Towards telomere-to-telomere genome assemblies of *Drosophila melanogaster*

Nicolas Altemose[1], Susan E. Celniker[2], Mahul Chakraborty[3], Cécile Courret[4], J.J. Emerson[3], Gary H. Karpen[1, 2], Bernard Y. Kim[5], Charles H. Langley[6], Sasha Langley[1], Amanda M. Larracuente[4], Barbara G. Mellone[7], Karen H. Miga[8], Danny E. Miller[9,10], Rachel J. O'Neill[7], Adam M. Phillippy[11], Brandon D. Pickett[11], Harsh G. Shukla[3]

1. University of California, Berkeley, Berkeley, CA 2. Lawrence Berkeley National Lab, Berkeley, CA 3. University of California, Irvine, Irvine, CA 4. University of Rochester, Rochester, NY 5. Stanford University, Stanford, CA 6. University of California, Davis, Davis, CA 7. University of Connecticut, Storrs, CT 8. University of California, Santa Cruz, CA 9. University of Washington, Seattle, WA 10. Seattle Children's Hospital, Seattle, WA 11. National Human Genome Research Institute, NIH, Bethesda, MD

Abstract

The goal of genome sequencing and assembly is to capture all sequence features that play critical roles in organisms and to join them together as parts of complete, gapless chromosome reference assemblies. However, in *Drosophila melanogaster*, the most complete assemblies are at least 15-20% smaller than the known genome size. The segments missing from assemblies are almost entirely composed of repetitive sequences, primarily transposable elements and satellite repeats concentrated in pericentromeric heterochromatin. These missing regions not only contain essential genes, they also harbor other elements that play crucial roles in genome stability, chromosome segregation, protein translation, and TE repression. To recover these important regions, we are building telomere-to-telomere assemblies of multiple *D. melanogaster* strains, including the genome reference strain iso-1. Our goal is to produce complete, extremely accurate (fewer than 1 error per megabase) assemblies for all three autosomes and the X and Y chromosomes. Efforts by the human Telomere-to-Telomere (T2T) Consortium have pioneered the completion of virtually gapless chromosome assemblies that span large repetitive arrays, including satellites, scrambled transposable elements, and ribosomal DNAs. These approaches leverage ultra-long sequencing reads to untangle assembly graphs derived from highly accurate long sequence reads. By applying this approach to additional strains, we can study variation in chromosome structures that have previously resisted scrutiny, like centromeres. This open, collaborative initiative aims to produce a gapless assembly of *D. melanogaster*, outline best practices for extending this approach to other strains and species, and support public accessibility of data releases and methodologies.

Goals

The initial phase of the project include the following:

1. Select initial stocks for sequencing
2. Ensure isogeny
2. Generate accurate long reads and ultralong reads
3. Generate T2T assemblies for multiple stocks
4. Share data, assemblies, and methods immediately upon generation

Challenges

Summary Large tandem arrays of repeats are refractory to genome assembly. Longer reads, more accurate reads, and material with no genetic variation make resolving such recalcitrant regions easier.

Telomere-to-telomere sequencing efforts are primarily limited by large spans of highly repetitive tandem arrays. These regions are recalcitrant to standard sequencing approaches, and result in discontinuities in chromosome sequences and/or misassemblies. To resolve such challenging regions, sequencing reads need either to span them outright or to span imperfections among the repeat units to form anchors that can be bridged by overlapping reads (**Figure 1**). Ideally, sequencing reads for T2T projects should have the following properties:

1. be long enough to span most repeats;
2. exhibit a sufficiently low error rate that unique anchors in longer repeats are not confounded by sequencing error
3. be long enough to span most unique anchors units in longer repeats.

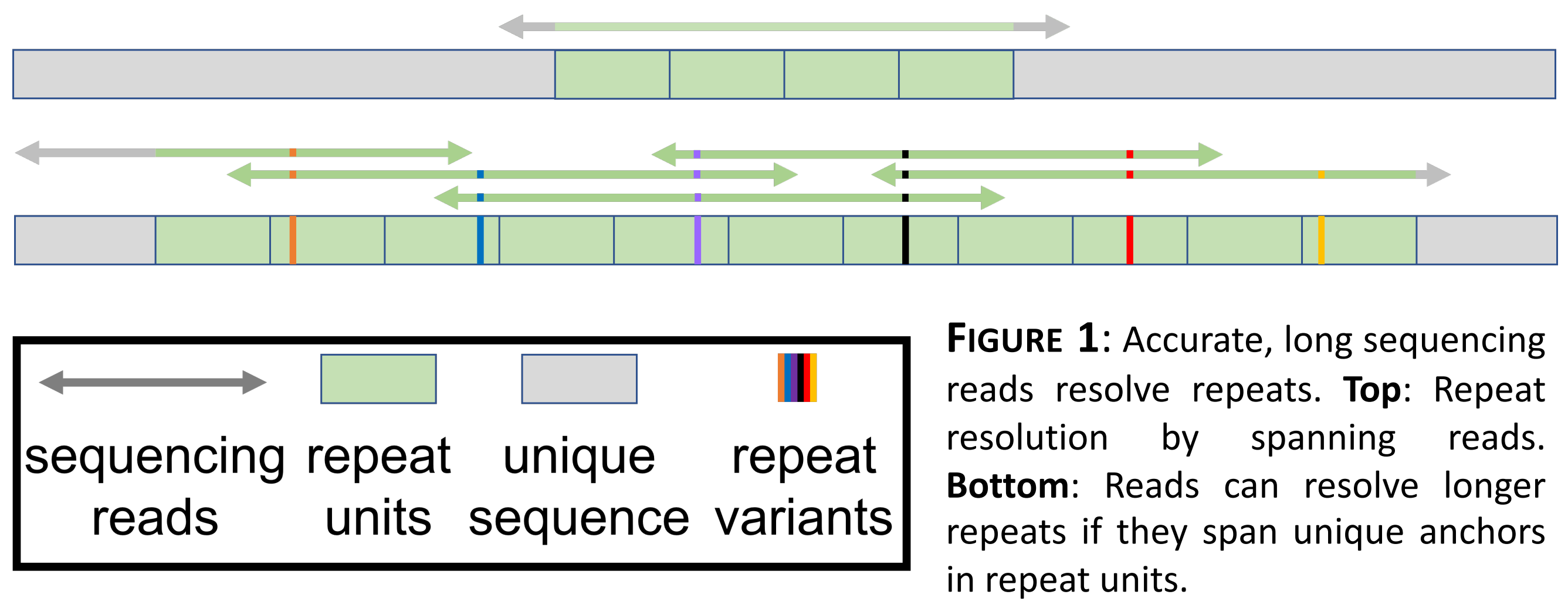

Figure 1: Accurate, long sequencing reads resolve repeats. **Top**: Repeat resolution by spanning reads. **Bottom**: Reads can resolve longer repeats if they span unique anchors in repeat units.

Repeat arrays are frequently much longer than typical long reads, spanning hundreds of kilobases or even megabases. Such repeats cannot be resolved within single reads (**Figure 1 Top**). In such regions, assemblies rely on accurate long reads to bridge variants within repeat arrays (**Figure 1 Bottom**). Even so, some repeats exhibit too few unique anchors, requiring even longer reads. Currently 1 & 2 can be satisfied with PacBio HiFi reads and 3 by Oxford Nanopore ultralong reads.

Strain Choice

Summary The first Fly T2T strain is FlyBase's reference strain iso-1[1]. An ideal second strain will have little genetic variation and be divergent to iso-1 in recalcitrant regions like those flanking centromeres.

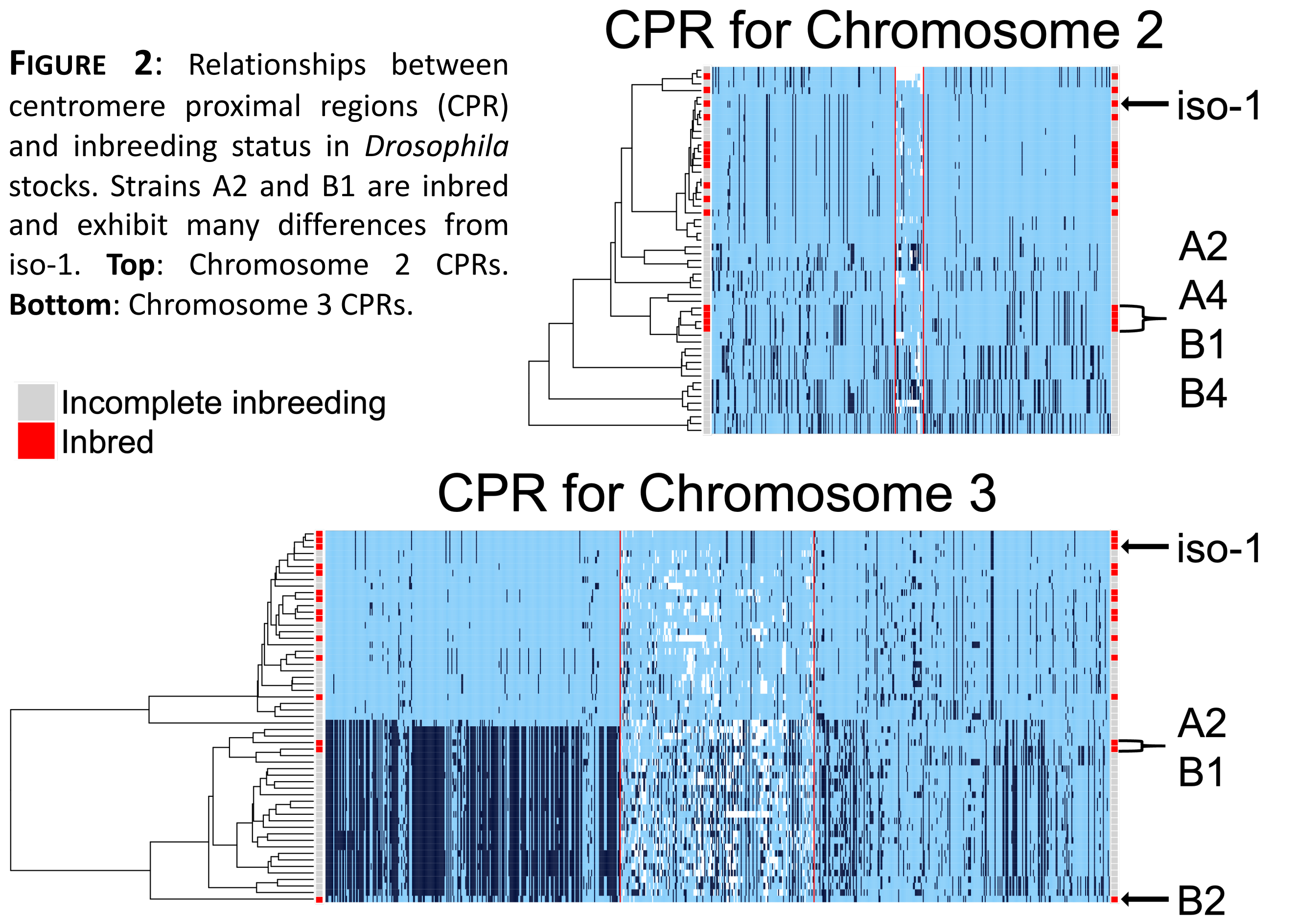

Figure 2: Relationships between centromere proximal regions (CPR) and inbreeding status in *Drosophila* stocks. Strains A2 and B1 are inbred and exhibit many differences from iso-1. **Top**: Chromosome 2 CPRs. **Bottom**: Chromosome 3 CPRs.

Many archived fly stocks maintain measurable genetic variation. This might be residual heterozygosity from their initial collection or mutation that has accumulated over time. Either way, segregating variation introduces challenges in resolving repeats, primarily by making it more difficult to distinguish intrachromosomal variation in repeat units from allelic variation. Eliminating such variation makes resolving repeats easier.

A survey centromere proximal regions among diverse strains indicates that the Drosophila Synthetic Population Resource[2] strains A2 and B1 possess divergent centromere proximal regions (CPR) relative to iso-1 and exhibit low heterozygosity (**Figure 2**). Because B1 has gone extinct, we will use A2. Other candidate strains are either more closely related to iso-1 in one more CPRs (e.g. Canton-S) or have measurable heterozygosity in stock center samples (e.g. Oregon-R) (**Figure 2**).

Satellite heterozygosity

Summary A major obstacle in resolving repeats (e.g. satellite repeats) is unrecognized heterozygosity that may be conflated with variation in repeat units. Verifying a low level of satellite heterozygosity means assembly approaches do not need to grapple with phasing of haplotypes in these recalcitrant genomic regions. The strains iso-1 and A2 have very low levels of satellite heterozygosity.

Common assembly methods operate on graphs constructed from sequencing reads. Any heterozygosity present in the sequenced sample is reflected in the form of more complex graphs. While diploid data does offer the prospect of resolving genetic variation in a single sample, the approaches for resolving isogenic or near-isogenic samples are more reliable and less prone to the types of assembly errors that are common when heterozygosity may be conflated with variation across repeat units.

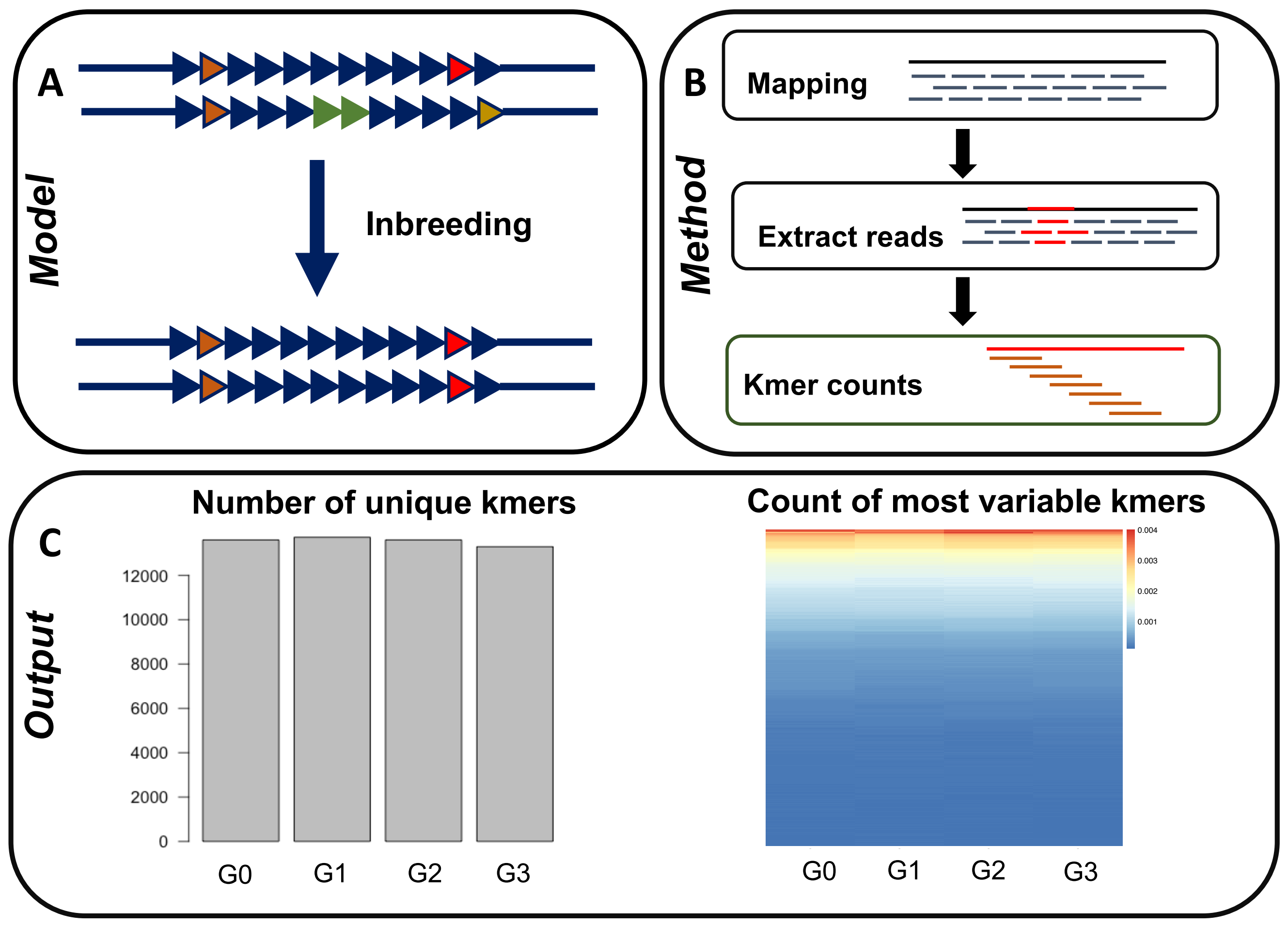

Figure 3: Estimating homozygosity following inbreeding with kmer count analyses. **A**: In the presence of measurable heterozygosity in repeat units, our model predicts the number of unique kmers will decrease across repeated rounds of inbreeding (G1-G3), eventually asymptoting as variation is depleted. **B**: To estimate satellite repeat diversity, we collected short reads mapping to specific repeats and counted kmers. **C**: For the *Responder* satellite in iso-1, kmer diversity does not change. The number of unique kmers (left) and the heatmap (right) show stable kmer counts across generations of inbreeding, suggesting that the original stock (G0) was already highly homozygous. We see similar results for other satellites on chromosomes 3 (*dodeca*, *353bp* and *356bp*) and X (*359bp* and *NTS*) in both iso-1 and A2.

Surveys of kmer diversity across inbreeding generations typically show drops when variation is segregating in the parental sample. In our samples (**Figure 3**), there is no significant change as a result of inbreeding. This may not be surprising since iso-1 was isogenized by genetic chromosome extraction[3] and A2 was subjected to intense sibling mating[2]. We're confident that the strains chosen for this phase will pose minimal problems related to heterozygosity.

T2T Assembly Approach

Overview of progress

Points below in black have been completed with preliminary data. Improvements to them are ongoing. Points in gray are anticipated to be completed in April or May.

1. Isolate pure, high molecular weight DNA from 2hr embryos
2. Sequence PacBio HiFi libraries (highly accurate long reads)
3. Generate string graph from HiFi reads
4. Sequence Oxford Nanopore ultralong reads (extremely long reads)
5. Lay out unresolved regions of the assembly by using ultralong reads to untangle repeat-induced snarls in the graph

As a test of T2T approaches on material from iso-1 stocks, we analyzed an existing HiFi dataset from 2021. The reads were trimmed to 18 Kb in homopolymer-compressed space then corrected and assembled into unitigs with Canu[4-5]. Unitigs were assembled into contigs using a custom string graph approach[6]. Assembly graphs were visualized with Bandage[7]. Node colors are assigned to chromosomes based on alignments to FlyBase Release 6[1]. Results are shown in **Figure 4**.

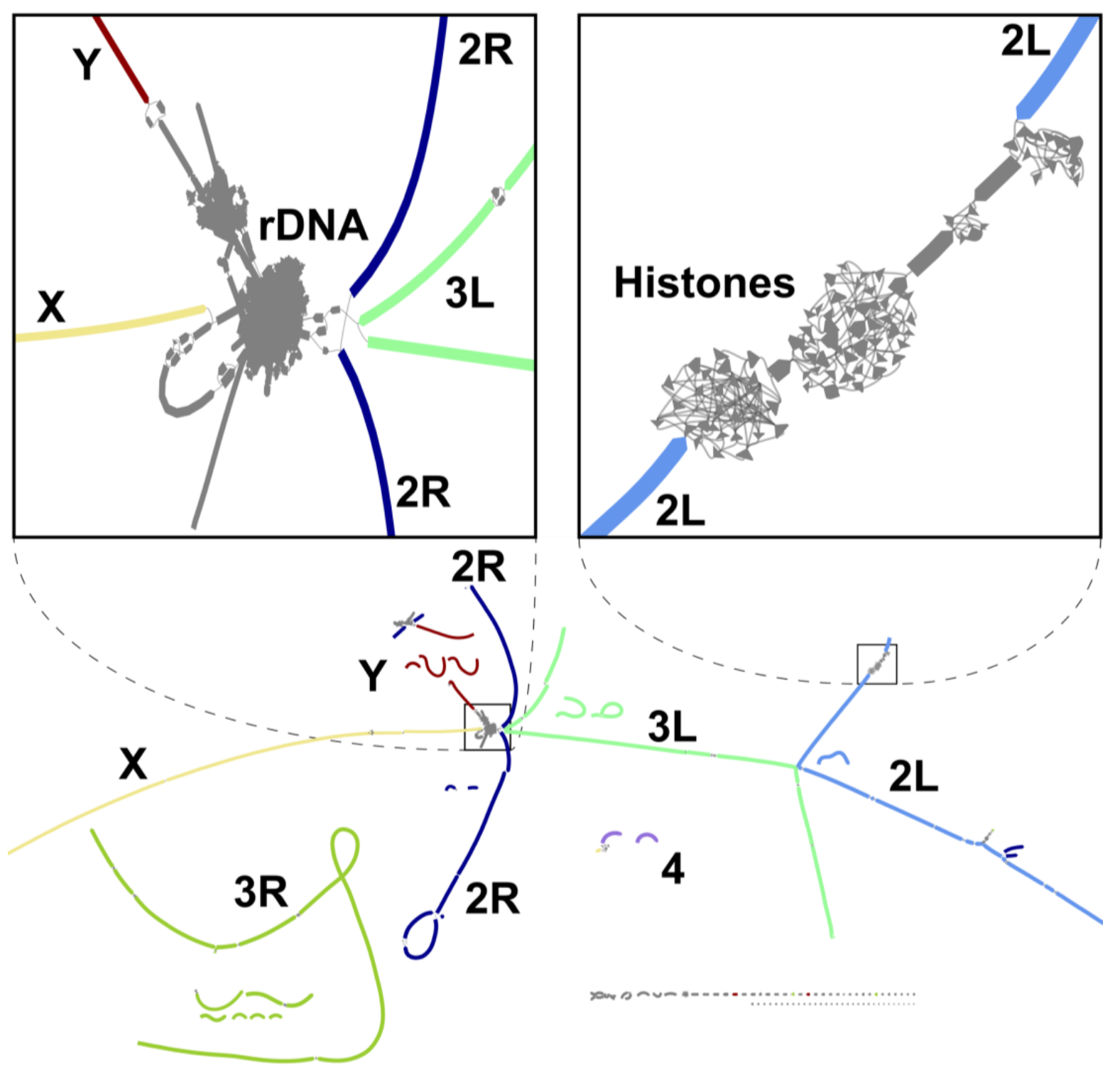

Figure 4: Bandage plot[7] of the assembly graph. **Bottom**: Plot of full genome. Most of the genome is assembled into very long contigs. Chromosome arms 2L, 2R, 3L, 3R, and X are each primarily contained in a few large, connected components. The principal challenges for resolving the rest of the genome include: the centromeres; the rDNA; the histone cluster; the Y chromosome; the dot chromosome. **Top left**: The rDNA regions on the X and Y chromosomes **Top right**: The histone cluster on 2L. This assembly approach is conservative and errs on the side of accuracy at the expense of contiguity Unresolved regions are likely to be resolved with a combination of longer modern HiFi libraries and ultralong Nanopore reads.

The conservative results presented in **Figure 4** suggest that ongoing work will lead to resolution of these challenging regions, consistent with other T2T approaches[6]. We are currently isolating gDNA from 2hr embryos, the completion of which is anticipated in late April or May. The addition of HiFi sequences from longer libraries and ultralong Nanopore reads will likely resolve these problematic regions.

We also assembled the same data using Hifiasm[8], which is more aggressive in resolving repeats. Like the results from HiCanu[5], the centromeres, the Y chromosome, and the rDNA could not be resolved (see **Poster 298C**). However, the histone cluster was assembled (**Figure 5A; Poster 298C**). Combined with data from another DSPR strain (A4) we were able to compare these preliminary results, revealing extensive variation in repeat unit content and arrangement between the alleles (**Figure 5B**). We will revisit these regions after T2T assembly.

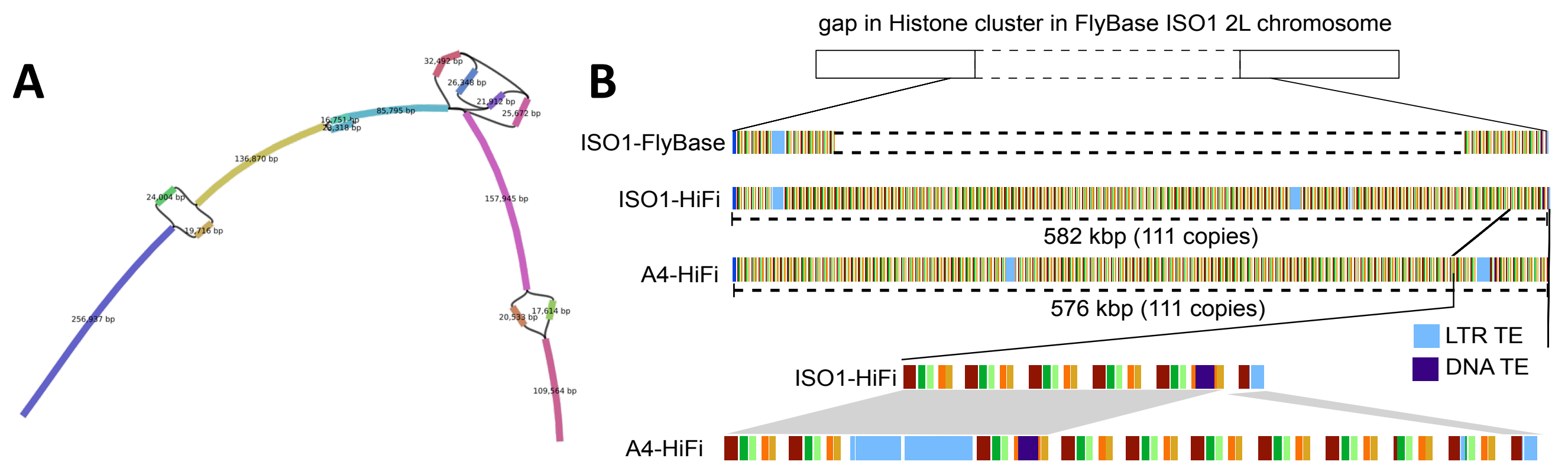

Figure 5: An alternative assembly of HiFi data in *Drosophila melanogaster*. More repetitive regions were recovered with HiFi data than in previous assemblies, including the histone cluster and part of the rDNA (**Poster 298C**). **A**: A Bandage[7] plot showing a much simpler graph structure in the histone cluster region. **B**: Comparison between iso-1 and DSPR's A4 in the histone cluster indicating both differences in the configuration and content of the array.

References

1. Hoskins RA et al. Genome Res. 2015; doi: 10.1101/gr.185579.114.
2. King EG et al. Genome Res. 2012; doi: 10.1101/gr.134031.111.
3. Brizuela et al. Genetics. 1994; doi: 10.1093/genetics/137.3.803.
4. Koren S et al. Genome Res. 2017; doi: 10.1101/gr.215087.116.
5. Nurk S et al. Genome Res. 2020; doi: 10.1101/gr.263566.120.
6. Nurk S et al. Science. 2022; doi: 10.1126/science.abj6987.
7. Wick RR et al. Bioinformatics. 2015; doi: 10.1093/bioinformatics/btv383.
8. Cheng H et al. Nature Methods. 2021; doi: 10.1038/s41592-020-01056-5.