Improving the Genome Assembly of the Upland Chorus Frog *Pseudacris feriarum*

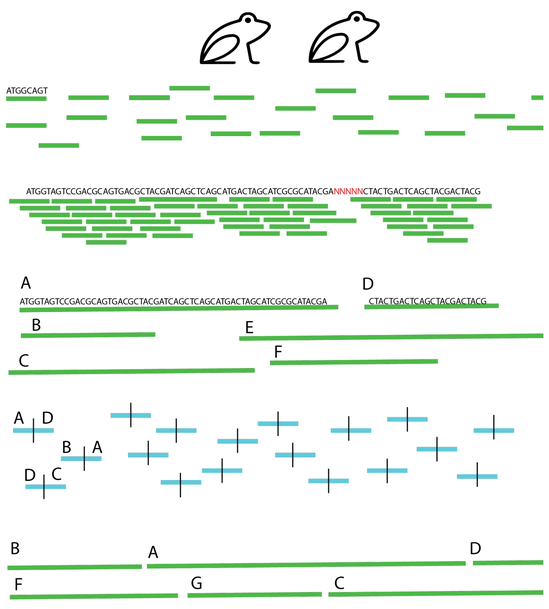
Kevin Ziegler[1], and Lemmon A.[1]

[1]Department of Scientific Computing, Florida State University, Tallahassee, United States

Abstract

We improve the draft de novo genome assembly of *Pseudacris feriarum* using HiC contact maps (High-throughput Chromosome Conformation Capture). The previous version of the assembly included large scaffolds, but not chromosome scale structures hundreds of megabases in size. Here we show the first step of a scaffolding algorithm, 3d-dna, which arranges contigs into pseudo chromosomes; these chromosomes are roughly the correct size and number for *Pseudacris feriarum* but have many mis-joins. After careful visualization of these mis-joins, we diagnose the root cause of mis-joins being multi-mapping HiC reads. Multi-mapping reads most commonly produce the distinct signature of heterozygosity, and less frequently cause mis-joins by mapping to "repeats" dispersed throughout the genome. We show steps used to filter out contigs containing the key signature of heterozygosity and remove multi-mapping repeat reads causing mis-joins. Once filtering is complete, future work will involve re-scaffolding using our cleaned dataset to produce an assembly containing less mis-joins.

Genome Assembly Review



1) Collect Organism

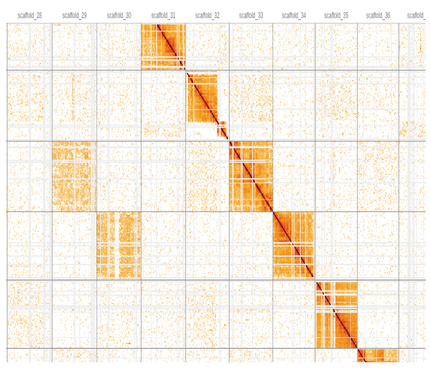
2) DNA Extraction and Sequencing

3) Overlap Reads into Continuous Sequences

4) Sequence Linking Fragments

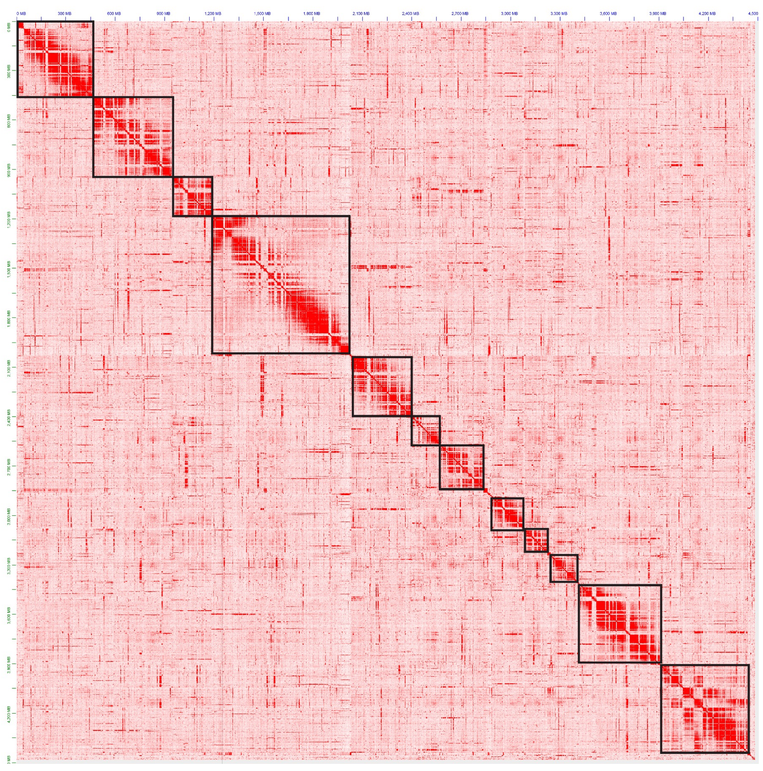
5) Order Contigs into Larger Scaffolds

Previous Draft Assembly



- Lacked large scale chromosome structures (100s of megabases)
- Large scaffolds were 10s of megabases
- Contained mis-joins between large pieces

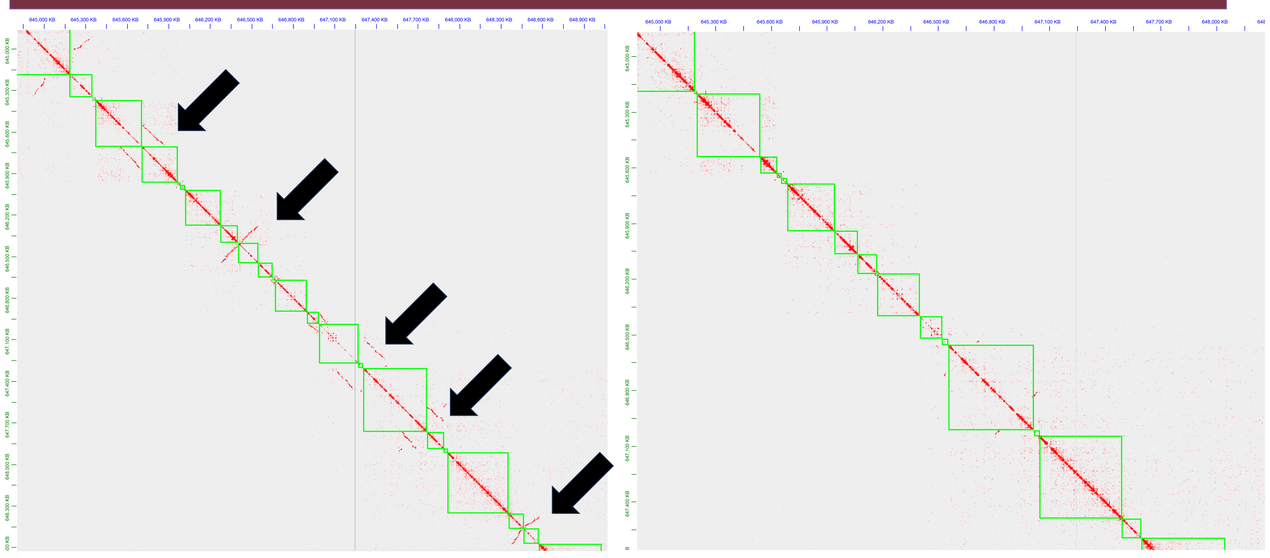
3d-DNA Scaffolder Using Juicer and Juicebox



Round One of Re-Scaffolding

- The red color represents the number of linking reads (HiC data) connecting contigs. Contigs were assembled using Pacbio CLR reads and Canu [7]
- We mapped hic reads with bwa [1] and Juicer [2], 3D-DNA [3] was used to scaffold, and visualization was done in Juicebox [4]
- Distance between fragments in linking reads follow a negative binomial distribution. Most reads will link regions close together; this is why the diagonal has the highest interaction frequency
- Large scale pseudo chromosomes structures are present (100s of Mbs in size). Karyotype from a related species, *Pseudacris triseriata*, finds 12 haploid chromones and an expected genome size of 4.37 Gb [6].
- Chromosomes aren't well defined
- Many mis-joins present (breaks in red signal)
- The ordering of contigs produced has far too many errors for downstream analysis.

Removing Excess Heterozygosity

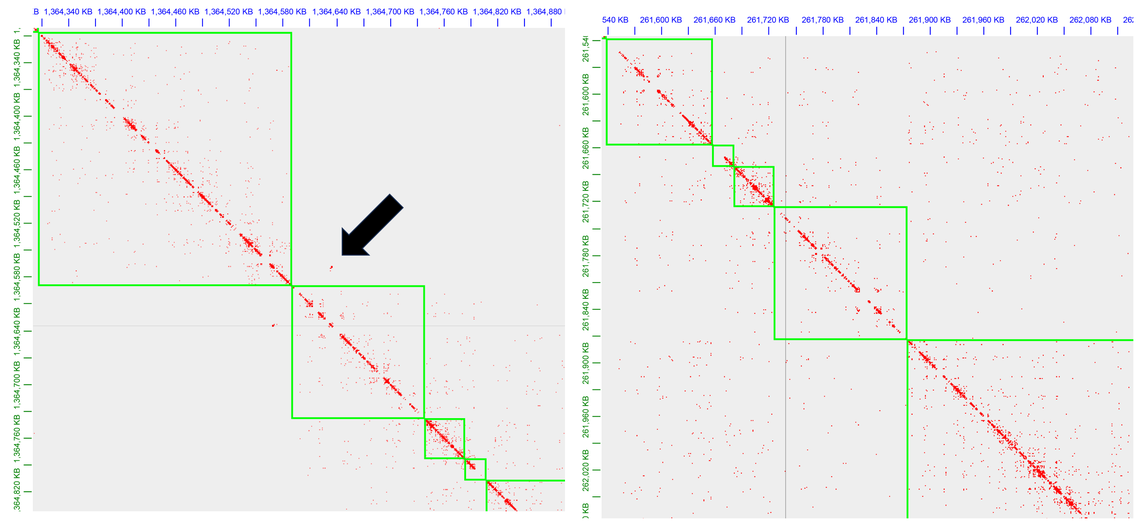


Left: Example of duplicated heterozygous sequences indicated by parallel or perpendicular lines next to the diagonal and **Right**: the same regions with duplicated heterozygous sequences removed.

- The sequences are duplicated, there are two copies one from each parent. The assembler failed to collapse these duplicate regions
- Our original assembly was almost twice as large we expected it to be due to heterozygosity
- We purged duplicate heterozygous sequences using a standard method, Purge Dups [8]. This dropped the duplication rate down from 34% to 7%
- We wrote our own procedure to purge the remaining large regions of heterozygosity. We dropped duplication rate from 7% to 1.8%

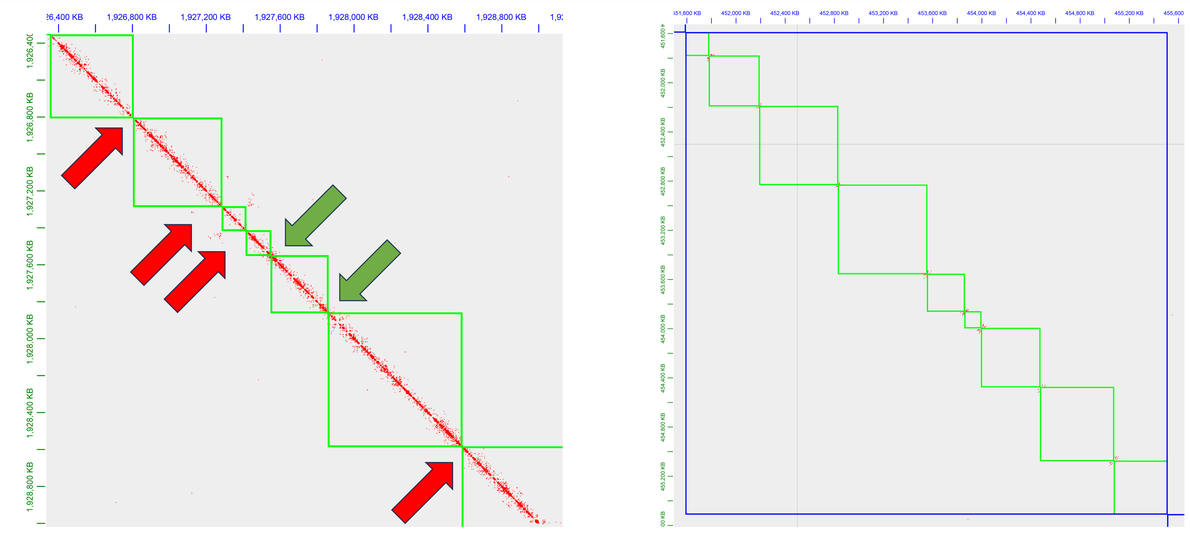
1. Filter for reads which map to two contigs
2. Keep contigs with over 3,000 connections
3. Ensure the connections span over 20kb (to avoid removing repeat regions)
4. Ensure the connections are distributed over at least 20% of the length of the smaller contig

Removing Multi-Mapping Reads



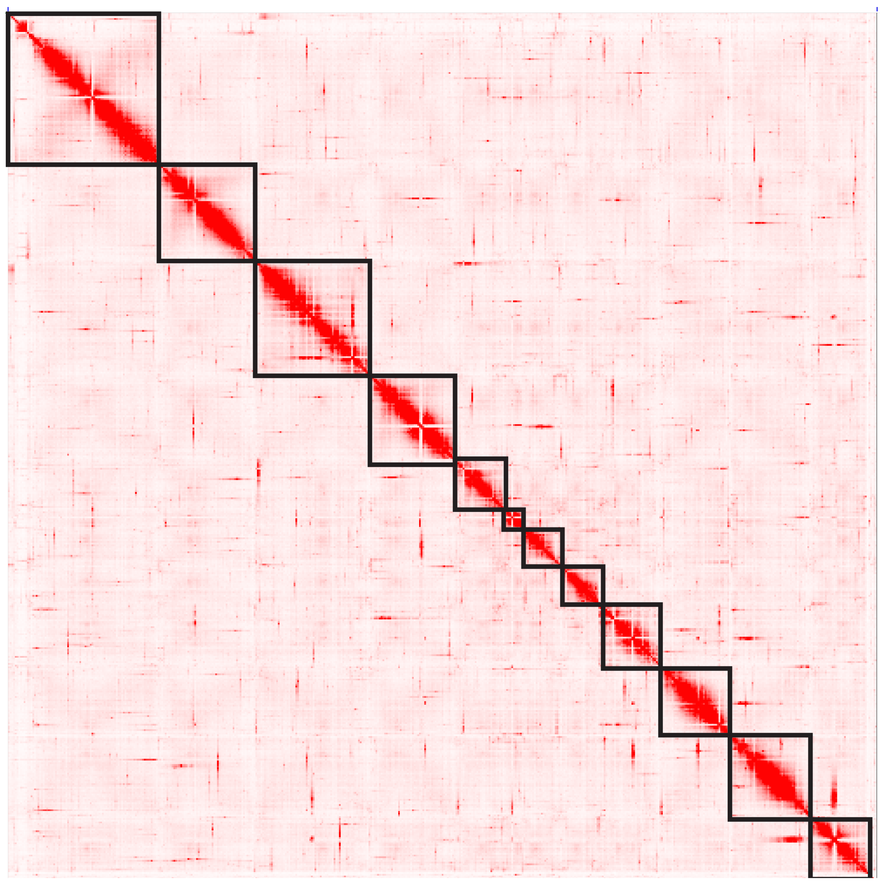
Left: Example of Multi mapping reads causing a mis-join. The sequence for the small region off the diagonal indicated by the arrow has about 100 hic reads connecting the two contigs. The sequence for this small stretch is similar and reads from the diagonal map ambiguously to this small region. **Right**: If we remove reads which map ambiguously to multiple locations we remove this potential cause of mis-joins. The same contig now has better interaction frequency with its new neighbors.

Scaffolding Intermediate Chicago Linking Data



Left: Noisy data throws off standard scaffolders relying on linking counts between contigs. Correct joins are represented by green arrows and incorrect joins are represented by red arrows. **Right**: We wrote our own scaffolding method which relies on not only the number of connections between contigs but also on the area they cover; we rely on area because it ensures the links are distributed more evenly; this mimics the normal interaction frequency better. We used the convex hull approach [5] to measure the area covered by linking reads which fall within the first 40,000 bases of the end of contigs.

Re-Scaffolding without Multi-Mapping Reads



Scaffolding with Only Multi-Mapped Reads Removed

- Rerunning scaffolding after removing multimapping reads improves assembly quality.
- There are still 12 chromosomes, but their stop and start boundaries are more refined. Centromeres are also present in many.
- Mis-joins are still present, but their frequency has decreased
- Scaffolding is currently being run again with a dataset which has heterozygosity removed and has intermediate Chicago data incorporated.
- Once this next round of scaffolding is complete. We will run the same downstream steps applied to the draft assembly:
  - Gap closing
  - Polishing
  - Repeat Annotation
  - Gene Annotation

References

[1] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.
[2] Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems*, *3*(1), 95-98.
[3] Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., ... & Aiden, E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92-95.
[4] Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems*, *3*(1), 99-101.
[5] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272.
[6] Gregory, T.R. (2024). Animal Genome Size Database. http://www.genomesize.com.
[7] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research, 27(5), 722-736.
[8] Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics, 36(9), 2896-2898.

Thank you to David Beamer for sample collection and thank you to members of the Lemmon lab for setting up the genome assembly project!