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 lower dipteran fungus gnat, *Bradysia* (*Sciara*) *coprophila*, has unusual chromosome biology. Chromosome imprinting was first discovered in this system. All paternal chromosomes are eliminated during spermatogenesis whereas both maternal X sister chromatids are retained. Embryos start out with three copies of the X chromosome, but 1-2 copies are eliminated from all somatic cells as a part of sex determination, and one is eliminated in the germline to restore diploidy. These developmentally normal features present opportunities to study unusual chromosome movements that may occur as rare/abnormal events in other systems. To help with such studies, we previously generated a highly contiguous optical-map-scaffolded long-read assembly (Bcop_v1) of the male somatic genome that contains four chromosomes. 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 chromatin conformation capture (Hi-C) data was collected and used to first produce a corrected Bcop_v1 assembly by breaking contigs at mis-joined regions, and then used to order and orient the corrected contigs into chromosome-scale scaffolds. The resulting assembly (Bcop_v2) had four chromosome-scale scaffolds as expected. Previously known chromosomal locations of locus-specific sequences were used to (i) identify the corresponding chromosome for each scaffold, and (ii) orient the chromosome scaffolds in the same order as published polytene chromosome maps. Finally, the gene

annotations produced for Bcop_v1 were lifted over to Bcop_v2 to allow a seamless transition in adopting the updated chromosome-scale assembly.

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]. Regarding somatic or germline, there are chromosomes referred to as the L chromosomes that are only found in germ cells [1]. Regarding male or female, half of all females, but no males, contain a variant of the X chromosome called the X' (X prime) [1]. 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, for example. The L chromosomes are eliminated from all somatic nuclei not set aside for the germline in the 5th-6th nuclear division [1]. 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]. There is a later point in development in which both the male and female germline are diploid for the X after eliminating one paternal copy [1]. In addition, one or more L chromosomes are eliminated from the germline around the same time, which prevents their accumulation [1]. 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]. 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]. 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 [2], and developmentally-programmed gene amplification [3,4]. Overall, fungus gnats present many opportunities to study chromosome biology.

Previously, we published a highly contiguous assembly of the somatic chromosomes (X, II, III, and IV) [5]. We targeted the somatic genome in males as it presented the lowest complexity version of the genome, lacking both the L and X' chromosomes. 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 [5]. 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 [5]. The assembly with the best evaluation scores was scaffolded with optical maps from BioNano Genomics [5]. The resulting assembly, Bcop_v1, had a contig NG50 of ~2.4 Mb [5] and scaffold NG50 of ~8.2 Mb (Table 1) [5]. 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 [5]. 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 [5]. Overall, nearly 50% of the expected genome size was anchored into chromosomes based on coverage and known sequence locations [5]. 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 cycle in *Bradysia coprophila*, our goal was to develop better models of the chromosome sequences. We approached this by using chromosome conformation capture with deep sequencing (Hi-C) to scaffold the chromosomes [6], as has been successfully done by many others [7-9]. Specifically, Hi-C is so powerful at putting chromosome sequences together from a collection of sub-chromosomal sequences because of two major attributes of in vivo chromosome interactions captured by Hi-C data: (i) because intra-chromosomal interactions are much more frequent than inter-chromosomal interactions, it is possible to use interaction frequencies to cluster contigs into 'chromosome groups', and (ii) because interaction frequencies decay with distance, it is possible to use them to order and orient the contigs in a chromosome group in the order they most likely appear along a chromosome.

Overall, we present here the first chromosome-scale genome assembly for *Bradysia coporophila*, 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 to seamlessly transition to the updated genome assembly. We expect that the chromosome-scale genome and gene annotations 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.

Methods and Results:

Male Pupae Collection

The dark-winged fungus gnat, *Bradysia coprophila* [10], was previously referred to as *Sciara coprophila* (Lintner, 1895) in chromosomal and molecular biology research papers since the 1920s [1,11], and is also known by other names such as *Sciara tilicola* (Loew, 1850) and *Sciara amoena* (Winnertz, 1867). In this study, the fungus gnats were from the HoLo2 line maintained in the International *Sciara* Stock Center at Brown University (https://sites.brown.edu/sciara/). *B. coprophila* females are monogenic, meaning they have either only male or only female offspring, which is determined by whether they harbor a variant of the X chromosome (X') or not. Specifically, X'X females are female producers and XX females are male producers. The Holo2 line has phenotypic marker gene on the X' chromosome called *Wavy.* Female producers (X'X) have the *Wavy* wing phenotype whereas male producers (XX) have wild-type straight wings. To collect only male pupae, crosses between straight-winged females (XX) and males (XO) were used to obtain strictly male progeny from which pupae were collected at the appropriate time approximately 4 weeks later. The collected male pupae were snap frozen in liquid nitrogen and stored at -80˚C until needed.

Hi-C Data Collection

The frozen male pupae were ground into a fine powder. Chromatin conformation capture data was then generated using a Phase Genomics (Seattle, WA) Proximo Hi-C 2.0 Kit, which is a commercially available version of the Hi-C protocol [6]. Following the manufacturer's instructions, intact cells from two samples of finely ground frozen male pupae were crosslinked using a formaldehyde solution, digested using the Sau3A1 restriction enzyme, and proximity-ligated with biotinylated nucleotides. This creates chimeric molecules made from DNA fragments that came close together *in vivo*, but that may not be close together along the genome sequence. As instructed by the manufacturer's protocol, the chimeric molecules were pulled down with streptavidin beads and processed into an Illumina-compatible sequencing library. Sequencing was performed on an Illumina NovaSeq, generating a total of 115.2 million 2x101 bp paired-end reads.

Hi-C assembly correction and scaffolding

The Hi-C data was processed, as described below, according to Phase Genomics recommendations found here: https://phasegenomics.github.io/2019/09/19/hic-alignment-and-qc.html. The paired-end Hi-C reads were first used to produce a corrected assembly. Briefly, the reads were aligned to the primary contigs of Bcop_v1 [5] using BWA-MEM [12] with the -5SP and -t 8 options specified, and all other options default. The output was piped into SAMBLASTER [13] to flag PCR duplicates that were later excluded from analyses. That output was subsequently piped into SAMtools [14] using "-F 2304" to remove non-primary and secondary alignments. Juicebox was used to identify and break contigs at 46 putative mis-joined regions based on disruptions in the expected pattern from Hi-C alignments along the contigs [7,15]. This transformed the input assembly (Bcop_v1) into the corrected assembly (Bcop_v1_corrected). The paired-end Hi-C reads

were then aligned to Bcop_v1_corrected using the same alignment procedure as above, and these alignments were the input to the Hi-C-based scaffolding described next.

The Phase Genomics' Proximo Hi-C genome scaffolding platform was used to create chromosome-scale scaffolds from the Bcop_v1_corrected as described in Bickhart et al. [8]. As in the LACHESIS method [9], the Proximo scaffolding process computes a contact frequency matrix from the aligned Hi-C read pairs that is normalized by the number of DPNII restriction sites (GATC) on each contig. Proximo then constructs scaffolds by optimizing the expected contact frequencies and other statistical patterns in Hi-C data. Using a brute-force approach, approximately 20,000 separate Proximo runs were performed to optimize the number of scaffolds and the concordance with the observed Hi-C data. Finally, Juicebox was again used to correct scaffolding errors, resulting in a final set of four primary chromosome-scale scaffolds each spanning 58-71 Mb (Table 1, Figure 1), a ~363 kb short scaffold, and 57 unscaffolded sequences with 46 marked as “debris”. The “debris” was largely comprised of optical-map gap sequences (NNNNNN) from the input assembly. Specifically, six unscaffolded debris sequences ranging from ~25-50 kb had 100% N content. These sequences were removed from the assembly. Furthermore, 24 sequences had either leading or trailing NNNNNs ranging from 1,489-36,461 bp, making up ~6-96% of the unscaffolded sequences they resided on. The leading and trailing N content was trimmed off these sequences. The remaining 52 cleaned up unscaffolded sequences summed to ~1.8 Mb, with just ~19.5 kb N content. 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 primary chromosome scaffolds that summed to ~297 Mb. The 539 associated contigs from Bcop_v1 that were not included during the Hi-C scaffolding process were also added back to the assembly. These sequences summed to 10.8 Mb, and were ~20 kb on average. Note that we also tried including the Bcop_v1 associated sequences during the Hi-C scaffolding process, but preferred the results when using only the primary assembly. We now refer to the four chromosome-scale scaffolds as the primary assembly, and the other 591 sequences as the associated contigs. Most of the associated contigs are haplotigs, and researchers interested mainly in the chromosome-scale scaffolds can opt to use or ignore them.

Overall, Phase Genomics Proximo Hi-C genome scaffolding yielded the updated genome assembly, Bcop_v2, that has four primary chromosome scaffolds of 58.3, 70.5, 71.0, and 97 Mb (Table 2, Figure 1), which make up ~96% of the entire assembly. The four primary scaffolds together with the 591 associated sequences have an NG50 of ~71 Mb pertaining to only 2 scaffolds, and an expected size of ~81 Mb (Table 1). The assembly retained 209 gaps from the original BioNano scaffolding process, that ranged from 25 bp to ~662 kb with a gap N50 of ~101.5 kb. The optical map gaps are represented as capital N’s in the assembly and are of variable size. Hi-C scaffolding introduced 241 gaps, 5 of which are on the short ~363 kb scaffold, the remainder being in the chromosome scaffolds. The Hi-C gaps are represented as lower-case n’s and are always 100 bp in length.

Anchoring and orienting the scaffolds

Sequences with known chromosomal locations based on previous *in situ* hybridization work were used with BLAST [16] to identify the corresponding chromosome for each scaffold (Table 3). The chromosome assignments were in general agreement with the expected sizes of each chromosome (Table 2) with chromosomes II and X being the smallest, followed by chromosomes III and IV. The total length of autosomes (226 Mb) was also concordant with expectations (218-231 Mb) (Table 2).

Since the locations of multiple sequences for each chromosome are known, it was also possible to orient the chromosome scaffold sequences in the same direction as locus numbers have been assigned on polytene maps [17]. There were no conflicts among the multiple anchors. In other words, each scaffold corresponded to one and only one chromosome, the 4 chromosome-scale scaffolds corresponded to different chromosomes, and when multiple anchors were present for a given chromosome, they all mapped in the expected order along the same scaffold. In addition to unique locus-specific landmarks, the centromere-associated sequence, Sccr, (*Sciara coprophila* centromeric repeat [18]) were in the expected positions. Specifically, chromosomes X, II, and III are acrocentric with their centromeres very close to the beginning of the chromosomes (as oriented). We found Sccr mapped within the first few megabases of the corresponding scaffolds. In contrast, chromosome IV is metacentric with a centrally-located centromere. We found Sccr mapped to more medial positions on the scaffold corresponding to IV. Overall, the agreement with expected sizes, the lack of anchor conflicts, the concordance with the sequence of loci on polytene maps, and the relative positioning of centromere-associated sequence suggests that the Hi-C scaffolding process accurately clustered contigs into chromosome groups and put them in the correct order.

Liftover of gene annotations to the chromosome-scale assembly (Bcop_v2).

There are presently two high quality gene annotation sets produced using the previous genome assembly (Bcop_v1) that was used as the input for Hi-C scaffolding. One was constructed by us [5] using Maker2 [19], which can be found at USDA Ag Data Commons [20] and the USDA i5k Workspace [21]. The other was made by the NCBI Eukaryotic Genome Annotation Pipeline for RefSeq, is called NCBI *Bradysia coprophila* Annotation Release 100, and can be found at NCBI [22-24]. While the FASTA files containing the transcript and protein sequences are in no need of updating, the GFF files that detail the coordinates of the gene models, including their exons, introns, CDS, and UTRs, on Bcop_v1 are not useful for the updated assembly, Bcop_v2. Since Bcop_v2 was produced from the Bcop_v1 sequences, a simple liftover procedure was adequate for creating new GFF files detailing the coordinates of the genes on Bcop_v2. Specifically, we used the program designed for this task, LiftOff [25] (version v1.6.3) with Minimap2 [26] (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.

Using the Maker2 annotation as an example of results, of the 23,117 genes and 28,870 transcripts, only 2 genes with 4 transcripts were not mapped to Bcop_v2. Of the 28,866 transcript models successfully lifted over to Bcop_v2, 28865 (all but one) had perfect full-length alignments to the original transcript sequences, although 53 (<0.2%) were a bit shorter than the original. We can visually confirm in a genome browser that the gene models at the same loci look the same between assemblies and that the lifted over gene models are highly concordant with RNA-seq datasets from various life stages mapped to Bcop_v2 (data not shown) [5].

Conclusions:

Our goal was to upgrade Bcop_v1 [5] to chromosome-scale scaffolds. Bcop_v1 is 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. Bcop_v1 was produced through an extensive evaluation process comparing nearly 100 assemblies. 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. Finally, two gene annotation sets were constructed for Bcop_v1. We produced one using Maker2 [5,19-21] and NCBI produced one for RefSeq [5,22-24]. As studies are ongoing with these gene sets, we also sought to lift over these gene annotation sets directly onto the chromosome-scale scaffolds so researchers can seamlessly transition current studies to the new assembly.

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 polyene maps produced in the 1960s [17] to be consistent with historical research. In the future, the polytene zones in those chromosome maps can roughly be mapped onto their corresponding Bcop_v2 chromosome-scaffold sequences using a combination of landmark sequences identified by FISH, genomic profiling techniques, and assumptions about the sizes of the polytene zones. Indeed, it may be possible to re-purpose the Hi-C data reported here to identify bands and interbands, which have been associated with topologically associated domains (TADs) in *Drosophila* polytene chromosomes using Hi-C data [27-28], across the Bcop_v2 chromosome sequences. Regardless, the male (XO) pupae Hi-C paired-end reads will continue to be useful for correcting mis-assemblies and scaffolding future *Bradysia coprophila* genome assemblies produced with newer long-read datasets (e.g. PacBio HiFi or ultra-long Q20 nanopore reads).

We have also provided the Bcop_v1 gene annotations directly lifted over to Bcop_v2 to enable researchers to seamlessly shift ongoing studies from Bcop_v1 to Bcop_v2. The chromosome-level nature of Bcop_v2 will allow researchers to design experiments and analyze data in a chromosome-specific way. The location and sequences of the telomeres and centromeres are now more obvious and open to future investigation. The X chromosome scaffold will prove useful in identifying the Controlling Element, an rDNA-associated sequence embedded in the short arm of the X upstream of

the centromere. Moreover, the X chromosome scaffold will benefit studies of the X chromosome variant, X-prime (X'), which differs from the X by a large paracentric inversion [1]. There are also interesting biological signals in the Hi-C maps that can be explored further. The notable example here are the three dots on the X chromosome Hi-C map (see the third chromosome cluster in Figure 1) that appear in the expected pattern of the three foldback repeats identified by Crouse [29]. Finally, the *B. coprophila* chromosome scaffolds will allow proper studies of synteny and chromosome evolution across closely and distantly related flies, and may prove useful for researchers working with closely related species (such as *B. odoraphaga*, *B. ocellaris*, *B. impatiens*) when they are working with fragmented genome assemblies by letting them cluster the contigs into likely chromosome groups, and even ordering and orienting them into pseudo-chromosome scaffolds.

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. This project is associated with BioProject PRJNA291918 and BioSample SAMN12533751. All data, including Hi-C sequencing data and lifted over gene models, will be available upon publication of this manuscript.

Abbreviations

Bcop: *Bradysia coprophila*; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; Gb: gigabase pairs; 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; Mb: megabase pairs; NCBI: National Center for Biotechnology Information; PacBio: Pacific Biosciences; PCR: polymerase chain reaction; Sccr: *Sciara coprophila* centromeric repeat ; 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. 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 chromosome identification, scaffold orientation, assembly evaluations, further Hi-C visualizations, annotation liftover, repeat

annotation, and synteny analyses. 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 the correspondents we interacted with at Phase Genomics, including Hayley Mangelson, Kayla Young, Shawn Sullivan, Brian Fan, Andrew Wiser, and Ivan Liachko. We would like to acknowledge Rob Baird and Laura Ross for being early adopters of Bcop_v2. We thank Jennifer Urban for proof-reading the manuscript.

References

1. Gerbi SA. Unusual chromosome movements in Sciarid flies. In: Hennig W, editor. Results and problems in cell differentiation, vol. 13 Germ Line - Soma Differentiation. Berlin, Heidelberg: Springer-Verlag; 1986. p. 71–104.
2. Rasch EM. DNA cytophotometry of salivary gland nuclei and other tissue systems in dipteran larvae. In: Wied BGF, editor. Introduction to quantitative cytochemistry, vol. 2. New York: Academic Press; 1970b. p. 357–97.
3. Gerbi SA, Strezoska Z, Waggener JM. Initiation of DNA replication in multicellular eukaryotes. J Struct Biol. 2002;140(1-3):17–30. https://doi.org/10.1016/S1047-8477(02)00538-5.
4. Rasch EM. Two-wavelength cytophotometry of *Sciara* salivary gland chromosomes. In: Wied BGF, editor. Introduction to quantitative cytochemistry, vol. 2. New York: Academic Press; 1970a. p. 335–55.
5. Urban JM, et al., High contiguity de novo genome assembly and DNA modification analyses for the fungus fly, *Sciara coprophila*, using single-molecule sequencing. BMC Genomics 22, 1-13 (2021).
6. Lieberman-Aiden, E., Van Berkum, N.L., Williams L., Imakaev M., Ragoczy T., Telling A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., lander, E.S., Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326, 289-293.
7. Suhas S.P. Rao*, Miriam H. Huntley*, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, Erez Lieberman Aiden. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” Cell 159, 2014.
8. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, Burton JN, Huson HJ, Nystrom JC, Kelley CM, Hutchison JL, Zhou Y, Sun J, Crisà A, Ponce de León FA, Schwartz JC, Hammond JA, Waldbieser GC, Schroeder SG, Liu GE, Dunham MJ, Shendure J, Sonstegard TS, Phillippy AM, Van Tassell CP, Smith TP. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 49(4):643-650 (2017).
9. 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. Biotech. 31, 1119 (2013).
10. Steffan WA. A generic revision of the family Sciaridae (Diptera) of America north of 892 Mexico. University of California publications in Entomology; 1966.
11. Metz CW. Chromosomes and sex in *Sciara*. Science. 1925;61(1573):212-214.
12. Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 589-595. [PMID: 20080505]
13. Gregory G. Faust, Ira M. Hall; SAMBLASTER: fast duplicate marking and structural variant read extraction, Bioinformatics, Volume 30, Issue 17, 1 September 2014, Pages 2503– 2505, https://doi.org/10.1093/bioinformatics/btu314

14. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]
15. Neva C. Durand*, James T. Robinson*, Muhammad S. Shamim, Ido Machol, Jill P. Mesirov, Eric S. Lander, and Erez Lieberman Aiden. "Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom." Cell Systems, July 2016.
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
17. Gabrusewycz-Garica N. Cytological and autoradiographic studies in *Sciara coprophila* salivary gland chromosomes. Chromosoma. 1964;15(3):312–44. https://doi.org/10.1007/BF00321517.
18. Escribá MC, Greciano PG, Méndez-Lago M, De Pablos B, Trifonov VA, Ferguson-Smith MA, et al. Molecular and cytological characterization of repetitive DNA sequences from the centromeric heterochromatin of *Sciara coprophila*. Chromosoma. 2011;120(4):387–97. https://doi.org/10.1007/s00412-011-0320-2.
19. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12(1):491. https://doi.org/10.1186/1471-2105-12-491.
20. https://doi.org/10.15482/USDA.ADC/1522618
21. https://i5k.nal.usda.gov/content/data-downloads
22. https://ftp.ncbi.nlm.nih.gov/genomes/all/annotation_releases/38358/100/
23. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Bradysia_coprophila/100/
24. https://ftp.ncbi.nlm.nih.gov/genomes/all/annotation_releases/38358/100/GCF_014529535.1_BU_Bcop_v1/
25. Shumate, Alaina, and Steven L. Salzberg. 2020. "Liftoff: Accurate Mapping of Gene Annotations." Bioinformatics , December. https://doi.org/10.1093/bioinformatics/btaa1016.
26. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34:3094-3100. doi:10.1093/bioinformatics/bty191
27. Eagen KP, Hartl TA, and Kornberg RD. Stable chromosome condensation revealed by chromosome conformation capture. Cell. 2015 November 5; 163(4): 934–946. doi:10.1016/j.cell.2015.10.026.
28. Spradling AC. Symposium - Germ Cells, Imprinting, Gene Dosage, and Regulation: Polytene Chromosome Structure and Somatic Genome Instability.Cold Spring Harb Symp Quant Biol 2017 82: 293-304; Published in Advance November 22, 2017, doi:10.1101/sqb.2017.82.033670
29. Crouse HV. X heterochromatin subdivision and cytogenetic analysis in Sciara coprophila (Diptera, Sciaridae): Centromere localization. Chromosoma. 1977;63(1):39-55.

468 **Table 1 – Assembly Contiguity Statistics comparing the input (Bcop_v1) into Hi-C**
469 **scaffolding to the resulting chromosome-scale assembly (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
Total length	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

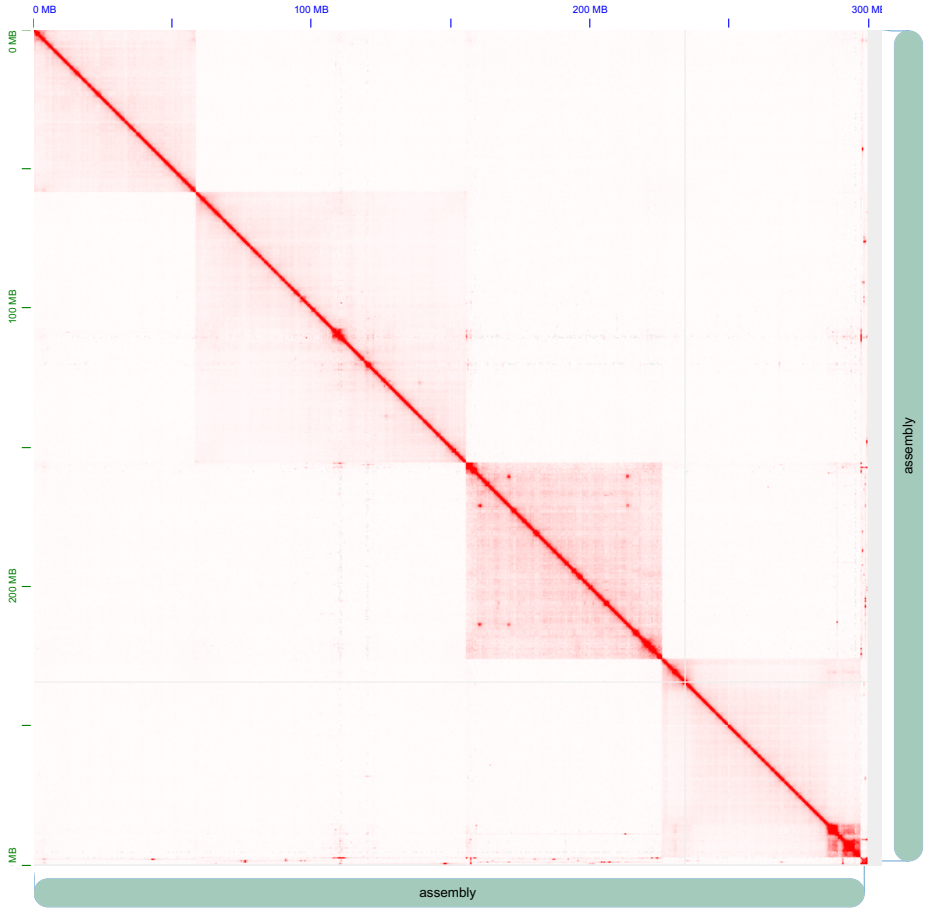
470 **Expected genome size used rather than assembly length: 280 Mb (Urban et al 2021).*
471
472

473 **Table 2**

	Bcop_v1 (Urban et al 2021)	Bcop_v2 (This paper)	Expected size*
Total primary length classified as autosomal	223.8-232.8 Mb	226.5 Mb	218-231 Mb
Total length classified as X	62-71 Mb	70.5 Mb	48.9-62.2 Mb
Total length classified as II	13.1-28.5 Mb	58.3 Mb	62.2-66 Mb
Total length classified as III	5.4-12.5 Mb	71.0 Mb	66.6-70.7 Mb
Total length classified as IV	34.7-46.4 Mb	97.1 Mb	88.8-94.2 Mb

474 ** Expected size calculations are from Supplemental Table S1 in Urban et al 2021*
475
476

477 **Figure 1 – Hi-C interaction frequency map shows four chromosomes**



478