaffolds of the four somatic *B. coprophila* chromosomes (X, II, III, IV).

The rich history of studying *B. coprophila* chromosomes cytogenetically with respect to polytene map zones [30] has been a major benefit in producing and interpreting these chromosome-scale scaffolds. For example, previous *in situ* hybridization studies that identified the

857 chromosomal addresses of cloned DNA sequences [17–29] has allowed us to identify the
858 chromosomal identity of each chromosome-scale scaffold, orient them according to the polytene
859 maps, test their structural accuracy, map their centromeric regions, and even to study known
860 features of the X corresponding Fold-Back Regions and breakpoints of a long paracentric inversion
861 on a variant of the X chromosome called the X' (X prime). Translating the rich history of
862 cytogenetic observations could be aided by more extensively annotating the polytene map zones
863 within the Bcop_v2 chromosome-scaffold sequences. In the future, the polytene zones can roughly
864 be mapped using a combination of landmark sequences identified by FISH, genomic profiling
865 techniques, and assumptions about the sizes of the polytene zones. It may also be possible to re-
866 purpose the Hi-C data reported here to identify the boundaries of bands and interbands seen in
867 polytenes, which would help interpret the boundaries of the historical polytene zones and sub-
868 zones. Polytene bands and interbands have been associated with topologically associated domains
869 (TADs) in *Drosophila* polytene chromosomes using Hi-C data [56,57]. The locations and
870 sequences of the polytene zones could help re-interpret older cytogenetic studies through a
871 genomic lens. Regardless, the male (XO) pupae Hi-C paired-end reads will continue to be useful
872 for correcting mis-assemblies and scaffolding future *Bradysia coprophila* genome assemblies
873 produced with newer long-read datasets (e.g. PacBio HiFi or ultra-long Q20 nanopore reads).
874
875 As Bcop_v2 is derived from scaffolding Bcop_v1, there were no significant changes to
876 conclusions about the types and amounts of repeats reported previously [9]. However, the repeats
877 in Bcop_v1 are scattered across hundreds of unanchored contig sequences whereas the scaffolds
878 of Bcop_v2 allow us to see where the repeats are located within the chromosomes. We found
879 recognizable repeat families concentrated around peri-centromeric and sub-telomeric regions with
880 different propensities. Retrotransposon families in general, and especially those classified as LINE
881 (non-LTR), were very enriched in pericentromeric regions, including the LINE/RTE sub-family
882 as shown to be enriched around centromeres previously [45]. Despite a major bias towards
883 centromeric regions, retrotransposon related sequences and reverse transcriptase BLAST hits are
884 found throughout scaffold terminal regions as well, including within 10 kb of most termini.
885 Retroelements have also been shown to be enriched in functional *Drosophila* centromeres [58],
886 and to make up and play an important role in maintaining *Drosophila* telomeres [15,16]. Helitron
887 elements were enriched in both terminal and centromeric regions. DNA transposon families were
888 biased more towards scaffold terminal regions, although they are likely sub-telomeric as they do
889 not appear as terminal sequences. Ultimately, defining the repeats that make up *B. coprophila*
890 telomeres and how those telomeric repeats are maintained is beyond the scope of this paper.
891 However, we checked for signatures of known telomeric systems. As with all Diptera, Telomerase
892 Reverse Transcriptase (TERT) was not found in the *B. coprophila* genome sequence nor proteome,
893 nor was the associated hexamer/pentamer (TTAGGG/TTAGG) or other variant found at scaffold
894 ends. Non-LTR retrotransposons (HeT-A, TART, TAHRE) make up and maintain Drosophila
895 telomeres [15,16], and TART copies were also found near the chromosome ends of more closely
896 related species, *Rhynchosciara americana* [16]. Although all these transposons are possibly too
897 far diverged to find matches in the terminal regions of *B. coprophila* chromosome scaffolds
898 (Bcop_v2), reverse transcriptase genes can certainly be found in the sub-telomeric regions, even
899 within the first 10 kb. Moreover, the left termini of the *B. coprophila* autosomal scaffolds all seem
900 to have arrays of retrotransposon-related sequences (RTRS) with units in the 1.7-3.3 kb range and
901 both termini of the X scaffold were associated with retrotransposons in the repeat annotation.
902 *Rhynchosciara* was also shown to have a "special type of short tandem repeats" (stSTRs) of 14-

903 22 bp [16]. Given their close evolutionary relationship, *B. coprophila* might be expected to have
904 the same system. However, we did not find short repeats at the *B. coprophila* chromosome scaffold
905 termini, although tandem arrays of short repeats may be found further inward in the sub-telomeric
906 regions. It is also possible *B. coprophila* has Long Complex Terminal Tandem Repeats (LCTTRs)
907 like those found in Chironomidae and Culicidae [15,16]. Indeed, we found a set of satellite
908 sequences with unit lengths of 161-176 bp associated with the right termini of autosomal scaffolds.
909 These terminal and sub-terminal satellites were also seen in peri-centromeric regions as was also
910 seen in some midges [15]. We note that the repeat lengths of these satellites are similar to
911 nucleosome spacing lengths. For example, the length of DNA per nucleosome in Drosophila is
912 ~175 bp [59], which includes the linker region, and two of these candidate repeats are exactly 175-
913 176 bp. Interestingly, all 1.7-3.3 kb RTRS units above have 1-4 partial hits to the 176 bp satellite
914 sequence, suggesting an ancient relationship between these satellites and transposons. This is also
915 reminiscent of the cycling between retroelements and satellite DNAs found in the centromeres and
916 telomeres across several *Drosophila* species [60]. Regardless, our current model for *B. coprophila*
917 telomeres based on the termini of these chromosome-scale scaffolds would appear to need to
918 include arrays of both RTRS and LCTTR satellites. It is possible that the centromere-proximal
919 (left) telomeres follow a different model than their distal (right) counterparts. In these scaffolds it
920 appears that autosomal centromere-proximal termini, or all autosomal left termini, are correlated
921 with RTRS arrays whereas the centromere-distal (or right side) termini are more correlated with
922 LCTTRs. In insects, different telomere sequences have been found on different chromosomes, and
923 even on the opposite sides of the same chromosome [15,16,47,48]. However, one or more of the
924 *B. coprophila* scaffold termini may not be representative of the true telomeric sequence, and a
925 definitive conclusion cannot yet be made for specific telomeres. The X scaffold in particular needs
926 further scrutiny and refinement as so much of the intriguing chromosome biology in *B. coprophila*
927 involves the short arm of the X chromosome specifically. In addition to experimental tests, such
928 as in situ hybridization, newer PacBio HiFi and Q20+ ultra-long nanopore data will also shed more
929 light on the telomere sequences in future work.
930
931 The chromosome-level nature of Bcop_v2 will allow researchers to design experiments
932 and analyze data in a chromosome-specific way. The location and sequences of the telomeres and
933 centromeres are now more obvious and open to future investigation. The X chromosome scaffold
934 will prove useful in identifying the Controlling Element, an rDNA-associated sequence embedded
935 in the short arm of the X upstream of the centromere. Moreover, the X chromosome scaffold has
936 already benefited studies of the X chromosome variant, X-prime (X'), which differs from the X
937 by a large paracentric inversion [10]. There are also interesting biological signals in the Hi-C maps
938 that can be explored further. The notable example here are the three dots on the X chromosome
939 Hi-C map that appear in the expected pattern of the three fold-back regions (FBRs) identified by
940 Crouse [22,23,31,49]. Now that the fold-back region sequences have been identified within the
941 scaffolds, it will be possible to begin defining what mediates their interactions and what function
942 is served by doing so. Finally, the *B. coprophila* chromosome scaffolds will allow proper studies
943 of synteny and chromosome evolution across closely and distantly related flies, and may prove
944 useful for researchers working with closely related species if they are working with fragmented
945 genome assemblies by letting them cluster the contigs into likely chromosome groups, and even
946 ordering and orienting them into pseudo-chromosome scaffolds.
947

948 **Methods**
949
950 **Male Pupae Collection**
951 The dark-winged fungus gnat, *Bradysia coprophila*, was previously referred to as *Sciara*
952 *coprophila* in chromosomal and molecular biology research papers since the 1920s [1,2,61], and
953 is also known by other names such as *Sciara tilicola* and *Sciara amoena* [2]. In this study, the
954 fungus gnats were from the HoLo2 line maintained in the International *Sciara* Stock Center at
955 Brown University (https://sites.brown.edu/sciara/). *B. coprophila* females are monogenic,
956 meaning they have either only male or only female offspring, which is determined by whether they
957 harbor a variant of the X chromosome (X') or not. Specifically, X'X females are female producers
958 and XX females are male producers. The Holo2 line has a phenotypic marker gene on the X'
959 chromosome called *Wavy* [1,53]. Female producers (X'X) have the *Wavy* wing phenotype whereas
960 male producers (XX) have wild-type straight wings. To collect only male pupae, crosses between
961 straight-winged females (XX) and males (XO) were used to obtain strictly male progeny.
962 Specifically, late-stage larvae from male-only matings were transferred from mating vials to agar
963 petri plates where food and debris were removed from them, and where they entered pupation
964 asynchronously over 2-4 days. The mixed-stage pupae were transferred to Eppendorf tubes (90-
965 100 per tube), washed twice with 2X SSC, and then were flash frozen in liquid nitrogen and stored
966 at -80˚C until needed. Approximately 1,000 frozen male pupae collected across 11 tubes were
967 shipped to Phase Genomics (Seattle, WA) on dry ice where they used the material as needed to
968 prepare a Phase Genomics Proximo Hi-C sequencing library.
969
970 **Hi-C Data Collection**
971 The frozen male pupae were ground into a fine powder. Chromatin conformation capture
972 data was then generated using a Phase Genomics (Seattle, WA) Proximo Hi-C 2.0 Kit, which is a
973 commercially available version of the Hi-C protocol [11]. Following the manufacturer's
974 instructions, intact cells from two samples of finely ground frozen male pupae were crosslinked
975 using a formaldehyde solution, digested using the Sau3A1 (^GATC) restriction enzyme, and
976 proximity-ligated with biotinylated nucleotides. This creates chimeric molecules made from DNA
977 fragments that came close together *in vivo*, but that may not be close together along the genome
978 sequence. As instructed by the manufacturer's protocol, the chimeric molecules were pulled down
979 with streptavidin beads and processed into an Illumina-compatible sequencing library. Sequencing
980 was performed on an Illumina NovaSeq, generating a total of 115.2 million 2x101 bp paired-end
981 reads.
982
983 **Hi-C assembly correction**
984 The input assembly to Hi-C correction was Bcop_v1 [9], a mega-base scale, optical-map-
985 scaffolded long-read assembly. Although Bcop_v1 consists of both scaffolds (two or more contigs
986 joined by optical maps) and singleton contigs, here and throughout, the term "contigs" is used
987 generically for both types of sequences from Bcop_v1, especially in the context as input to Hi-C
988 guided contig correction and scaffolding. The Hi-C data was processed, as described below,
989 according to Phase Genomics recommendations [62]. To produce a corrected assembly, the reads
990 were aligned to the 205 primary contigs of Bcop_v1 [9] using BWA-MEM [63,64] with the -5SP
991 and -t 8 options specified, and all other options default. The output was piped into SAMBLASTER
992 [65] to flag PCR duplicates (to later exclude from analyses) and subsequently piped into SAMtools
993 [66] using "-F 2304" to remove non-primary and secondary alignments. Juicebox [67–69] was

994 used to manually identify and break contigs at putative mis-joined regions based on disruptions in
995 the expected pattern from Hi-C alignments along the contigs [11,67–69]. Such putatively mis-
996 joined regions were cut out and marked as "debris". The cut-out debris ranged from ~1.5-50 kb.
997 Regarding the number of mis-joins detected and corrected (see Results), we leaned toward over-
998 correction rather than under-correction at this step since the Phase Genomics Proximo Hi-C
999 scaffolding in the next step can "repair" these breaks by joining the sub-contigs back together if
1000 breaking them apart was in error.
1001
1002 **Hi-C assembly scaffolding**
1003 The paired-end Hi-C reads were then aligned to Bcop_v1_corrected using the same
1004 alignment procedure as above, and these alignments were the input to the Phase Genomics'
1005 Proximo Hi-C genome scaffolding platform, which was used to create chromosome-scale scaffolds
1006 from the Bcop_v1_corrected as described in Bickhart et al [13]. As in the LACHESIS method
1007 [14], the Proximo scaffolding process computes an interaction frequency matrix from the aligned
1008 Hi-C read pairs that is normalized by the number of DPNII restriction sites (^GATC) on each
1009 contig. Proximo then constructs scaffolds by optimizing the expected interaction frequencies and
1010 other statistical patterns in Hi-C data. Using a brute-force approach, approximately 20,000 separate
1011 Proximo runs were performed to optimize the number of scaffolds and the concordance with the
1012 observed Hi-C data. Finally, Juicebox [67] was again used to manually inspect for mis-join signals
1013 corresponding to scaffolding errors. The associated contigs from Bcop_v1 that were not included
1014 during the Hi-C scaffolding process were added back. Note that we also tried including the
1015 Bcop_v1 "associated contigs" (haplotigs, etc) during the Hi-C scaffolding process. In doing so, we
1016 found that although the resulting scaffolds were extremely similar to those produced from using
1017 only the primary contigs, it resulted in erroneously stitching together primary and associated
1018 contigs previously identified as haplotigs rather than keeping them separate. Therefore, the results
1019 of using only the primary assembly were preferred.
1020
1021 At the end of the scaffolding and correction process, there were 58 separate sequences not
1022 placed into the four chromosome-scale scaffolds ("unplaced"): 57 unscaffolded input contigs and
1023 a short (363 kb) Hi-C scaffold of 6 input contigs. Of the 57 unscaffolded input contigs, 46 were
1024 marked as "debris" (the regions cut out during the Hi-C correction step above). The "debris" was
1025 largely comprised of optical-map gap sequences from the input assembly. Six unscaffolded debris
1026 sequences ranging from ~25-50 kb that had 100% N content were removed from the assembly
1027 (leaving 52 unplaced sequences). Furthermore, 24 unscaffolded sequences had leading or trailing
1028 NNNNN-content ranging from 1,489 - 36,461 bp, making up ~6 - 96% of the unscaffolded
1029 sequences in which they resided, that was trimmed off. Only ~19.5 kb of internal N content
1030 remained in the cleaned-up debris regions after removal and trimming of N-gap sequence. The
1031 primary assembly for Bcop_v2 consists of the four chromosome-scale scaffolds whereas the
1032 associated assembly for Bcop_v2 consists of the 52 cleaned-up unplaced sequences as well as all
1033 the "associated contigs" from Bcop_v1. For simplicity, even though the associated assembly of
1034 Bcop_v2 contains a scaffolded sequence, they are all referred to as 'associated contigs' throughout.
1035 In Bcop_v2, the retained BioNano Genomics optical map gap sequences are represented as capital
1036 N's in the assembly and are of variable estimated lengths whereas the Hi-C gap sequences are
1037 represented as lower-case n's and are always arbitrarily set to 100 bp in length.
1038
1039

1040 **Anchoring and orienting the chromosome-scale scaffolds**
1041 Sequences with known chromosomal locations based on previous *in situ* hybridization
1042 work [17–29] [17–21,23–29] were used with BLAST [70] to identify the corresponding
1043 chromosome for each scaffold (Figures 2-3, Tables 2-3). For localization of centromeric and
1044 pericentromeric sequences within the chromosome-scale scaffolds, the short Sccr sequences [45]
1045 were used to identify Sccr-enriched Repeat Families in our custom *de novo* repeat libraries learned
1046 by RepeatModeler on the previous version of this genome (Bcop_v1) [9]. Sccr and Sccr-associated
1047 Repeat Families were mapped to Bcop_v2 using RepeatMasker [71]. To visualize the coverage
1048 and lengths of Sccr-associated alignments in IGV [72], and to identify the most likely centromere
1049 positions in the scaffolds, the lengths of all Sccr-repeat-family-associated alignments over a given
1050 position were summed for each position in the genome. The proportion of bases covered by Sccr-
1051 related or F4-related sequence in 100 kb bins was also computed with BEDtools [73] and
1052 visualized in IGV [72]. Approximated centromere positions within the scaffolds were used to
1053 compare to known centromere positions within the polytene chromosomes [22,30,31,45]. The
1054 anchor sequences and centromere positions were used to orient the chromosome-scale scaffold
1055 sequences to be concordant with conventional polytene maps [30]. Polytene map orientation was
1056 the final processing step for Bcop_v2, which was deposited to NCBI (GenBank
1057 GCA_014529535.2; WGS VSDI02; release date 2023-01-04). For researchers interested in how
1058 the contigs in Bcop_v1 correspond to Bcop_v2, we have also deposited AGP and BED files
1059 mapping these two assemblies together to: https://github.com/JohnUrban/Bcop_v2.
1060
1061
1062 **Other Hi-C analyses on Bcop_v2**
1063 After correction, scaffolding, post-processing, cleaning, anchoring, polytene orientation,
1064 and otherwise all finishing steps to produce the final updated reference genome (Bcop_v2), the
1065 paired-end Hi-C reads were realigned to Bcop_v2 and interaction frequencies were computed and
1066 visualized similar to the above pipelines for correction and scaffolding: the output of mapping
1067 reads with BWA-MEM (-5SP -t 16; version 0.7.17-r1198-dirty) [63,64] was piped into
1068 SAMBLASTER (--removeDups; version 0.1.26) [65] to mark and remove duplicates, and further
1069 piped into SAMtools "view" (-F 2316; version 1.15.1 using htslib 1.15.1) [66] to filter out
1070 alignments that were marked as supplemental or secondary and/or where one or both mates did
1071 not map. BEDtools "bamtobed" (-bedpe; version v2.30.0) [73] was used to convert BAM records
1072 to "BEDPE" format. AWK was used to convert BEDPE format to Juicer's "extra short" format ,
1073 from which a HIC file (".hic") was created using JuicerTools (min. MAPQ of 10; v2.20) [69]. The
1074 HIC file was visualized in JuiceBox (v2.15) [67]. Note that the procedure described above for
1075 aligning, filtering, and visualizing Hi-C data was used to create all Hi-C map figures of Bcop_v1,
1076 Bcop_v1_corrected, and Bcop_v2 that appear in this manuscript.
1077
1078 JuiceBox [67] visualization aided manual inspection of the Hi-C signal across Bcop_v2
1079 corresponding to individual chromosomes and individual contig joins within chromosomes. BED
1080 files for optical map gaps were used to visualize their locations and help interpret drops in Hi-C
1081 signal and depth of coverage on the Hi-C maps. BED files of Bcop_v1_corrected contig locations
1082 across Bcop_v2 were used to visualize them as squares across individual chromosome-scale
1083 scaffolds for manual inspection of the Hi-C-guided contig joins. To visualize and manually inspect
1084 contig joins in IGV [72], we extracted just the interaction frequencies between the start or end of
1085 a selected contig (left-most or right-most 1 kb or 10 kb of the contig) with 1 kb (or 10 kb) bins

corresponding to the rest of the chromosome it resided on. Briefly, to do so we extracted "target-linked mates", which are Hi-C paired-end reads that had at least one of the mates mapped within the targeted bin. Target-linked mate locations (interactions) in BEDPE format were visualized as arcs or links in IGV, which all emanate from the target region bin. However, although visualizing the links this way helps see all regions that are connected to the target bin by paired reads, it vastly misrepresents the interaction frequencies. Therefore, to visualize interaction frequencies across the scaffold, BEDtools was used to count the number of target-linked mates in bins that were constructed on the contig-bearing scaffold starting from the star/stop boundaries of the target region. Since interaction frequency decays with distance, log10 was also used (after adding a pseudocount of 1 to all bins) to better visualize long-distance interactions, such that a log10 score of 0 represented 0 interactions.

Testing coverage expectations of the X chromosome compared to autosomes

Sequencing and mapping coverage across the four chromosome-scale scaffolds was visualized using a variety of orthogonal genomic datasets produced with different technologies and different life stages from previous studies [9,10]. In sum, we used Oxford Nanopore MinION and PacBio long reads from embryonic gDNA [9], BioNano Genomics long-range optical maps from male pupae gDNA [9], and Illumina paired-end reads from male and female adult gDNA [10]. Paired-end short reads were mapped as in Baird et al (2023) [10]. All long reads were mapped with Minimap2 with technology-specific settings [74]. The optical maps were aligned with Maligner [75] with the following considerations: (i) the Bcop_v2 reference genome sequence was converted to '*in silico* optical maps' (make_insilico_map with CACGAG recognition sequence); (ii) the *in silico* Bcop_v2 maps and raw optical map data were "smoothed" using smooth_maps_file with "-m 1500" and "-m 600", respectively; and (iii) smoothed optical maps were aligned to the smoothed *in silico* Bcop_v2 maps with maligner_dp ("--max-alignments 1" and otherwise default parameters). Genomic DNA sequencing coverage was computed for each dataset using BEDtools [73] and converted to bigWig using UCSC Kent Utilities [76] for visualization in IGV [72]. Coverage from long reads was computed after filtering for a minimum read length of 5 kb and a minimum MAPQ of 40. Coverage from BioNano optical maps was computed after filtering for molecules with a minimum length of 150 kb that aligned to Bcop_v2 with a maximum Maligner M-score of -10.

Gene annotations for the chromosome-scale assembly (Bcop_v2).

To lift-over gene annotation sets originally produced on Bcop_v1, we used the program designed for this task, LiftOff [77] (version v1.6.3) with Minimap2 [74] (version 2.24-r1122). Along with the Bcop_v1 and Bcop_v2 assembly FASTA files and the Bcop_v1 GFF file, the following parameters were given to LiftOff: -p 16 -u unmapped.txt -polish -flank 0.25 -a 0.9 -s 0.9. The "polished" and "unpolished" outputs were identical for the Maker2 annotation, and were identical except for one gene for NCBI, which was extended 115 bp. Manual inspection showed the "unpolished" version to be correct in terms of matching Bcop_v1. Therefore, the "unpolished" versions are provided. GFF files for both gene annotation sets were deposited to: https://github.com/JohnUrban/Bcop_v2.

To compare the lifted-over gene models to RNA-seq evidence, RNA-seq datasets were first mapped genome-wide to Bcop_v2 without reference to any gene models using STAR [78] with the following parameters: --runThreadN ${P} --genomeDir ${STARIDX} --outSAMtype BAM

SortedByCoordinate --readFilesCommand zcat --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --outSAMattributes NH HI AS NM MD --sjdbScore 1 --readFilesIn ${R1} ${R2}. Resulting BAM files were indexed with SAMTools [66]. Genome-wide coverage was computed two ways for comparison: (i) with BEDtools [73] using the genomecov (-split -bg -ibam) and sortBed functions to produce a bedGraph that was converted to bigWig using bedGraphToBigWig from UCSC Kent Utilities [76]; and (ii) with RSEM [79] utilities using rsem-bam2wig (--no-fractional-weight) to produce a Wig file that was converted to bigWig using wigToBigWig [76]. The gene model GFF files produced by LiftOff [77] and the bigWigs for RNA-seq coverage profiles across Bcop_v2 were viewed in IGV [72]. The gene model examples shown for RNA-seq coverage are as follows, adorned with Maker2 (Bcop_v1) [9,34] and NCBI RefSeq (gene-LOC) [35] identifiers, either of which can be used with the lifted-over GFFs to find the coordinates in Bcop_v2. Bcop-PALP1: Bcop_v1_g007065; gene-LOC119085384. AGO1-like: Bcop_v1_g003309; gene-LOC119067568. Aub-like: Bcop_v1_g01302; gene-LOC119068172. CYP53A57: Bcop_v1_g020678; gene-LOC119075328. CYP3718B1: Bcop_v1_g007236; gene-LOC119074427. CYP4721A1: Bcop_v1_g009136; gene-LOC119078932.

To compare male and female expression of lifted-over gene models grouped by chromosome, we used published RNA-seq data for males and females across various life stages [9]. Transcript level quantification of all genes in the NCBI annotation set [35] lifted over to Bcop_v2 was carried out by RSEM (v1.3.1) [79] coupled with STAR (version 2.7.10a_alpha_220818) [78] using the "rsem-calculate-expression" pipeline. The RSEM-calculated expected counts were used with EdgeR [80] for between-sample normalization and differential expression analysis. Plots of EdgeR log2 fold-change results were made in Python.

Using comparative genomics and evolutionary expectations to test the groupings of Bcop_v1 contigs into chromosome-scale scaffolds (Bcop_v2):

We identified orthologous relationships across all proteins within the proteomes of the five species using OrthoFinder [81,82], which also estimated the proteome-wide rooted species tree. For *B. coprophila*, we used NCBI *Bradysia coprophila* Annotation Release 100 protein set for Bcop_v1 (NCBI GenBank GCA_014529535.1; RefSeq GCF_014529535.1) [35] and the corresponding gene coordinates lifted over to Bcop_v2 described above. For the other species, we used the protein sets and genomic GTF files associated with the following NCBI accessions: *Drosophila melanogaster* (NCBI GenBank GCA_000001215.4; RefSeq GCF_000001215.4) [40]; *Pseudolycoriella hygida* (NCBI GenBank GCA_029228625.1) [42]; yellow fever mosquito (Aedes aegypti; NCBI GenBank GCA_002204515.1; RefSeq GCF_002204515.2) [41]; African malaria mosquito (Anopheles gambiae; NCBI GenBank GCA_000005575.1; RefSeq GCF_000005575.2) [39]. The proteome-wide rooted species tree was visualized using icy tree [83] by providing the Newick code returned by OrthoFinder [81,82]: "(D. mel:0.243236,((A. gam:0.26685,A. aeg:0.227034)0.954721:0.17008,(P. hyg:0.177818,B. cop:0.123002)0.969863:0.240218)1:0.243236);".

OrthoFinder [81,82] creates distinct "Ortho Groups" containing all orthologs for a given gene from all given species. Thus, one can use the genes within an Ortho Group to compare the

genomic locations of the "same" gene across all species. For simplicity and higher specificity, we used only the subset of Ortho Groups that are composed of a single ortholog from each and all of the five species, otherwise known as Single Copy Orthologs (SCOs). Thus, using the genomic location of a single gene from each species in an orthogroup of SCOs, we can make dot plots (or scatter plots), that show the location of the "same" gene in one species' genome on the X-axis and another species' genome on the Y-axis. Such dot plots were made in R using the assembalign.dotplot function from Lave [9,84] (https://github.com/JohnUrban/lave). Note that we used "conserved fly genes" in other analyses, which were *B. coprophila* genes that had one or more orthologs in all five Dipteran species, as opposed to only SCOs (which have one and only one copy in all five species).

We quantified the amount of "gene shuffling" using entropy, and implemented this analysis in Python3. Briefly, entropy can be thought of as a measure of uncertainty, information, disorder, or "mixed-up-ness", and is often used to study systems that evolve from a highly ordered to a highly disordered state or from a non-random to a random state [44]. Mathematically, entropy is defined as -1 times the sum of p*log(p) for all probabilities, p, in a set of probabilities, P. That is, -1*sum(p*log(p)). When log base 2 is used, entropy is expressed in "bits". For "gene shuffling", the initial non-random (lowest entropy) state corresponds to how genes are grouped on chromosomes within a given species. As genes get shuffled around into different chromosomal groupings across evolutionary time, they approach randomized (high entropy) states with respect to some target species (e.g., *B. coprophila*). To calculate the entropy between any two species, we used only SCOs that appeared on the chromosome-scale scaffolds of each species, which is to say we excluded those that appeared on unplaced contigs. We counted the number of SCOs that appeared on each pair of inter-species chromosomes (i.e. chromosome "i" from species A and chromosome "j" from species B), added a pseudo-count of 0.1 to the SCO count of each inter-species pair, then divided each inter-species pair SCO count by the sum of SCO counts across all pairs of inter-species chromosomes (including pseudo-counts). Note that, for *D. melanogaster and A. gambiae,* the SCO counts from the left and right arms of chromosomes 2 and 3 (2L, 2R, 3L, 3R) were summed to get the full counts for those chromosomes (2L+2R, 3L+3R). For each pair of species, this gave a table of joint probabilities of a SCO appearing on chromosome "i" from species A and chromosome "j" from species B (see Table 6 for an example). The inter-species entropy was computed from these joint probabilities using the entropy equation: -1*sum(p*log2(p)), where p is the matrix of joint probabilities. However, as each pairwise species comparison has a different lowest possible entropy state (the non-random initial state) and a different highest possible (fully random) entropy state given their SCO counts per chromosome, we needed to use min-max normalization ((X-min)/(max-min)) in order to compare entropy values across pairwise inter-species comparisons. To calculate the minimum (initial) entropy state for each species within the context of a pairwise comparison, we marginalized (i.e. summed) over the joint probabilities to get the marginal probabilities that a SCO came from each chromosome for a given species. The entropy of the marginal probabilities of both species was then computed separately, using the equation above, and averaged together to get the lowest entropy value the joint probabilities could take. When comparing a species to itself, the average simply returns the entropy of the marginal probabilities for that species. To calculate the maximum (fully random) entropy state between two species, the marginal probabilities for each chromosome of a given species were multiplied by the marginal probabilities for each chromosome of the other species. This yielded a table of fully-random joint probabilities where the probability of a SCO appearing on chromosome "i" from

species A and chromosome "j" from species B was simply the product of the marginal probability of a SCO appearing on chromosome "i" from species A and the marginal probability of a SCO appearing on chromosome "j" from species B. For the given chromosomal SCO counts for each species, this produces the joint probabilities expected at random, and maximum entropy is computed on it as above: -1*sum(p*log2(p)), where p is the matrix of random joint probabilities. The Seaborn library [85] in Python3 was used to visualize the min-max normalized entropy scores calculated against *B. coprophila* as a bar plot and all pairwise entropy scores as a clustered-heatmap.

Testing gene locations of "alien" P450 genes within the chromosome-scale scaffolds using a variety of genomic datasets

If the "alien" cytochrome P450 ("CYP") genes reported recently [7] are contained within the nuclear chromosomes, then they should exhibit expected Hi-C interaction frequencies [11] within the scaffolds. If they are external contamination erroneously stitched into the scaffold sequences, there should be severe drops in the Hi-C interaction frequency signal with these genes and the rest of the scaffold they reside in. To obtain all Hi-C interactions between a given "alien" P450 gene and all other sites in the genome, the same procedure described above for contig ends was used, here using Hi-C paired-end reads that had at least one of the mates mapped within the gene. For each "alien" P450 gene, the closest highly conserved fly gene that had a similar length (within 1% difference) was also interrogated the same way. Interactions and interaction frequencies in 100 kb bins were computed and visualized in IGV similar as described above, but with important differences. The number of interactions between a gene and the rest of the genome, and whether interactions are sparse or dense, is a function of gene length. To make valid comparisons with nearby conserved fly genes, gene length was controlled for in two ways. The first was to use nearby conserved fly genes of similar size (within 1%), as stated above. This helped match the observed interaction sparsity or density. The second was to normalize counts to each gene's length, then multiply by 10,000 such that the magnitude of counts for all genes corresponded to a gene length of 10 kb, which was chosen since the "alien" gene lengths spanned from ~1.8-7.8 kb.

Read depth can also differ significantly at sites of contamination, especially across different biological samples. Some samples may show low or no sequencing depth over sites where contamination was integrated into a genome assembly. Other samples with an abundance of the contaminant may show spikes in read depth there. It is unlikely for such sites of integrated contamination to have no disruptions in the depth signal across multiple samples. To investigate the DNA-sequencing coverage over the regions bearing "alien" P450 genes, we queried the sequencing depth (computed as above) of a diverse set of biological samples, including different sexes and developmental stages [9,10], using datasets from orthogonal technologies, including Illumina-paired end reads [10], confidently-mapped (MAPQ≥40) long (≥5 kb) Nanopore and PacBio reads [9], and stringently-mapped ultra-long (≥150 kb) optical maps from BioNano Genomics [9]. See above "coverage" Methods subsection for details.

Repeat annotations for the chromosome-scale assembly (Bcop_v2).

RepeatMasker [71] was used to re-annotate the repeats across Bcop_v2 using comprehensive repeat libraries as done previously for Bcop_v1 [9], rather than using a lift-over

process. As previously, the comprehensive repeat library contained (i) species-specific repeat libraries learned with RepeatModeler [86], (ii) previously known repeat sequences from *Bradysia coprophila* found at NCBI prior to Bcop_v1 [9], and (iii) all Arthropod repeats in the RepeatMasker Combined Database, which included Dfam_Consensus-2018/10/26 [87] and RepBase-2018/10/26 [88]. The corresponding repeat annotation GFF file for Bcop_v2 was deposited to: https://github.com/JohnUrban/Bcop_v2. RepeatMasker [71] was also used to map just the arthropod repeats alone. TandemRepeatFinder (TRF) [89] was used independently to map and visualize just tandem repeats in the genome (trf ${FA} 2 7 7 80 10 50 2000 -f -d -m). Repeat densities in 100 kb bins across the genome were computed using BEDtools [73]. Briefly, either all repeats in the RepeatMasker GFF or targeted repeats (e.g. a given family like LINEs) were extracted and converted to BED; the proportion of bases corresponding to those repeat intervals within 100 kb windows was calculated using the "coverage" BEDtool [73]. Repeat density bedGraphs were visualized in IGV [72]. Retrotransposon-related sequences and satellite sequences were found by inspecting the repeat annotation and tandem repeat finder results in the first 10 kb (or up to 100 kb) of the scaffold termini using IGV [72], and looking at dot plots of the first 10-100 kb. Their distributions across the scaffolds were further explored by BLASTing [70] them back to the assembly.

Chromosome-wide and genome-wide dot plots were made to visualize the repeat structure across the scaffolds. Briefly, Minimap2 [74] was used to map chromosomes against themselves or the entire genome against itself using parameters optimized for dot plots (-PD -k19 -w19 -m200 -t8). The PAF output files were visualized as dot plots in R using the assembalign.dotplot function from Lave [9,84], which presents the pairwise-aligned query and target intervals as line segments. BED files for the locations of anchor sequence midpoints +/- 500 kb, estimated centromere midpoints +/- 500 kb, and estimated locations of Fold-Back Regions 1-3 and the X' long paracentric inversion breakpoints were provided to annotate the plot margins for reference points.

Extracting and analyzing locus sequences corresponding to long-range Hi-C interactions on the X chromosome.

The locations of the loci on Fold-Back Regions on the X chromosome that appear as three off-diagonal dots corresponding to long-range interacting loci on the Hi-C maps, as well as the "main breakpoints" corresponding to a long paracentric inversion with the X' chromosome variant, were learned by analyzing the paired Hi-C reads that mapped to the X with long distances in between them. Two datasets were explored: male (XO) pupae data from this study as well as female X'X Hi-C data from a recent analysis of the X' chromosome [10]. Briefly, paired-end Hi-C reads (in Juicer's Extra Short format [69]) that mapped to the X chromosome using the above pipeline were read into R where map locations were further processed to filter for paired reads with (i) inter-mate distances of at least 1 Mb, and (ii) minimum mapping qualities (MAPQ) of 10, although a range of MAPQs were explored, including no filtering, and gave similar end results. The 2D Binned Kernel Density Estimate (from the 'KernSmooth' package) was then taken using a 1024x1024 grid (~68.9 kb bins) with 250 kb smoothing bandwidths on both the x- and y- axes corresponding to where mate1 and mate2 mapped across the X. During optimization and exploratory analyses, kernel-smoothed 2D grids were visualized using levelplot from the 'lattice' package. Peaks in the Z (heat) dimension of the 2D grid were called by identifying x,y-coordinates where the Z dimension was above the 99.8th percentile across the entire grid. In other words, Z-peaks corresponded to pairs of distant loci (>1 Mb) with the highest numbers of Hi-C reads

connecting them. The x,y-coordinates of peaks in the 2D grid were then hierarchically clustered, and the tree was cut to identify k=12 peak clusters on the 2D grid. The median, minimum, and maximum x- and y- coordinates of each cluster were obtained to help define the centers and the boundaries of interacting pairs of loci on the X chromosome, as was the coordinates of the peak summit bin (the bin with the highest interaction frequency in a peak cluster). The peak clusters were filtered for those whose median x-coordinates were separated by 5 Mb from the median y-coordinates. In other words, these were loci pairs separated by at least 5 Mb. Three of the four remaining peak clusters separated by at least 5 Mb corresponded to the interactions between putative FBRs 1 and 2, putative FBRs 1 and 3, and putative FBRs 2 and 3. The fourth remaining peak cluster corresponded to the breakpoint dots for the long paracentric inversion with the X'. The peak clusters for loci separated by less than 5 Mb may correspond to (i) other smaller structural variations between the X' and X, and (ii) other shorter Mb-range interactions of distal loci. The regions corresponding to the three FBRs and the two main X' breakpoints were visualized on Hi-C maps using JuiceBox [67]. Note that using inter-mate distances of 5 Mb instead of 1 Mb and k=4 clusters gave the same end results without needing further filtering. With these parameters peak clusters ranged from 551 kb to 1.1 Mb (8-16 bins) and had summit bins of 68.9 kb. We also identified parameters for much narrower estimates of FBR locations using MAPQ of 10, intermate distance minimum of 5 Mb, a 2048x2048 grid (~34.4 kb bins) with a smoothing bandwidth of 50 kb, a peak quantile threshold of 0.9999, and k=4 peak clusters, which produced four interaction peak clusters in the ~103-551 kb range (3-16 bins) with summit bin estimates of ~34.4 kb.

The nucleotide distances between the mid-points of the FBR regions, X' breakpoints, and centromeres observed on the X scaffold were compared to expected nucleotide distances given corresponding polytene map distances in proportion to the length of the X chromosome scaffold. The X polytene map is organized into 14 zones of similar lengths, each with 3 subzones (A, B, and C) for a total of 42 sub-zones. Assuming nucleotide lengths of zones are also similar, the ~70.5 Mb X-scaffold would indicate zone lengths of ~5 Mb and sub-zone lengths of ~1.68 Mb, which we used to compute expected lengths. The FBR coordinates were used to define both the sequences of the FBRs as well as the Genomic Complement (the genome sequence excluding the FBRs). FBR and Genomic Complement sequences were extracted using BEDtools [73]. JellyFish [90] was used for kmer operations on both sets of sequences. BLAST [70] was used to compare FBR sequences to each other and to the Genomic Complement sequences. FBR dotplots were made as described for the entire genome and individual chromosomes above. BEDtools [73] was used to extract and analyze genes and repeats in the FBRs. The FBR coordinates were also used to annotate the X chromosome dotplot, IGV traces showing repeats, and Hi-C interaction frequency maps.

Data Availability

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession VSDI00000000. The version described in this paper, referred to throughout as Bcop_v2, is version VSDI02000000 (deposited 2022; released Jan 4, 2023). This project is associated with BioProjects PRJNA291918 and PRJNA672144; and BioSamples SAMN12533751 and SAMN20343824. Male Pupa Hi-C data was deposited to SRA: SRR23335771. Lifted over gene model GFFs, repeat annotation GFF, *B. coprophila*-specific repeat libraries, as well as AGP and BED files mapping Bcop_v1 to Bcop_v2 were deposited to a github repository: https://github.com/JohnUrban/Bcop_v2. The RNA-seq samples from *B. coprophila* from different life stages used to analyze dosage compensation and test gene models are from Urban et al (2021) [9] and associated with BioProject PRJNA748150. The RNA-seq samples from the radiation study used to test gene models and visualize exon coverage are from Urban et al (2023) [37] and associated with BioProject PRJNA928089. The long reads from Oxford Nanopore and PacBio, and optical maps from BioNano Genomics are from Urban et al (2021) [9] and associated with BioProject PRJNA291918, and the Illumina reads from male and female adults as well as X'X Hi-C data are from Baird et al (2023) [10] and associated with BioProject PRJNA953429.

Abbreviations

AGP: a golden path (file format); Bcop: *Bradysia coprophila*; bp: base pairs; BED: Browser Extensible Data (file format); BUSCO: Benchmarking Universal Single-Copy Orthologs; CDS: coding sequence; FBR: foldback region; Gb: gigabase pairs; GFF: General Feature Format (file format); GTF: General Transfer Format (file format); Hi-C: High dimensional Chromosome conformation capture; i5k: initiative to sequence 5000 insect (and other arthropod) genomes (http://i5k.github.io/about) ; kb: kilobase pairs; LCTTR: long complex terminal tandem repeats ; LINE: long interspersed nuclear element ; LTR: long terminal repeats; Mb: megabase pairs; MMNE: Min-Max Normalized Entropy; NCBI: National Center for Biotechnology Information; PacBio: Pacific Biosciences; PCR: polymerase chain reaction; RTE : retrotransposable element; RTRS: retrotransposon related sequence; Sccr: *Sciara coprophila* centromeric repeat; ScRTE: *Sciara coprophila* retrotransposable element as defined in Escribá et al (2011) [45]; SMRT: single-molecule real-time. ONT: Oxford Nanopore Technologies.

Competing Interests

Authors declare no competing interests.

Funding

Financial support has come from National Institutes of Health (NIH) NIH/GM121455 to SAG. ACS is an Investigator of the Howard Hughes Medical Institute (HHMI) in the HHMI unit at the Carnegie Institute for Science (CIS). JMU is an HHMI/CIS Research Associate under ACS.

Authors' Contributions

JMU, SAG, and ACS conceived the project. SAG collected male pupae, coordinated with Phase Genomics, and helped with chromosome maps and map positions. Hi-C library preparation, sequencing, and scaffolding was performed as a service provided by Phase Genomics. JMU performed the original Bcop_v1 assembly, and corresponded with Phase Genomics during the optimization process for Bcop_v2 scaffolding. JMU performed all subsequent analyses and

operations, including subsequent Hi-C data processing, analysis, exploration and visualizations in all figures, chromosome identification, scaffold orientation, assembly evaluations, gene annotation lift-overs, comparative genomics and entropy analyses, HGT gene validations, repeat annotation and analyses, scaffold termini characterization, and fold-back repeat sequence extraction and characterization; JMU developed the ideas, analyses, and python/R code for Lave dot plots, for entropy/MMNE analyses, and for the long-range Hi-C interaction peak detection method to discover FBR and breakpoint locations; JMU made all figures and tables. ACS helped conceive of analyses, figures, and tables. All authors participated intellectually in the development and execution of this project. All authors wrote, read, and edited the manuscript.

Acknowledgements

We thank Jacob Bliss who helped with *Bradysia coprophila* maintenance during data collection. We thank our correspondents at Phase Genomics, including Hayley Mangelson, Kayla Young, Shawn Sullivan, Brian Fan, Andrew Wiser, and Ivan Liachko. We acknowledge and thank Rob Baird and Laura Ross for being early adopters of Bcop_v2 for collaborative studies of the X' chromosome. We thank Jennifer Urban for proof-reading the manuscript. We thank Ava Filiss for the X and X' polytene chromosome images. We thank all members of the Spradling Laboratory for feedback during lab meetings, and especially Robert Levis for discussions on telomeres. Polytene chromosome maps in figures 3, 7, and 8 were reproduced from plates 1, 2 and 3 of Gabrusewycz-Garcia (1964) [30] with permission from Springer Nature under permission number: 5490830617309. We thank bioRxiv for hosting our 2022 preprint that preceded this paper (https://doi.org/10.1101/2022.11.03.515061; version 1, Nov 2022; version 2, Dec 2024).

References

1. Gerbi SA. Unusual chromosome movements in sciarid flies. In: Hennig W, editor. *Results Probl Cell Differ*. Springer, Berlin, Heidelberg;

2. Gerbi SA. Non-random chromosome segregation and chromosome eliminations in the fly Bradysia (Sciara). *Chromosom Res 2022 302*. Springer; 2022; doi: 10.1007/S10577-022-09701-9.

3. Rasch EM. DNA cytophotometry of salivary gland nuclei and other tissue systems in dipteran larvae. In: Wied GL, Bahr GF, editors. *Introd to Quant Cytochem*. New York: Academic Press; p. 357–97.

4. Gerbi SA, Strezoska Z, Waggener JM. Initiation of DNA replication in multicellular eukaryotes. *J Struct Biol*. 140:17–302002;

5. Rasch EM. Two-wavelength cytophotometry of *Sciara* salivary gland chromosomes. In: Wied GL, Bahr GF, editors. *Introd to Quant Cytochem*. New York: Academic Press; p. 335–55.

6. Hodson CN, Jaron KS, Gerbi S, Ross L. Gene-rich germline-restricted chromosomes in black-winged fungus gnats evolved through hybridization. *PLOS Biol*. Public Library of Science; 2022; doi: 10.1371/JOURNAL.PBIO.3001559.

7. Feyereisen R, Urban JM, Nelson DR. Aliens in the CYPome of the black fungus gnat, Bradysia coprophila. *Insect Biochem Mol Biol*. Pergamon; 2023; doi: 10.1016/J.IBMB.2023.103965.

8. Chen X, Urban JM, Wurlitzer J, Wei X, Han J, O'Connor SE, et al.. Canonical terpene synthases in arthropods: Intraphylum gene transfer. *Proc Nat Acad Sci*. 2024;

9. Urban JM, Foulk MS, Bliss JE, Coleman CM, Lu N, Mazloom R, et al.. High contiguity de novo genome assembly and DNA modification analyses for the fungus fly, Sciara coprophila, using single-molecule sequencing. *BMC Genomics 2021 221*. BioMed Central; 2021; doi: 10.1186/S12864-021-07926-2.

10. Baird RB, Urban JM, Mongue AJ, Jaron KS, Hodson CN, Grewoldt M, et al.. Recent Evolution of a Maternally Acting Sex-Determining Supergene in a Fly with Single-Sex Broods. *Mol Biol Evol*. Oxford Academic; 2023; doi: 10.1093/MOLBEV/MSAD148.

11. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al.. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-)*. American Association for the Advancement of Science; 2009; doi: 10.1126/SCIENCE.1181369/SUPPL_FILE/LIEBERMAN-AIDEN.SOM.PDF.

12. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al.. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. Cell Press; 2014; doi: 10.1016/j.cell.2014.11.021.

13. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al.. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet 2017 494*. Nature Publishing Group; 2017; doi: 10.1038/ng.3802.

14. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013; doi: 10.1038/nbt.2727.

15. Kamnert I, López CC, Rosén M, Edström JE. Telomeres Terminating with Long Complex Tandem Repeats. *Hereditas*. John Wiley & Sons, Ltd; 1997; doi: 10.1111/J.1601-

1467 5223.1997.00175.X.
1468 16. Kuznetsova V, Grozeva S, Gokhman V. Telomere structure in insects: A review. *J Zool Syst*
1469 *Evol Res*. John Wiley & Sons, Ltd; 2020; doi: 10.1111/JZS.12332.
1470 17. Pardue M Lou, Gerbi SA, Eckhardt RA, Gall JG. Cytological localization of DNA
1471 complementary to ribosomal RNA in polytene chromosomes of Diptera. *Chromosoma*. Springer-
1472 Verlag; 1970; doi: 10.1007/BF00325943.
1473 18. Gerbi SA. Localization and characterization of the ribosomal RNA cistrons in Sciara
1474 coprophila. *J Mol Biol*. Academic Press; 1971; doi: 10.1016/0022-2836(71)90367-6.
1475 19. Foulk MS, Liang C, Wu N, Blitzblau HG, Smith H, Alam D, et al.. Ecdysone induces
1476 transcription and amplification in *Sciara coprophila* DNA puff II/9A. *Dev Biol*. 2006; doi:
1477 10.1016/j.ydbio.2006.07.018.
1478 20. Greciano PG, Ruiz MF, Kremer L, Goday C. Two new chromodomain-containing proteins
1479 that associate with heterochromatin in *Sciara coprophila* chromosomes. *Chromosoma*. 2009; doi:
1480 10.1007/s00412-009-0203-y.
1481 21. Foulk MS, Waggener JM, Johnson JM, Yamamoto Y, Liew GM, Urnov FD, et al.. Isolation
1482 and characterization of the ecdysone receptor and its heterodimeric partner ultraspiracle through
1483 development in *Sciara coprophila*. *Chromosoma*. 2013; doi: 10.1007/s00412-012-0395-4.
1484 22. Crouse H V.. X heterochromatin subdivision and cytogenetic analysis in *Sciara coprophila*
1485 (Diptera, Sciaridae): Centromere localization. *Chromosoma*. Springer-Verlag; 1977; doi:
1486 10.1007/BF00292941.
1487 23. Crouse H V., Gerbi SA, Liang CM, Magnus L, Mercer IM. Localization of ribosomal DNA
1488 within the proximal X heterochromatin of Sciara coprophila (Diptera, Sciaridae). *Chromosoma*.
1489 Springer-Verlag; 1977; doi: 10.1007/BF00294938.
1490 24. Kerrebrock AW, Srivastava R, Gerbi SA. Isolation and characterization of ribosomal DNA
1491 variants from *Sciara coprophila*. *J Mol Biol*. Academic Press; 1989; doi: 10.1016/0022-
1492 2836(89)90286-6.
1493 25. DiBartolomeis SM, Gerbi SA. Molecular characterization of DNA puff II/9A genes in *Sciara*
1494 *coprophila*. *J Mol Biol*. 210:531–401989;
1495 26. Bienz-Tadmor B, Smith HS, Gerbi SA. The promoter of DNA puff gene II/9-1 of *Sciara*
1496 coprophila is inducible by ecdysone in late prepupal salivary glands of *Drosophila melanogaster*.
1497 *Cell Regul*. 2:875–881991;
1498 27. Wu N, Liang C, DiBartolomeis SM, Smith HS, Gerbi SA. Developmental progression of
1499 DNA puffs in *Sciara coprophila*: amplification and transcription. *Dev Biol*. 1993; doi:
1500 10.1006/dbio.1993.1287.
1501 28. Mok EH, Smith HS, DiBartolomeis SM, Kerrebrock AW, Rothschild LJ, Lange TS, et al..
1502 Maintenance of the DNA puff expanded state is independent of active replication and
1503 transcription. *Chromosoma*. 110:186–962001;
1504 29. Urnov FD, Liang C, Blitzblau HG, Smith HS, Gerbi SA. A DNase I hypersensitive site
1505 flanks an origin of DNA replication and amplification in *Sciara*. *Chromosoma*. 2002; doi:
1506 10.1007/s00412-002-0194-4.
1507 30. Gabrusewycz-Garcia N. Cytological and autoradiographic studies in Sciara coprophila
1508 salivary gland chromosomes. *Chromosoma*. 15:312–441964;
1509 31. Crouse HV. Translocations in Sciara: their bearing on chromosome behavior and sex
1510 determination. *Univ Missouri Res Bull*. 379:1–751943;
1511 32. Rasch EM. Genome size and determination of DNA content of the X chromosomes,
1512 autosomes, and germ line-limited chromosomes of *Sciara coprophila*. *J Morphol*. 2006; doi:

1513 10.1002/jmor.10478.
1514 33. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management
1515 tool for second-generation genome projects. *BMC Bioinformatics*. 2011; doi: 10.1186/1471-
1516 2105-12-491.
1517 34. Urban JM: Bradysia coprophila genome annotations Bcop_v1.0. USDA Ag Data Commons.
1518 https://doi.org/10.15482/USDA.ADC/1522618 (2021).
1519 35. NCBI: Bradysia coprophila Annotation Release 100. NCBI Eukaryot. Genome Annot.
1520 Pipeline. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Bradysia_coprophila/100/
1521 36. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. BUSCO: assessing
1522 genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*.
1523 2015; doi: 10.1093/bioinformatics/btv351.
1524 37. Urban JM, Bateman JR, Garza KR, Borden J, Jain J, Brown A, et al.. Bradysia (Sciara)
1525 coprophila larvae up-regulate DNA repair pathways and down-regulate developmental regulators
1526 in response to ionizing radiation. *Genetics*. Oxford University Press (OUP); 2023; doi:
1527 10.1093/GENETICS/IYAD208.
1528 38. da Cunha PR, Granadino B, Perondini AL, Sánchez L. Dosage compensation in sciarids is
1529 achieved by hypertranscription of the single X chromosome in males. *Genetics*. Oxford
1530 Academic; 1994; doi: 10.1093/genetics/138.3.787.
1531 39. Sharakhova M V., Hammond MP, Lobo NF, Krzywinski J, Unger MF, Hillenmeyer ME, et
1532 al.. Update of the Anopheles gambiae pest genome assembly. *Genome Biol*. 2007; doi:
1533 10.1186/gb-2007-8-1-r5.
1534 40. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al.. The Release 6
1535 reference sequence of the Drosophila melanogaster genome. *Genome Res*. Cold Spring Harbor
1536 Laboratory Press; 2015; doi: 10.1101/gr.185579.114.
1537 41. Matthews BJ, Dudchenko O, Kingan SB, Koren S, Antoshechkin I, Crawford JE, et al..
1538 Improved reference genome of Aedes aegypti informs arbovirus vector control. *Nature*. Nature
1539 Publishing Group; 2018; doi: 10.1038/s41586-018-0692-z.
1540 42. Trinca V, Carli S, Uliana JVC, Garbelotti CV, Mendes da Silva M, Kunes V, et al..
1541 Biocatalytic potential of Pseudolycoriella CAZymes (Sciaroidea, Diptera) in degrading plant and
1542 fungal cell wall polysaccharides. *iScience*. Elsevier; 2023; doi: 10.1016/J.ISCI.2023.106449.
1543 43. Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim JW, Lambkin C, et al.. Episodic
1544 radiations in the fly tree of life. *Proc Natl Acad Sci U S A*. National Academy of Sciences; 2011;
1545 doi: 10.1073/pnas.1012675108.
1546 44. Seitz W, Kirwan AD. Mixed-Up-Ness or Entropy? *Entropy 2022, Vol 24, Page 1090*.
1547 Multidisciplinary Digital Publishing Institute; 2022; doi: 10.3390/E24081090.
1548 45. Escribá MC, Greciano PG, Méndez-Lago M, De Pablos B, Trifonov VA, Ferguson-Smith
1549 MA, et al.. Molecular and cytological characterization of repetitive DNA sequences from the
1550 centromeric heterochromatin of *Sciara coprophila*. *Chromosoma*. 2011; doi: 10.1007/s00412-
1551 011-0320-2.
1552 46. Smith CD, Edgar RC, Yandell MD, Smith DR, Celniker SE, Myers EW, et al.. Improved
1553 repeat identification and masking in Dipterans. *Gene*. Gene; 2007; doi:
1554 10.1016/J.GENE.2006.09.011.
1555 47. Lukhtanov VA. Diversity and evolution of telomere and subtelomere DNA sequences in
1556 insects. *bioRxiv*. Cold Spring Harbor Laboratory; 2022; doi: 10.1101/2022.04.08.487650.
1557 48. Lukhtanov VA, Pazhenkova EA. Diversity and evolution of telomeric motifs and telomere
1558 DNA organization in insects. *Biol J Linn Soc*. Oxford University Press (OUP); 2023; doi:

10.1093/BIOLINNEAN/BLAD068.
49. Crouse H V.. X heterochromatin subdivision and cytogenetic analysis in *Sciara coprophila* (diptera, sciaridae): The Controlling Element. *Chromosoma*. Springer-Verlag; 1979; doi: 10.1007/BF00292274.
50. Zhimulev IF, Belyaeva ES, Semeshin VF, Koryakov DE, Demakov SA, Demakova OV, et al.. Polytene Chromosomes: 70 Years of Genetic Research. *Int Rev Cytol*. Academic Press; 2004; doi: 10.1016/S0074-7696(04)41004-3.
51. Metz CW. Sex Determination in Sciara. *Am Nat*. 63:487–961929;
52. Metz CWW, Schmuck MLL. Studies on sex determination and the sex chromosome mechanism in Sciara. *Genetics*. Genetics Society of America; 16:225–531931;
53. Metz CW, Smith HB. Further Observations on the Nature of the X-Prime (X') Chromosome in Sciara. *Proc Natl Acad Sci U S A*. National Academy of Sciences; 17:195–81931;
54. Metz CWW. Chromosome Behavior, Inheritance and Sex Determination in Sciara. *Am Nat*. 72:485–5201938;
55. Baird RB, Mongue AJ, Ross L. Why put all your eggs in one basket? Evolutionary perspectives on the origins of monogenic reproduction. *Hered 2023 1312*. Nature Publishing Group; 2023; doi: 10.1038/s41437-023-00632-7.
56. Eagen KP, Hartl TA, Kornberg RD. Stable Chromosome Condensation Revealed by Chromosome Conformation Capture. *Cell*. Cell Press; 2015; doi: 10.1016/J.CELL.2015.10.026.
57. Spradling AC. Polytene Chromosome Structure and Somatic Genome Instability. *Cold Spring Harb Symp Quant Biol*. Cold Spring Harbor Laboratory Press; 2017; doi: 10.1101/sqb.2017.82.033670.
58. Chang CH, Chavan A, Palladino J, Wei X, Martins NMC, Santinello B, et al.. Islands of retroelements are major components of Drosophila centromeres. *PLOS Biol*. Public Library of Science; 2019; doi: 10.1371/JOURNAL.PBIO.3000241.
59. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, et al.. Nucleosome organization in the Drosophila genome. *Nature*. Nature Publishing Group; 2008; doi: 10.1038/nature06929.
60. Courret C, Hemmer LW, Wei X, Patel PD, Chabot BJ, Fuda NJ, et al.. Turnover of retroelements and satellite DNA drives centromere reorganization over short evolutionary timescales in Drosophila. *PLOS Biol*. Public Library of Science; 2024; doi: 10.1371/JOURNAL.PBIO.3002911.
61. Metz CW. CHROMOSOMES AND SEX IN SCIARA. *Science*. American Association for the Advancement of Science; 1925; doi: 10.1126/science.61.1573.212.
62. PhaseGenomics: Aligning and QCing Phase Genomics Hi-C Data. Phase Genomics, Seattle, WA. https://phasegenomics.github.io/2019/09/19/hic-alignment-and-qc.html (2019).
63. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. Oxford University Press; 2009; doi: 10.1093/bioinformatics/btp324.
64. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*. Oxford Academic; 2010; doi: 10.1093/BIOINFORMATICS/BTP698.
65. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*. Oxford Academic; 2014; doi: 10.1093/BIOINFORMATICS/BTU314.
66. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al.. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; doi: 10.1093/bioinformatics/btp352.

67. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al.. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst*. NIH Public Access; 2016; doi: 10.1016/J.CELS.2015.07.012.
68. Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, et al.. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv*. Cold Spring Harbor Laboratory; 2018; doi: 10.1101/254797.
69. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al.. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. Cell Press; 2016; doi: 10.1016/j.cels.2016.07.002.
70. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. Academic Press; 1990; doi: 10.1016/S0022-2836(05)80360-2.
71. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. :http://www.repeatmasker.org2013;
72. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013; doi: 10.1093/bib/bbs017.
73. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010; doi: 10.1093/bioinformatics/btq033.
74. Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. Oxford University Press; 2018; doi: 10.1093/bioinformatics/bty191.
75. Mendelowitz LM, Schwartz DC, Pop M. Maligner: a fast ordered restriction map aligner. *Bioinformatics*. 2015; doi: 10.1093/bioinformatics/btv711.
76. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*. 2010; doi: 10.1093/bioinformatics/btq351.
77. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. *Bioinformatics*. Oxford Academic; 2021; doi: 10.1093/BIOINFORMATICS/BTAA1016.
78. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al.. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013; doi: 10.1093/bioinformatics/bts635.
79. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. BioMed Central; 2011; doi: 10.1186/1471-2105-12-323.
80. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. Oxford University Press; 2010; doi: 10.1093/bioinformatics/btp616.
81. Emms DM, Kelly S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol*. BioMed Central Ltd.; 2019; doi: 10.1186/S13059-019-1832-Y/FIGURES/5.
82. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. BioMed Central Ltd.; 2015; doi: 10.1186/S13059-015-0721-2/FIGURES/7.
83. Vaughan TG. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*. Oxford Academic; 2017; doi: 10.1093/BIOINFORMATICS/BTX155.
84. Urban JM: Lave: A collection of functions in R for genomic tinkering, processing, and plotting. https://github.com/JohnUrban/lave
85. Waskom ML. seaborn: statistical data visualization. *J Open Source Softw*. The Open Journal; 2021; doi: 10.21105/JOSS.03021.

86. Smit A, Hubley R. RepeatModeler Open-1.0. :http://www.repeatmasker.org2008;
87. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al.. The Dfam database of repetitive DNA families. *Nucleic Acids Res*. Oxford University Press; 2016; doi: 10.1093/nar/gkv1272.
88. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015; doi: 10.1186/s13100-015-0041-9.
89. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. Oxford Academic; 1999; doi: 10.1093/NAR/27.2.573.
90. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. Oxford Academic; 2011; doi: 10.1093/BIOINFORMATICS/BTR011.
91. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al.. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*. 2012; doi: 10.1101/gr.131383.111.

1668 **Table 1: Contiguity Statistics comparing the assembly before (Bcop_v1) and after (Bcop_v2)**
1669 **Hi-C scaffolding (Bcop_v2).**

	Bcop_v1 (Urban et al 2021)	Bcop_v1 Primary Only (Urban et al 2021)	Bcop_v2 (This paper)	Bcop_v2 Primary Only (This paper)
Number of sequences	744	205	595	4
Sum of all lengths	309,775,056	298,965,442	309,636,011	296,980,291
Max sequence length	23,039,227	23,039,227	97,081,274	97,081,274
Min sequence length	671	4,264	671	58,343,183
Mean sequence length	416,364	1,458,368	520,397	74,245,073
Median sequence length	21,428	94,665	17,801	70,777,917
Sequence length N50	6,790,317	6,790,317	71,047,972	71,047,972
Sequence length L50	14	14	2	2
Expected sequence length*	7,982,431	8,269,986	73,791,194	76,933,973
Sequence length NG50**	8,288,951	8,288,951	71,047,972	71,047,972
Sequence length LG50**	12	12	2	2
Normalized expected sequence length	8,831,279	8,830,144	81,601,468	81,599,549

1670 *- The terms "sequence" and "sequences" covers both contigs and scaffolds.*
1671 *- Bcop_v1 and Bcop_v2 are each comprised of a "primary" and "associated" assembly where the*
1672 *"primary" is a single complete haploid representation of the entire genome and the "associated"*
1673 *contains "haplotigs", repeat variants, and other redundant loci; and only represents a small*
1674 *subset of the genome. The primary assembly of Bcop_v1 (Bcop_v1_primary) was the input for Hi-*
1675 *C correction and scaffolding for Bcop_v2, which produced four chromosome-scale scaffolds and*
1676 *some unplaced contigs. The primary assembly for Bcop_v2 is comprised of only the four*
1677 *chromosome-scale scaffolds, and the associated assembly for Bcop_v2 contains the unplaced*
1678 *Bcop_v1_primary contigs and the Bcop_v1 associated assembly. Thus, the total length of*
1679 *Bcop_v2_primary is a bit shorter than Bcop_v1_primary.*
1680 **Expected sequence length as described elsewhere* [91].

1681 ***NG50, LG50, and "Normalized expected sequence length" use the expected somatic genome*
1682 *size of 280 Mb* [9] *rather than assembly length.*
1683

1684 **Table 2: Chromosome anchoring statistics compared to the previous reference assembly.**
1685

	Bcop_v1 (Urban et al 2021)	Bcop_v2 (This paper)	Expected size based on Gabrusewycz-Garcia (1964)*	Expected size based on Crouse (1943)	Gabrusewycz-Garcia (1964) Size Relationships	Crouse 1943 Size Relationships
Total primary length classified as autosomal	223.8-232.8 Mb	226.5 Mb	218-231 Mb	214-231 Mb	.	.
Total length classified as X	62-71 Mb	70.5 Mb	48.9-62.2 Mb	65.8-71.0 Mb **	Shortest and similar to II	Similar length as III
Total length classified as II	13.1-28.5 Mb	58.3 Mb	62.2-66 Mb	54.9-59.2 Mb	Shortest and similar to X	Shortest
Total length classified as III	5.4-12.5 Mb	71.0 Mb	66.6-70.7 Mb	65.8-71.0 Mb	Longer than X and II	Similar length as X
Total length classified as IV	34.7-46.4 Mb	97.1Mb	88.8-94.2 Mb	93.3-100.6	Longest	Longest

1686 ** Expected size calculations based on Gabrusewycz-Garcia (1964)* [30] *are from Supplemental*
1687 *Table S1 in Urban et al 2021* [9]*, and are based on genome size estimates from Rasch (2006)*
1688 [32]. *Chromosome lengths were predicted according to the number of polytene zones for each*
1689 *chromosome assigned by Gabrucewycz-Garcia 1964 (14, 14, 15, and 20 for X, II, III, and IV,*
1690 *respectively)* [30]. *Rasch estimated the X to be 48.9 Mb* [32]*, which is reflected in the range*
1691 *given. Autosomal size ranges reflect calculations using the estimated size of the entire genome*
1692 *(279.789 Mb)* [9] *or just the autosomal complement after subtracting out Rasch's expected size*
1693 *of the X (230.889).*
1694 *** Chromosome length calculations were predicted according to the number of polytene sections*
1695 *for each chromosome assigned by Crouse (1943) (12, 10, 12, and 17 for X, II, III, and IV,*
1696 *respectively)* [31]. *The size range of X is kept identical to that of III based on Crouse's*
1697 *expectation that they are the same length* [31]*, although the range could go as low as 48.9 Mb*
1698 *given Rasch's estimate* [32].
1699
1700
1701
1702
1703

1704 **Table 3: Chromosome-scale scaffold accessions and attributes.**

Chromosome	GenBank	GC Content*	# Genes (Maker2)**	# Genes (RefSeq)**
X	CM025456.2	35.0	4,797	3,146
II	CM025453.2	36.1	4,144	3,214
III	CM025454.2	35.7	6,213	4,426
IV	CM025455.2	35.6	7,126	5,340
Bcop_v2* **	GCA_014529535.2	35.5	23,115	16,542

1705 *GC computed using only A, C, G, T, and the sum of these 4 bases. N-content was ignored and
1706 its length not used in the sum.
1707 ** # protein-coding genes lifted over from Bcop_v1 assembly.
1708 *** Includes unplaced associated assembly sequences as well as the four chromosome scaffolds.

1709 **Table 4: Lifted-over gene annotation statistics.**

	Maker2 (Urban et al 2021)*	NCBI r100 (RefSeq)**
# protein-coding genes Bcop_v1	23,117	16,546
# genes lifted over to Bcop_v2	23,115 (99.99%)	16,542 (99.98%)
# genes not lifted over to Bcop_v2	2 (0.01%)	4 (0.02%)
# transcripts Bcop_v1	28,870	28,473
# transcripts lifted over to Bcop_v2	28,866 (99.99%)	28,464 (99.97%)
# transcripts not lifted over to Bcop_v2	4 (0.01%)	9 (0.03%)
# lifted over transcript sequences identical after liftover	28,813 (99.82%)	28,388 (99.73%)
# lifted over transcripts that yield identical protein sequence after liftover	28,817 (99.83%)	28,396 (99.76%)
# lifted over transcript sequences changed after liftover	53 (0.18%)	76 (0.27%)
# lifted over transcripts that yield changed protein sequence after liftover	49 (0.17%)	68 (0.24%)
# lifted over genes where all transcript isoform sequences are identical before and after liftover	23,083 (99.86%)	16,512 (99.82%)
# lifted over genes where all transcript isoforms yield same protein sequences after liftover	23084 (99.87%)	16516 (99.84%)
# lifted over genes where a subset of transcript isoforms sequences is identical before and after liftover	10 (0.04%)	3 (0.02%)
# lifted over genes where a subset of transcript isoforms yield same protein sequences after liftover	10 (0.04%)	4 (0.02%)
# lifted over genes where all transcript isoform sequences differ before and after liftover	22 (0.1%)	27 (0.16%)
# lifted over genes where all transcript isoforms yield changed protein sequences after liftover	21 (0.09%)	22 (0.13%)

1710 * Urban et al (2021) [9,34]
1711 ** NCBI *Bradysia coprophila* Annotation Release 100 [9,35]
1712

1713 **Table 5: Chromosomal counterparts in other species.**
1714

	Chromosome(s) from other species with highest percent of SCOs from given *B. coprophila* chromosome			
***B. coprophila* chromosomes**	***P. hygida***	***A. aegypti***	***A. gambiae***	***D. melanogaster***
X	X (97.3%)	1 (83.5%)	2 (54.3%); X (34.8%)	3 (57.1%); X (29.7%)
II	C (97.3%)	3 (81.9%)	2 (84.2%)	3 (59.0%); X (27.9%)
III	B (97.0%)	2 (85.1%)	2 (80.0%)	3 (47.2%); 2 (45.2%)
IV	A (98.2%)	2 (50.7%); 3 (44.8%)	3 (81.8%)	2 (70.3%); 3 (23.9%)

1715
1716
1717
1718

1719 **Table 6: Joint probabilities of SCOs on chromosomes from *B. coprophila* and *P. hygida* as**
1720 **an example.**
1721

		***P. hygida* chromosomes**				
		X	**C**	**B**	**A**	***Marginal B. cop***
***B. coprophila* chromosomes**	**X**	0.253	0.002	0.002	0.003	*0.261*
	II	0.000	0.166	0.001	0.003	*0.170*
	III	0.002	0.001	0.241	0.004	*0.248*
	IV	0.002	0.002	0.002	0.315	*0.321*
	Marginal P. hyg	*0.258*	*0.170*	*0.247*	*0.325*	

1722 This is an example, showcasing the two fungus gnats, of the input used for computing the raw
1723 entropy score before min-max normalization (see Methods), including the observed marginal
1724 probabilities of finding a single copy ortholog (SCO) on a given chromosome from a given species
1725 and the observed joint probabilities of finding a SCO on a given pair of chromosomes between the
1726 species. *B. coprophila* and *P. hygida* chromosomes are ordered by corresponding chromosomes as
1727 defined in Table 5, such that the highest joint probabilities are along the diagonal (left-to-right,
1728 top-to-bottom). The joint probabilities correspond to the proportions of SCOs that appear on (or
1729 the probabilities that any given SCO will appear on) each pair of inter-species chromosomes.
1730 Marginal probabilities (italicized) are obtained by summing the non-rounded values in the rows
1731 for *B. coprophila* and the columns for *P. hygida* (summed values rounded to 3 digits)*. The raw
1732 entropy score is computed on the observed joint probabilities. To get the minimum possible
1733 entropy the joint probabilities between these species could take given their SCO distributions,
1734 entropy is computed on each set of marginal probabilities separately, then averaged. To get the
1735 maximum entropy, entropy is computed on the set of joint probabilities expected at random, which
1736 are defined as the products of each inter-species pair of marginal probabilities. Min-max
1737 normalized entropy (MMNE) scores are then computed as: (observed entropy – minimum
1738 entropy)/(maximum entropy – minimum entropy). See the Methods for further details.
1739
1740 * *Note that joint probabilities were rounded to 3 digits for simplifying this table, but not for*
1741 *computing the entropy. This results in a difference between the sum of rounded joint probabilities*
1742 *in a column or row and the reported marginal probability computed from summing non-rounded*
1743 *joint probabilities, then rounding the sum. The difference is only in the third position after the*
1744 *decimal.*
1745

1746 **Table 7: Repeat characterization of Bcop_v2 using RepeatMasker.**
1747

Type	Class	Sub-Class	Number of elements	Length occupied (bp)	Percent of Genome Length
Retroelements	-	-	39,488	15,747,087	5.09
	SINEs	-	6,623	1,245,502	0.40
	Penelope	-	1,629	562,250	0.18
	LINEs	-	21,699	9,895,637	3.2
		L2/CR1/Rex	4,955	2,617,504	0.85
		R1/LOA/Jockey	1,103	318,765	0.10
		R2/R4/NeSL	417	157,115	0.05
		RTE/Bov-B	3,378	1,938,253	0.63
		L1/CIN4	346	57,084	0.02
	LTR elements	-	11,166	4,605,948	1.49
		BEL/Pao	2,186	1,353,548	0.44
		Ty1/Copia	1,472	639,157	0.21
		Gypsy/DIRS1	5,659	1,935,158	0.62
		Retroviral	1,552	646,987	0.21
DNA transposons	-	-	51,336	10,756,047	3.47
	hobo-Activator	-	16,061	2,460,594	0.79
	Tc1-IS630-Pogo	-	878	288,734	0.09
	En-Spm	-	0	0	0
	MuDR-IS905	-	0	0	0
	PiggyBac	-	227	77,358	0.02
	Tourist/Harbinger	-	71	9,265	0.00
	Other	-	397	127,501	0.04
Rolling-circles	-	-	12,582	4,339,527	1.40
Unclassified	-	-	512,007	91,528,113	29.56
Total interspersed repeats	-	-	-	11,803,1247	38.12
Small RNA			1,336	451,128	0.15
Satellites*			1,734	225,658	0.07
Simple repeats*			49,847	2,952,948	0.95
Low complexity*			10,001	459,419	0.15
Total bases masked**	-	-	-	126,310,607	40.79

1748 * “Satellites”, “Simple repeats”, and “Low complexity” here are as defined and reported by
1749 RepeatMasker [71].
1750 **Number of bases identified for masking by RepeatMasker [71], though not necessarily masked
1751 out by default in the assembly.

1752
1753 **Table 8: Sequence of satellites found at or near scaffold termini.**

Satellite	Unit Length	Percent AT	Sequence
orange-NLTR	175 bp	70.9%	ACTAAACTCGAAATTATTTGCTTTACGTGTAGCGTCAAATTTTTAATTTCTTGTCTTAATGTTTCAGTCTTGATTCATTTTACTAAATAAAATTCATCTCAGCGACGTCGTGACCATTTTTGATATATTTTTCATTCAATTATTAAGAAAATGTTTGCTACGCCTATGCTTGTGA
red-NLTR	161 bp	76.4%	TCCAAATTCCATGAAATTTTAACTGTATGTATGAAATGCTAATCAAACCATATATCCAAAATTTCATGCGATTCTGAGACAAAAAATTTTTAATTTTACTTTTTTGACCATACATTATGACCAAAAAATGGCATTATTACAAAAATTTAAAAAAAATTCTC
blue-NLTR	176 bp	67.6%	AATTCAATTAGCTTTCTAACCATACCAACTTTAGATCACCAGACTTCATTTTCGAAGAGCGAGAGCAGTTTAAAGGAAAATTGTGAAGCAAAAATTGACAAGTTTCAACTTATTTCATTGATTTGACACAGAAGATAGCGTTTTGCTCTTCTACAAGTTTCGCTTATTAAATTCTA
brown-NLTR	191 bp	67%	AGCTTCAGAGCTACATATCCATGTGAAAAATACCTATTAATTTTGGATGTTTTCTACCTTCACGTTATATTATCTCCCAAGCAACGACTTCGATCGTAATGTTTGTTTCACATTTATTTATGAATCGAAGTAGCGCGTTCGGGCTTATCCGTCGTTTGAAAATCAAAGAGAATTTTTTTTTAAATTTTTTC
B56-MLTR	56 bp	58.9%	GTTTACTCTTTGACAGCCAATCATCGACTGAAGTATGTCCTTGTTTTGGCAGAGAT

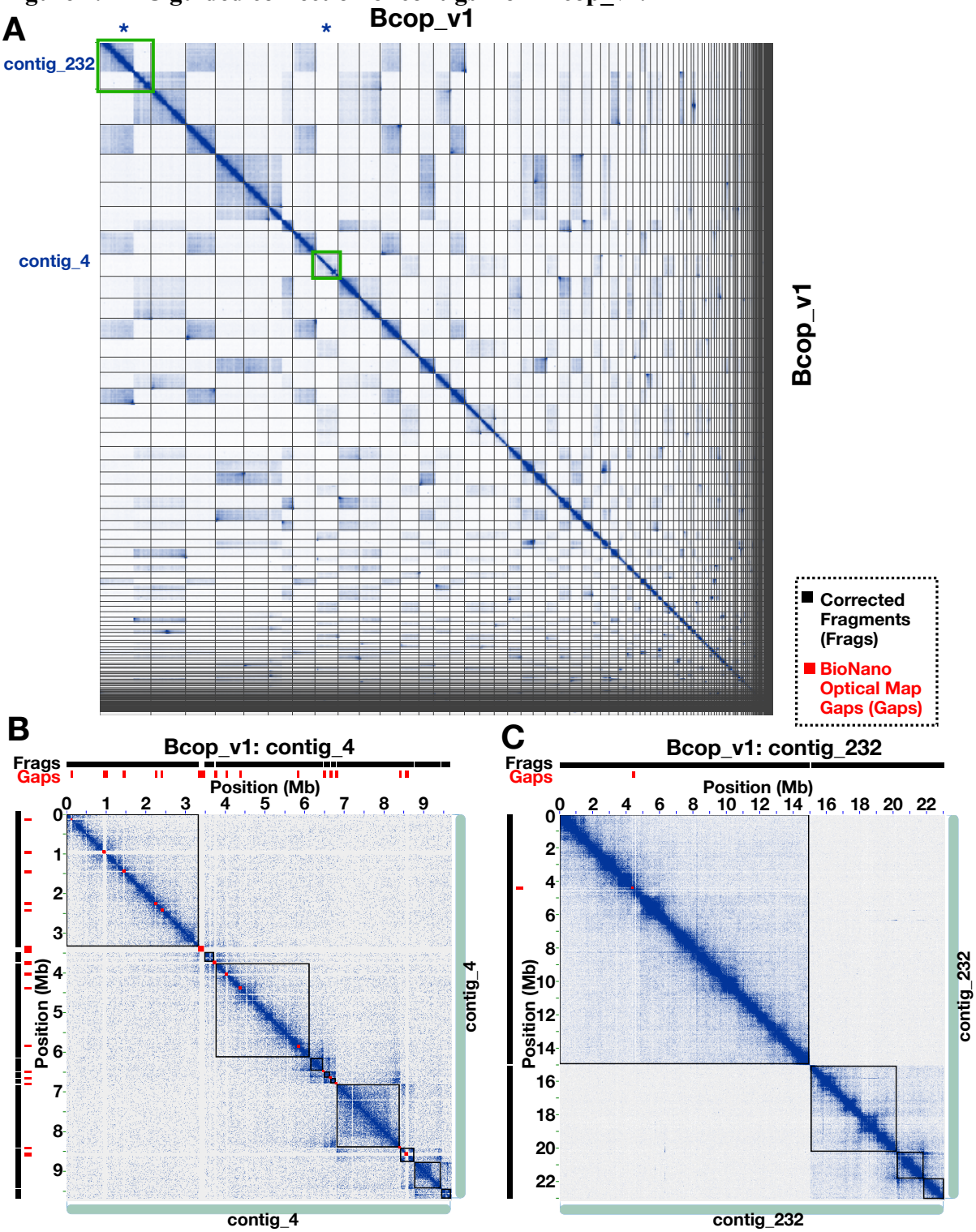
1754 NLTR = Nucleosome-length tandem repeat; MLTR = Medium-length tandem repeat.
1755
1756

1757 **Table 9: Nucleotide distances between FBRs and other features of the X chromosome are**
1758 **concordant with expectations.**

Locus 1	Locus 2	Approx. Polytene Map Distance (# sub-zones)	Expected Distance (Mb)	Observed Distance (Mb)
Centromere	FBR1	2	3.4	3.49
5' X' break point	FBR1	0.5	0.839	0.620
FBR1	FBR2	7	11.8	10.3
FBR2	FBR3	26	43.6	42.7
FBR3	3' X' break point	3	5.0	4.65
FBR3	3' end of chromosome	6	10.1	12.3

1759 The X polytene map [30] is organized into 14 zones of similar lengths, each with 3 subzones (A,
1760 B, and C) for a total of 42 sub-zones. Assuming nucleotide lengths are also similar, the ~70.5 Mb
1761 X-scaffold would indicate zone lengths of ~5 Mb (5.036276) and an average sub-zone length of
1762 ~1.68 Mb (1.678759), which we used to compute expected nucleotide distances.
1763
1764

1765 **Figure 1. Hi-C guided correction of contigs from Bcop_v1.**



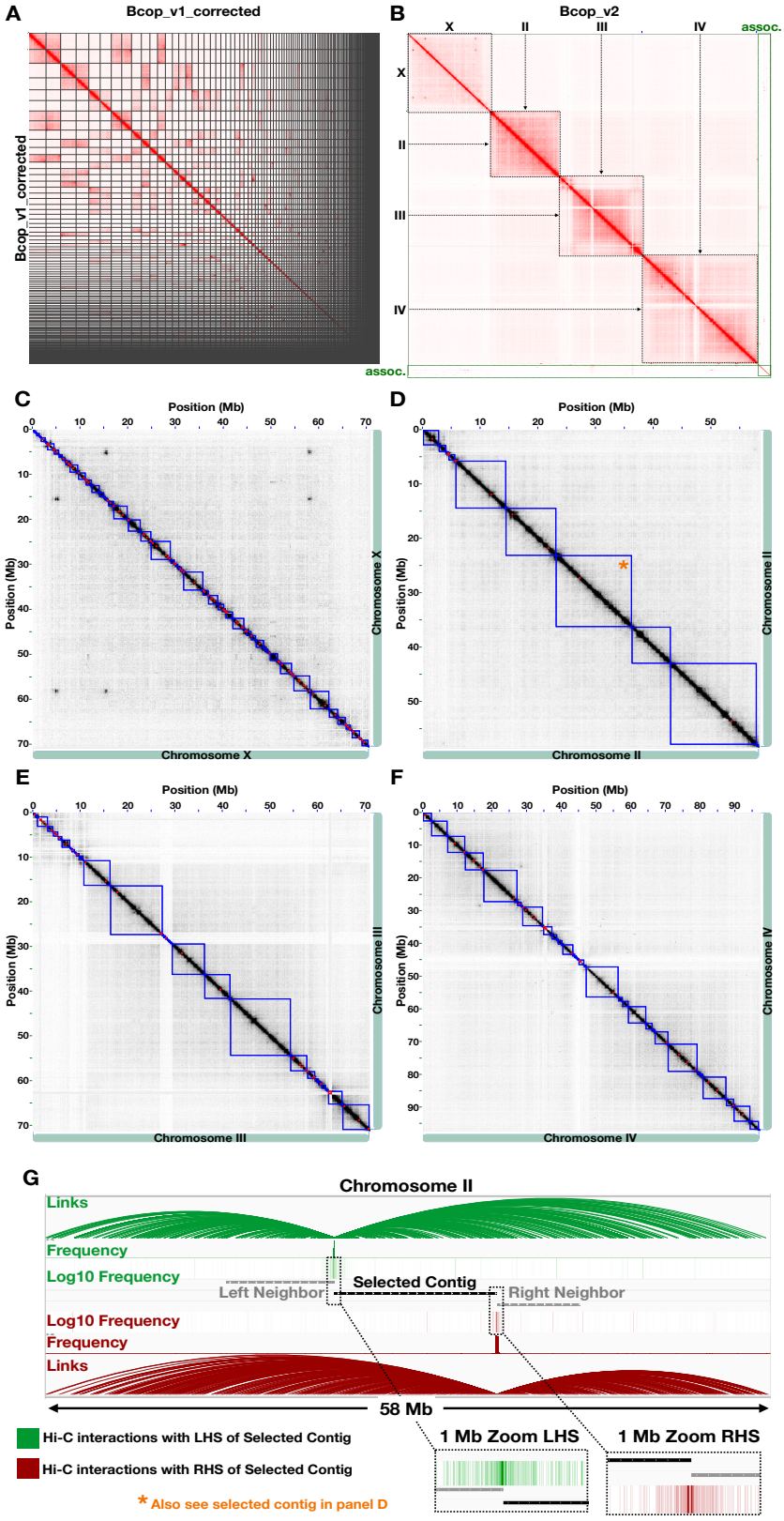
1766

1767 **Figure 1. Hi-C guided correction of contigs from Bcop_v1**

1768 **(A)** The entire Hi-C map of all primary scaffolds and contigs in Bcop_v1 ordered from longest
1769 to shortest before correction and scaffolding. Scaffolds from B and C are highlighted by green
1770 boxes. **(B)** Shows example of disrupted Hi-C signals related to joins made by scaffolding with

1771 BioNano Genomics optical maps and associated gaps. This is contig_4, a ~9.7 Mb contig with 14
1772 optical map gaps that was broken into 10 corrected fragments with the 9 breaks correlated with
1773 gap locations. New corrected contigs are outlined by black squares and optical map gaps are
1774 represented by red squares. Both are also displayed above and to the left of the Hi-C map. The
1775 bottom and right show seafoam-colored bars representing the full original Bcop_v1 contig. **(C)**
1776 Example of disrupted Hi-C signal that occurred independent of optical map gaps within
1777 contiguous sequence. The Hi-C signal forms two obvious sequences that interact within
1778 themselves but that are significantly depleted of interactions with each other. The black squares
1779 outline the new contigs after correction. The self-interacting sequence on the bottom right has
1780 additional Hi-C signals that suggested possible structural errors; hence it being broken up into
1781 additional fragments. Fragments, optical map gaps, and the original contig are outlined and
1782 represented as in B.
1783

1784 **Figure 2: Hi-C signal shows high 'chromosome specificity' and 'structural accuracy' of the**
1785 **chromosome-scale scaffolds.**



1786

1787 **Figure 2: Hi-C signal shows high 'chromosome specificity' and 'structural accuracy' of the**
1788 **chromosome-scale scaffolds.**
1789 **(A)** Interaction frequencies are visualized across Bcop_v1_corrected (Bcop_v1 after Hi-C guided
1790 correction), which was the input to Hi-C scaffolding. **(B)** Interaction frequencies visualized
1791 across the final version of Bcop_v2, after polytene orientation and with associated/unplaced
1792 contigs. Squares are drawn around boundaries of the four chromosome-scale scaffolds. **(C, D, E,**
1793 **F)** Interaction frequencies across the X, II, III, and IV chromosome scaffolds, respectively. Blue
1794 squares correspond to individual contigs from Bcop_v1_corrected that make up the scaffolds.
1795 The asterisk in D corresponds to the selected contig in G. **(G)** An example of the interaction
1796 frequencies between the ends of a selected contig and the rest of the chromosome-scale scaffold.
1797 The selected contig shown was from chromosome II (see asterisk in panel D). The selected
1798 contig is colored black whereas the two neighboring contigs are grey. Links (arcs) tracks
1799 correspond to a summary of loci that were linked together by paired-end reads where at least one
1800 mate maps to the given contig end. The "Frequency" tracks are bar plots corresponding to the
1801 number of times each bin across the chromosome was linked to the given contig end. However,
1802 interaction frequency decays exponentially with distance. Thus, the "Log10 Frequency" across
1803 the chromosome is also shown as a heatmap, and zoom-ins of the Log10 Frequency heatmaps in
1804 1 Mb regions centered on each contig end are shown for higher resolution. Green and red tracks
1805 correspond to interactions and interaction frequencies with the left end and the right end of the
1806 selected contig, respectively.
1807
1808

Figure 3: Chromosome identity, orientation, and assessment of the chromosome-scale scaffolds.

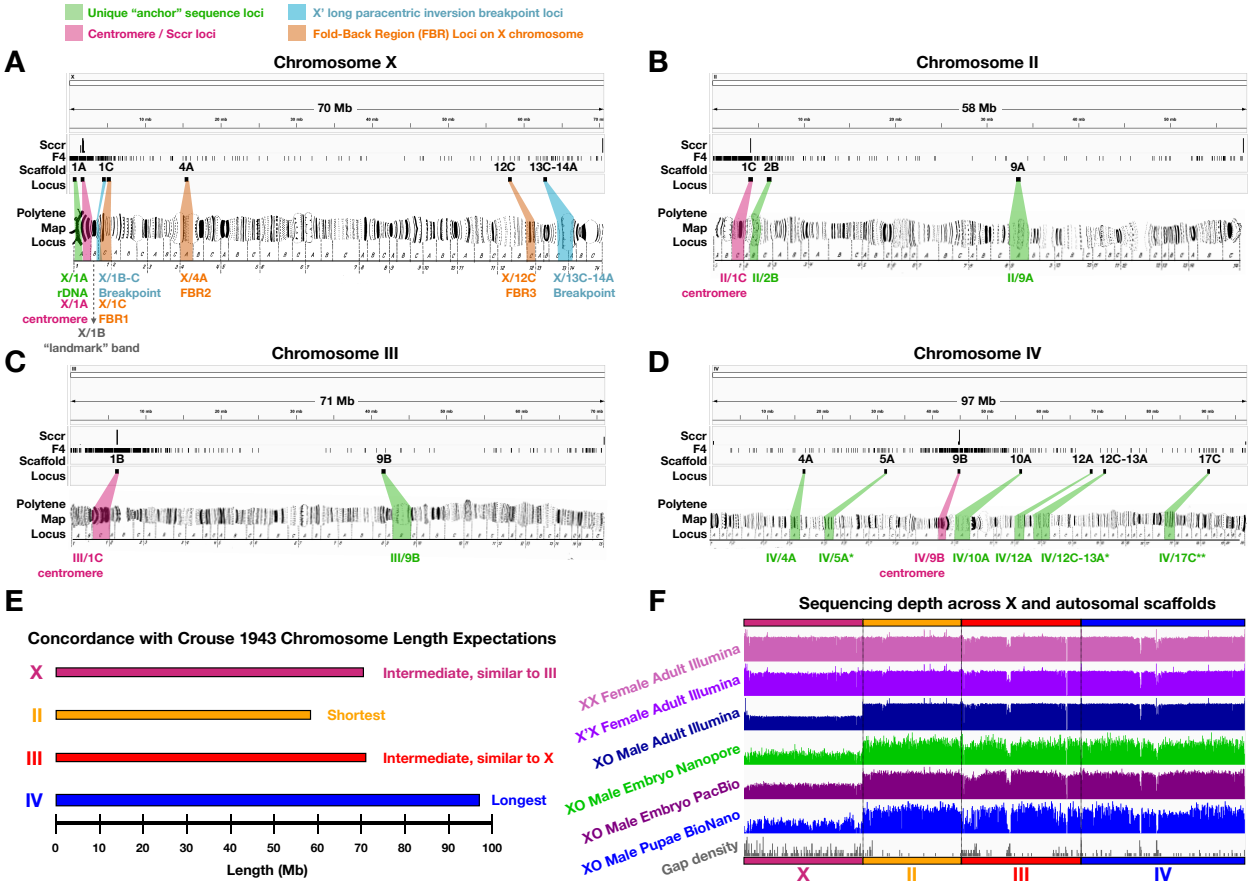
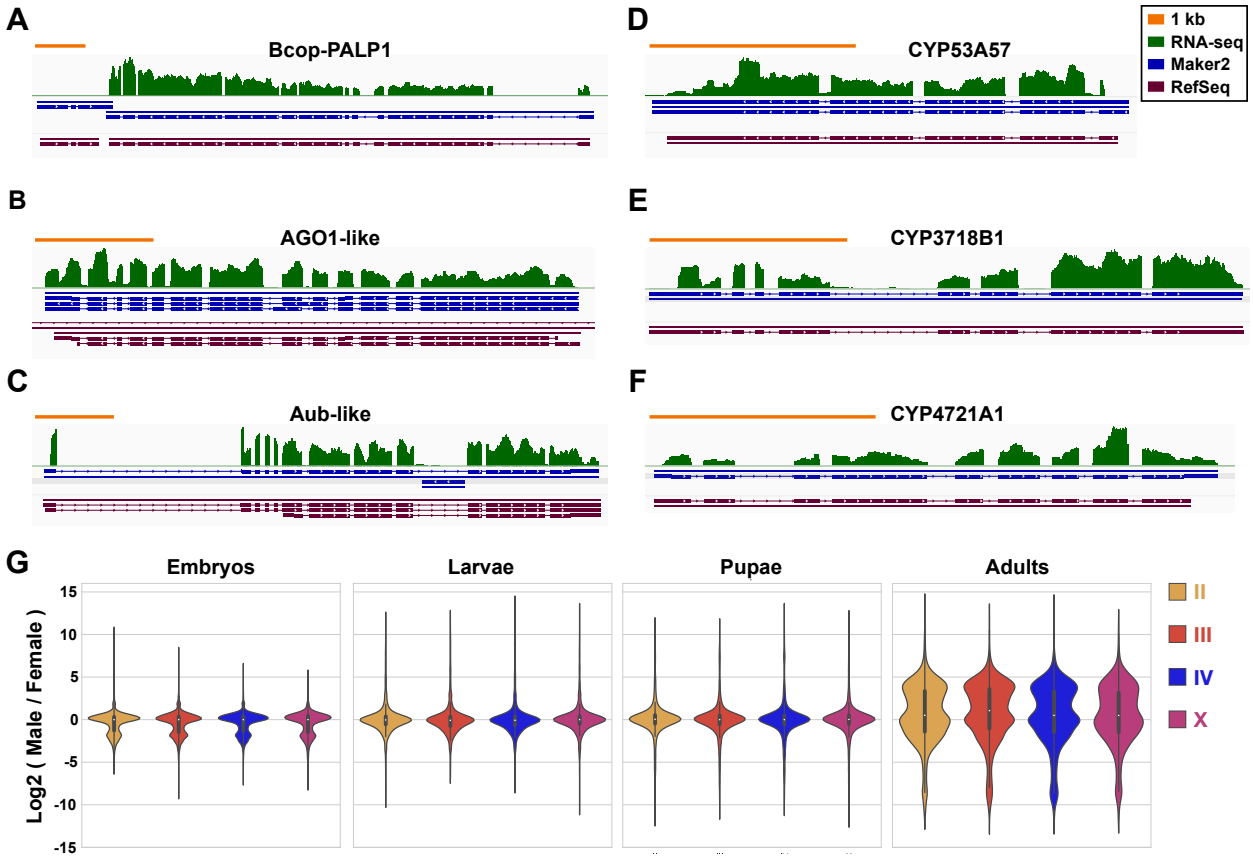


Figure 3: Chromosome identity, orientation, and assessment of the chromosome-scale scaffolds.

(A-D) Anchoring, polytene map orientation, scaffold structural verification, and loci of interest on the chromosome-scale scaffolds for chromosomes **(A)** X, **(B)** II, **(C)** III, and **(D)** IV. The chromosome identity and specificity of each scaffold, their consistent orientations with polytene maps, and structural verifications were obtained using 10 unique "anchor" sequences with known chromosomal addresses mapped previously using *in situ* hybridization [17–29]. Centromeric sequences identified the expected locations of centromeres [22,30,31,45] on acrocentric (X, II, and III) and meta-centric (IV) chromosomes, which were further supported by the density of a probe sequence (F4 from ScRTE) shown previously to be pericentromeric in all four chromosomes [45]. The scaffold for X also shows the locations found for the fold-back regions and long paracentric inversion breakpoints on the X'. Each chromosome-scale scaffold had at least one known unique sequence location anchoring it to a specific chromosome. There was no conflicting evidence with regard to chromosome identity nor with regard to the order in which sub-chromosomal sequences mapped along the scaffolds. Polytene chromosome maps were reproduced from plates 1, 2 and 3 of Gabrusewycz-Garcia (1964) [30] with permission from Springer Nature under permission number: 5490830617309. **(E)** Chromosome scaffold sizes are consistent with chromosome length expectations from Crouse (1943) [31]. **(F)**

1832 Chromosome scaffolds are consistent with read depth expectations on the X versus autosomes in
1833 males and females. Overall, this evidence suggests that Hi-C accurately ordered and oriented the
1834 contigs into chromosome-scale scaffolds corresponding to the expected four somatic
1835 chromosomes.
1836

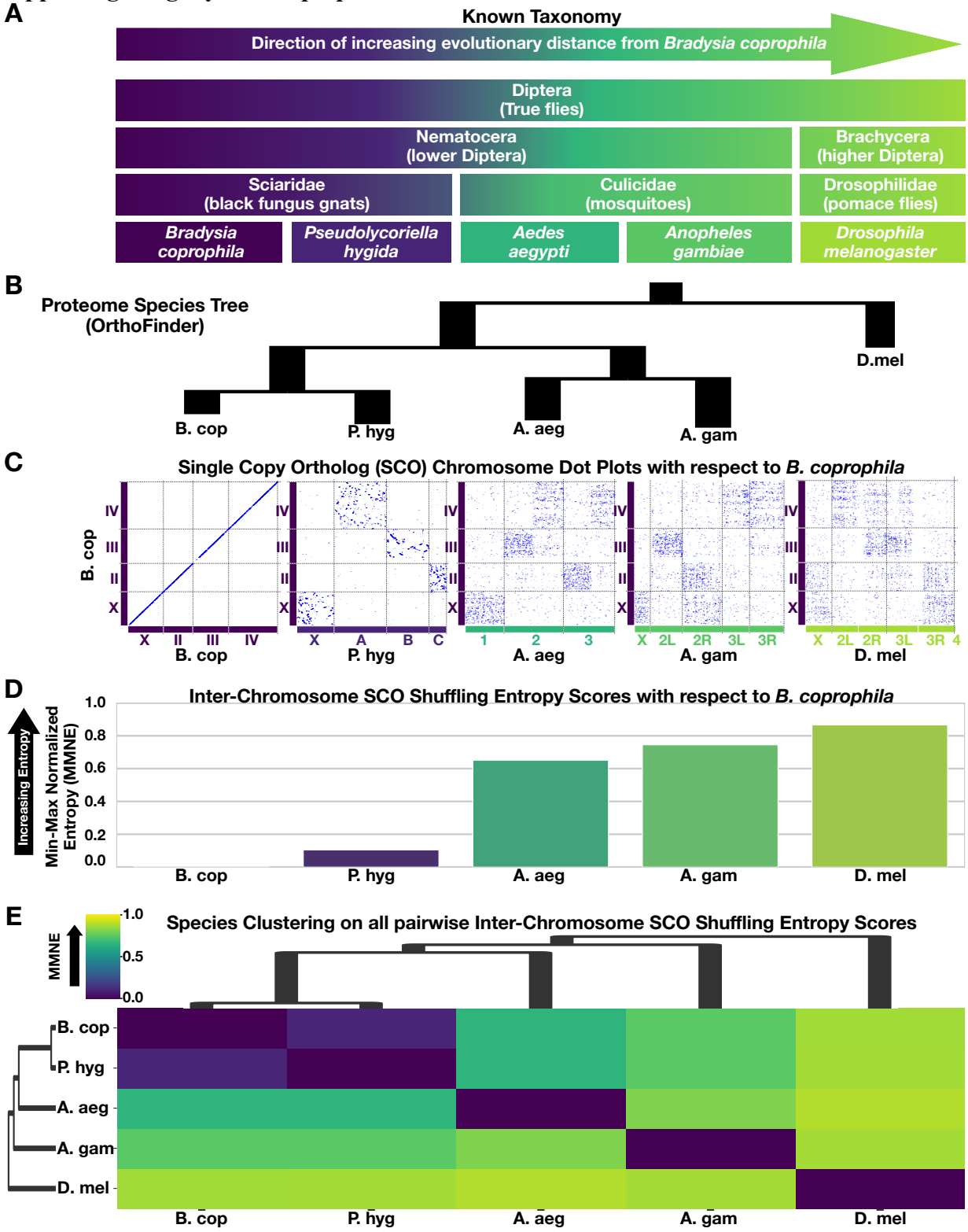
1837 **Figure 4: Analysis of gene models on Bcop_v2 lifted over from Bcop_v1 gene annotation**
1838 **sets.**
1839



1840
1841
1842 **Figure 4: Analysis of gene models on Bcop_v2 lifted over from Bcop_v1 gene annotation**
1843 **sets.**
1844 **(A-F)** Example Bcop_v1 gene models were chosen to show concordance with RNA-seq datasets
1845 after liftover to Bcop_v2. For simplicity, only one RNA-seq dataset is shown (irradiated fourth
1846 instar pre-eyespot larvae, replicate 1 [37]). However, lifted over gene models were consistent
1847 with all RNA-seq datasets checked, including those from all life stages and both sexes used in
1848 the original gene annotation [9]. RNA-seq concordance is illustrated here using genes of interest
1849 from recent *B. coprophila* studies, including **(A-C)** PARP and Argonaute family genes that arose
1850 in response to radiation [37], and **(D-F)** example "alien" cytochrome P450 genes [7]. For all six
1851 gene panels **(A-F)**, as summarized in the pictorial legend box on the top right: the orange scale
1852 bar represents 1 kb, the green trace (top level) shows RNA-seq coverage, the blue gene models
1853 (middle level) are from Maker2 [9,34], and the purple gene models (bottom level) are from
1854 NCBI RefSeq [9,35]. In all cases, the RNA-seq trace shows genome-wide coverage that was
1855 computed agnostic to the location of gene structures, and that was not artificially restricted to
1856 exons. Thus, the correspondence of higher RNA-seq coverage over exons shows that fidelity of
1857 the liftover process. **(G)** NCBI gene models lifted over to Bcop_v2 were used for chromosome-
1858 specific differential expression analysis of RNA-seq data from males versus females across four
1859 life stages [9]. Log2 fold-changes are shown for genes grouped by chromosome-scale scaffolds.
1860 The distribution of male vs. female fold-changes across X-linked genes match the distributions

1861 across each autosomal chromosome rather having a 2-fold down-shifted X-linked fold-change
1862 distribution corresponding to the single male X chromosome compared to two in females. This
1863 indicates there is dosage compensation.
1864

Figure 5. Interchromosomal gene shuffling increases with evolutionary distance, further supporting integrity of *B. coprophila* chromosome-scale scaffolds.



1869 **Figure 5. Interchromosomal gene shuffling increases with evolutionary distance, further**
1870 **supporting integrity of *B. coprophila* chromosome-scale scaffolds.**
1871
1872 **(A)** Known taxonomic and phylogenetic information regarding *B. coprophila* and four other
1873 Dipteran species. With respect to *B. coprophila,* the species are presented here from left to right
1874 in order of the most closely related to most distantly related. The color scheme from left to right
1875 (dark purple to light yellow-green) reflects the entropy scores between *B. coprophila* and these
1876 species as in C-E. The species are presented in this order in each sub-figure below as well. **(B)**
1877 Species tree produced by OrthoFinder [81,82] using the proteomes of these five species
1878 reproduces the expected relationships amongst these Dipteran flies. **(C)** Scatter plots of SCO
1879 locations along the chromosomes of each of the five species (x-axis) plotted against *B.*
1880 *coprophila* (y-axis). Inter-chromosomal SCO scattering increases with evolutionary distance as
1881 expected. Blue dots represent SCO locations. Colored boxes along x- and y- axes represent
1882 chromosomes. The chromosome box color scheme reflects the entropy score color scheme in D-
1883 E. **(D)** The pairwise min-max normalized entropy scores for *B. coprophila* computed against all
1884 five species. The entropy scores increase with evolutionary distance as expected. Although the
1885 entropy scores approach a fully random state with evolutionary distance from *B. coprophila*,
1886 pairwise relationships of chromosomal SCO groupings still display a bit of non-randomness even
1887 in the most distantly related fly species tested (*D. melanogaster*). **(E)** Matrix of all pairwise min-
1888 max normalized entropy (MMNE) scores among the five species presented as a hierarchically-
1889 clustered heatmap. The dendrogram from pairwise entropy scores reproduces the expected
1890 relationships among these Dipteran flies. Altogether, as the SCO dispersion and entropy of
1891 chromosomal gene sets follows evolutionary expectations, the location of genes in the
1892 chromosome-scale scaffolds we report here for *B. coprophila* are non-random and suggest the
1893 contigs were grouped correctly by chromosome.
1894

Figure 6: Diverse genomic datasets support the existence of "alien" genes within the chromosomes.

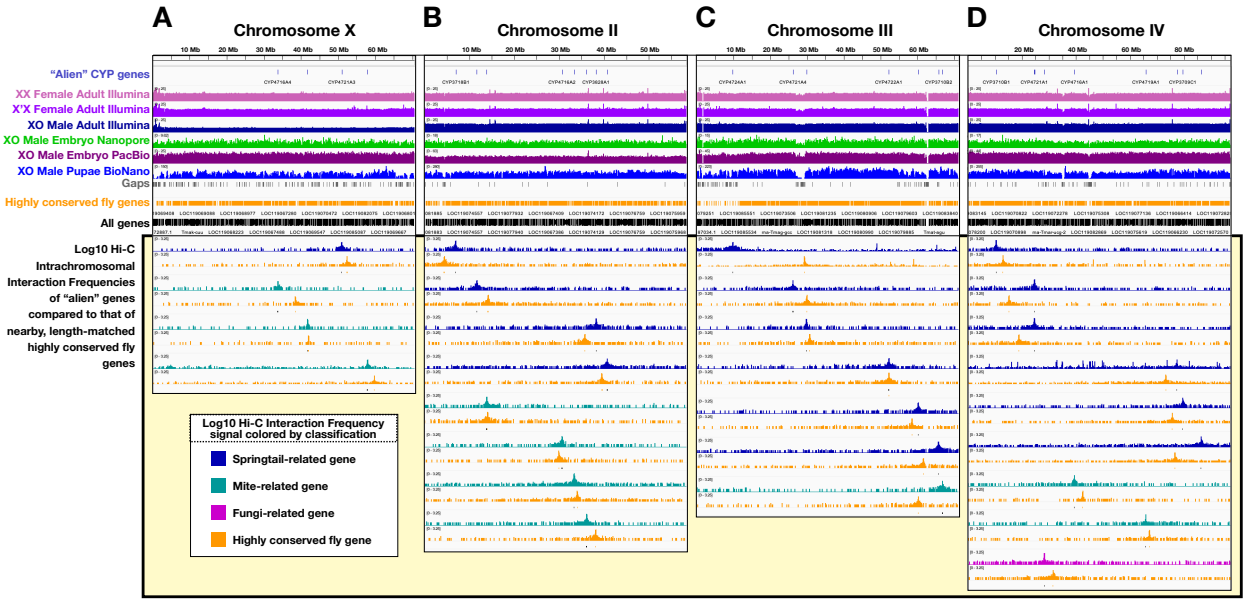
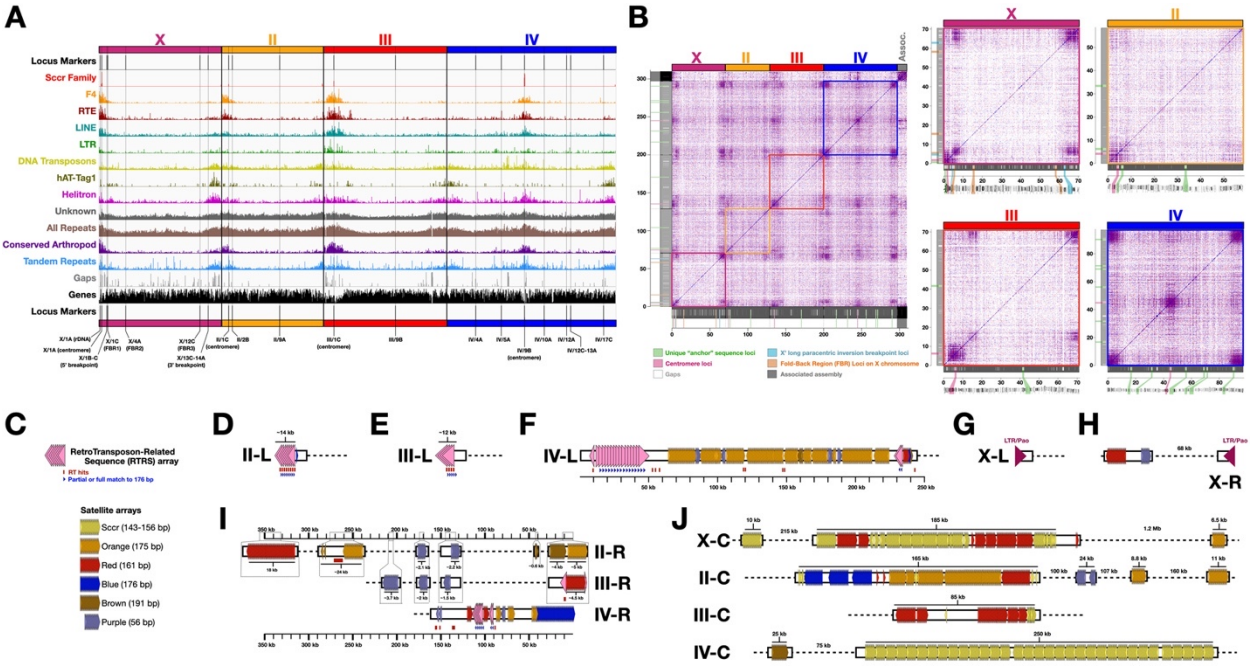


Figure 6: Diverse genomic datasets support the existence of "alien" genes within the chromosomes.

IGV traces for chromosome **(A)** X, **(B)** II, **(C)** III, and **(D)** IV. From the top track downward, each shows the location of (1) "alien" CYP genes described by Feyersen et al [7], (2-7) coverage tracks from various genomic technologies and samples [9,10], (8) gaps in the assembly where coverage signal is expected to drop, (9) locations of all highly conserved fly genes, and (10) locations of all genes. In each, the 11th track from the top and all tracks beneath it highlighted by the yellow box correspond to the log10 intrachromosomal interaction frequency (IIF) signal emanating from selected "alien" genes (dark blue, cyan, or magenta) or the closest by highly conserved fly gene that is approximately the same length (orange). All log10 IIF tracks have the same y-axis range (0-3.25), allowing direct comparison of peak heights and shapes. Under each pair of log10 IIF signals is a track showing the exact location of both genes. Overall, the genes reported to be "aliens" by Feyersen et al [7] show no differences in coverage or Hi-C interactions compared to highly conserved fly genes, demonstrating their existence inside the scaffolds is not a result of contamination and/or assembly errors.

Figure 7: Repeat regions, centromere locations, and terminal repeats across the *B. coprophila* chromosome-scale scaffolds.

JPEG



1923 **Figure 7: Repeat regions, centromere locations, and terminal repeats across the *B.***
1924 ***coprophila* chromosome-scale scaffolds.**
1925 **(A)** Repeat densities in 100 kb bins across each chromosome, and locations of known locus
1926 markers. Sccr = Sciara coprophila centromeric repeat [45]. F4 = probe sequence for ScRTE [45].
1927 hAT-Tag1 = a DNA transposon sub-family. RTE = a LINE sub-family. F4, RTE, LINE, and
1928 LTR are retrotransposon families. Helitron elements are DNA transposons thought to amplify
1929 through a rolling-circle mechanism. Unknown = repeats that were modeled but not classified as a
1930 known repeat class. All Repeats = all in comprehensive repeat library. Conserved Arthropod =
1931 matches to known Arthropod Repeats (excluding *B. coprophila*). Tandem Repeats were found
1932 with Tandem Repeats Finder [89]. Gaps = assembly gaps. **(B)** Dot plots corresponding to (i) the
1933 entirety of Bcop_v2, including all primary chromosome scaffolds as well as the associated
1934 contigs, and (ii) each individual chromosome mapped against itself. The locations of various loci
1935 markers are shown in the margins, are color-coded as shown in the legend, and are extended to
1936 axes. Gap locations are shown as white rectangles in the grey margins. Polytene chromosome
1937 maps were reproduced from plates 1, 2 and 3 of Gabrusewycz-Garcia (1964) [30] with
1938 permission from Springer Nature under permission number: 5490830617309. **(C)** Legend to help
1939 interpret D-J. **(D-G)** Pictorial models of the left termini for scaffolds corresponding to **(D)**
1940 chromosome II (II-L), **(E)** chromosome III (III-L), **(F)** chromosome IV (IV-L), and **(G)**
1941 chromosome X (X-L). **(H-I)** Pictorial models of the right termini for scaffolds corresponding to
1942 **(H)** chromosome X, and **(I)** chromosomes II, III, and IV. **(J)** Pictorial models of satellite
1943 sequences in centromeric and pericentromeric regions of each chromosome scaffold. Models
1944 represent the approximate positions and lengths of RTRS and satellite arrays. RTRS copy
1945 number is shown. Satellite copy number can be approximated by the total length divided by the
1946 satellite unit length.
1947

Figure 8: Putative Fold-Back Regions (FBRs) and X' Breakpoints further confirm accuracy of the X scaffold.

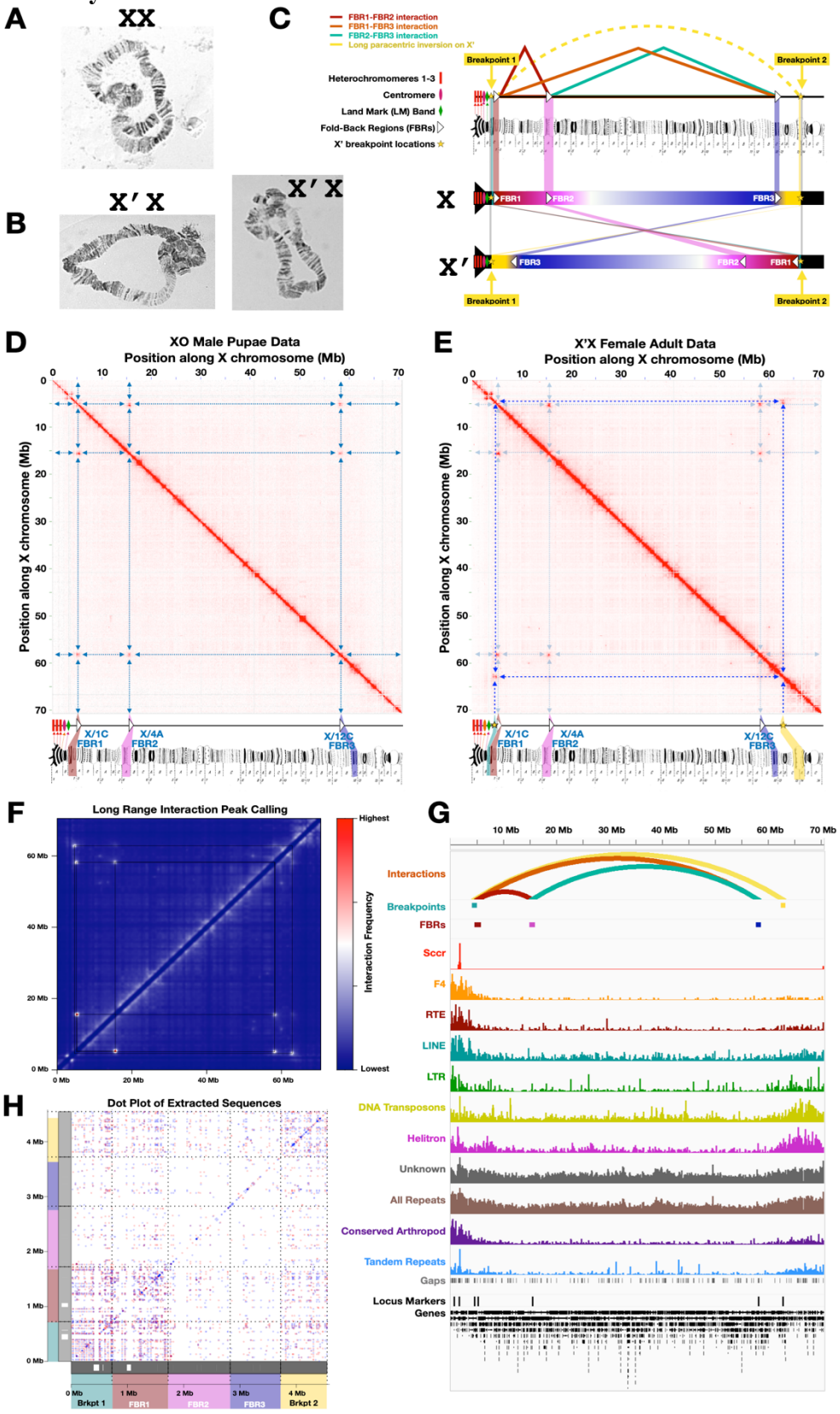


Figure 8: Putative Fold-Back Regions (FBRs) and X' Breakpoints further confirm accuracy of the X scaffold.

(A-B) Polytene chromosome spreads demonstrating how the X and X' chromosomes are usually found: folded back on itself with specific loci physically interacting. **(C)** Pictorial representation of what is known about the X and X' chromosomes, FBRs, and inversion breakpoints. **(D)** Hi-C map from male pupae data showing three interacting loci and how they appear to correspond to the expected locations of FBRs along polytene maps. **(E)** Hi-C map from X'X adult female data [10] that shows the FBRs as well as two additional loci that correspond to the long paracentric inversion breakpoints. **(F)** An intermediate scatter plot heat map visualizing part of the long-range interaction peak cluster calling pipeline with squares demonstrating long-range interactions found by this method. **(G)** IGV Trace of the X chromosome with special attention to the locations identified as FBRs and breakpoints (top 3 tracks), as defined by our long-range Hi-C interaction peak detection method, and how they relate to repeat-rich and gene-rich regions. **(H)** Dotplot of extracted sequences corresponding to FBRs and breakpoint regions, showing sequence matches within and between each sequence. Blue dots represent matches on the same strand. Red dots represent matches on opposite strands. Polytene chromosome maps in C-E were reproduced from plates 1, 2 and 3 of Gabrusewycz-Garcia (1964) [30] with permission from Springer Nature under permission number: 5490830617309.