Draft Genome 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

Introduction

The Upland Chorus Frog, *Pseudacris feriarum*, is a small Anuran in the family Hylidae that can primarily be found throughout the Eastern US. *P. feriarum* provides a unique opportunity to study the causes and mechanisms of speciation. Decades of work by Drs. Emily and Alan Lemmon have shown specific populations of *P. feriarum* within the southeastern US are likely undergoing incipient speciation based on female call preference[1]. The assembled genome of *P. feriarum* will be an important tool in pinpointing exactly which genes drive female call preference and therefore speciation.

We performed the first de novo genome assembly of *P. feriarum*, only the 13th Anuran and first member of the family Hylidae to have its genome assembled. The current results are reported in this poster, but we are still working to improve the assembly over time. The ideal end point is having one-character strings representing each of the 12 chromosomes. Future work will entail using the newly assembled genome in a genome wide association study to narrow down which specific genes are responsible for speciation in specific populations of *P. feriarum.*

Methods

The project to assembly the genome of P. *feriarum* has been in the works since 2015. An overview of the methodology and data used for this assembly are shown in **Figure 1**. Previous assemblies have been very fragmented with short contigs (assembly sequences). This new assembly benefits heavily from recent technological innovations in both algorithms and sequencing. My contribution started with the third orange box; the current assembly would not have been possible without much work from others in the Lemmon lab and David Beamer. The workflow shown here stems from reading other papers about vertebrate genome assembly and copying their workflow using open-source software applicable to the types of sequencing already collected by the Lemmon Lab[2]. Presented here is a simplified version of the pipeline used for our genome assembly generated from dozens of attempts with different software, data and parameters.

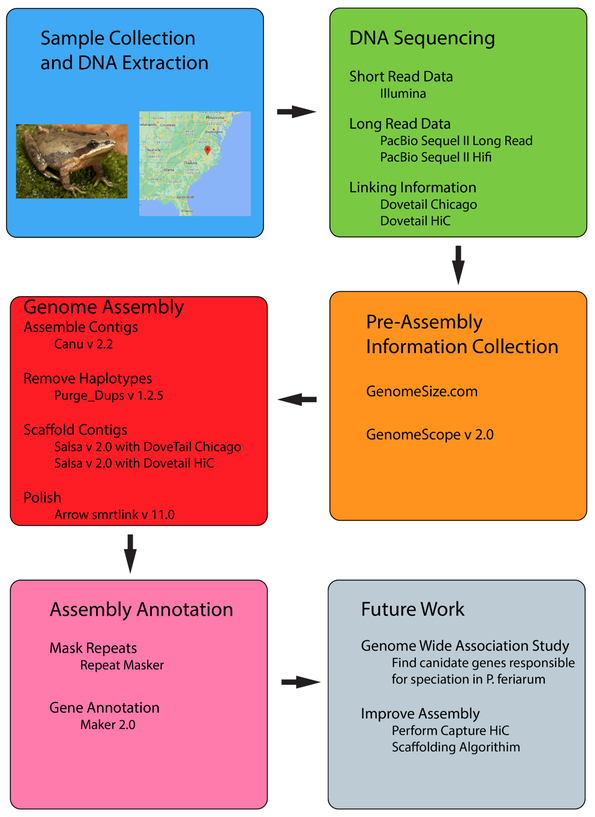

Figure 1. Workflow/Pipeline of Genome Assembly. The process starts with sample collection and sequencing. The genome is first assembled into smaller pieces called contigs; these contigs are then iteratively corrected and grown. The steps here are simplified as most steps have pipelines of their own.

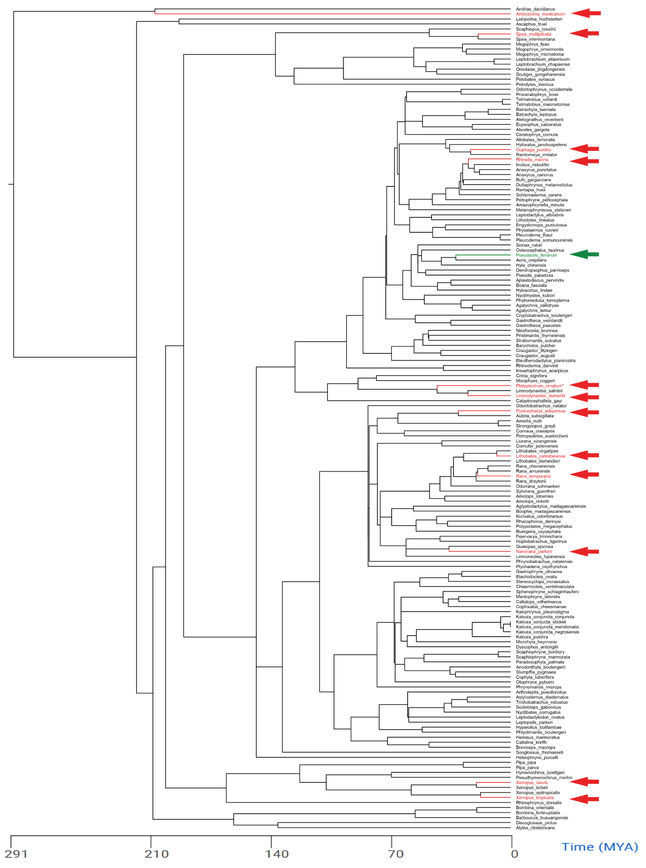

Figure 2. Previous Anuran genome assemblies[3]. Previous assemblies are represented with red arrows; 13th Anuran genome assembly, Pseudacris feriarum, green arrow. There are over 6,000 species of Anurans, only ~150 are shown here. There are thousands and thousands of undiscovered genes and biological pathways that serve undiscovered purposes.

Results

The basic results are provided in **Table 1**; the table reports the size of pieces making up the assembly at different stages. Judging the results of this assembly is very challenging. Statistical properties of the genome being assembled and the number of resources available to perform the assembly greatly impact what the expected level of completeness should be. The genome of *P. feriarum* is quite large ~ 4.4 billion base pairs, it has high heterozygosity ~1.4%, and high repeat content ~45%; previous results have shown that these three statistics correlate with increasingly fractured assemblies[2]. Out of the currently published Anuran assemblies our scaffold sizes are about in the middle of the pack. Whether or not the assembly is correct is a separate difficult to answer question. Our BUSCO scores (a metric at assessing completeness) finds about 83.1% of ancestral genes; this is on par with previous Anuran Assemblies [4][5].

Stage	NG 50	NG 70	NG 90	NG 100	Assembly Size	BUSCO
Canu Contigs	171,603 bp in 11,265 Pieces	98,147 bp in 22,735 Pieces	44,862 bp in 44,901 Pieces	1,083 in 71,716 Pieces	7,421,933,773	-
Purge Dups	243,385 bp in 4,949 Pieces	137,053 in 9,864 Pieces	55,196 bp in 13,797 Pieces	1,100 in 34,439 Pieces	4,482,116,604	C:75.1%[S:68.1%,D:7.0%], F:10.2%,M:14.7%,n:5310
Scaffold Chicago	885,256 bp in 1,382	439,435 bp in 2,804 Pieces	103,353 bp in 6,821 Pieces	1,113 bp in 18,371 Pieces	4,482,115,504	-
Scaffold HiC	3,082,713 bp in 300 Pieces	949,910 bp in 848 Pieces	156,683 bp in 3191 Pieces	1,113 in 12,874 Pieces	4,482,115,504	-
Polished Arrow	3,084,769 bp in 300 Pieces	950,494 bp in 849 Pieces	156,725 bp in 3193 Pieces	1,105 bp in 12874 Pieces	4,487,389,031	C:83.1%[S:77.1%,D:6.0%], F:6.2%,M:10.7%,n:5310

Table 1. Statistics of the genome assembly. NG X means X% of the genome must be NG X size or larger. The final scaffolds contain half the genome (NG 50) in only 300 pieces of size at least 3 million base pairs.

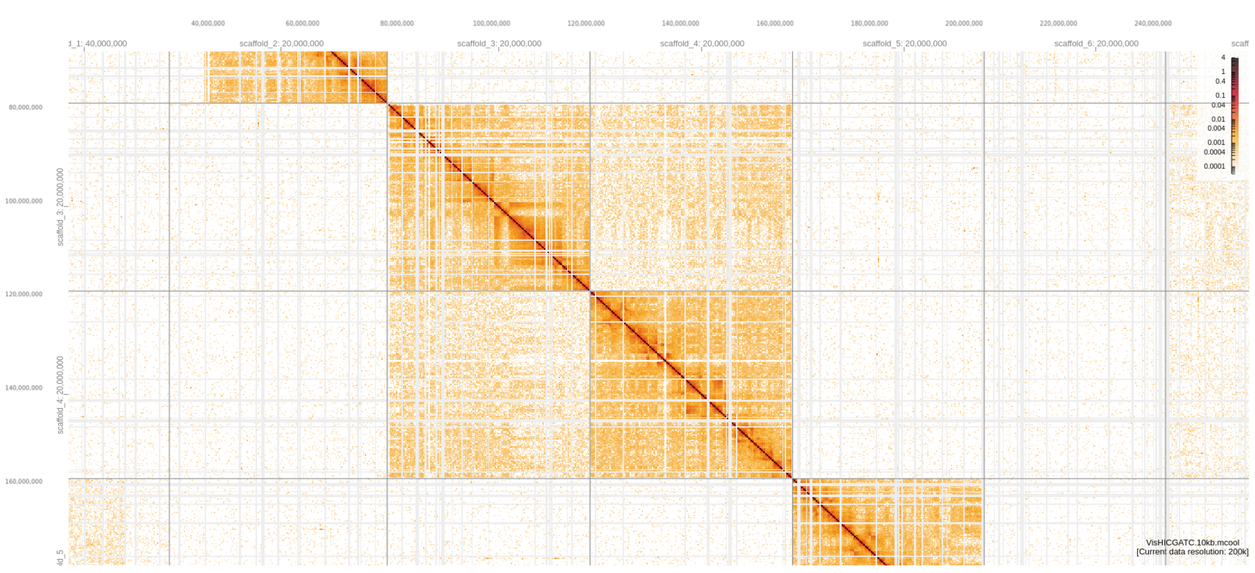

Figure 3. Correct Scaffolding Visualized with HiGlass[6]. Scaffolding is performed using long-range linking data; HiC measures the frequency at which DNA fragments associate in 3d space. Regions close together within a chromosome will interact frequently. The highest interaction frequency of any region will be with itself (x=y). This association can be seen on the diagonal of figure 3. Contigs which have interaction frequencies much higher than background noise should be joined together. This is correctly done within scaffold 3 and within scaffold 4. Scaffold 3 and 4 are not joined, but based off interaction frequency they are close together or adjacent to one another.

Results

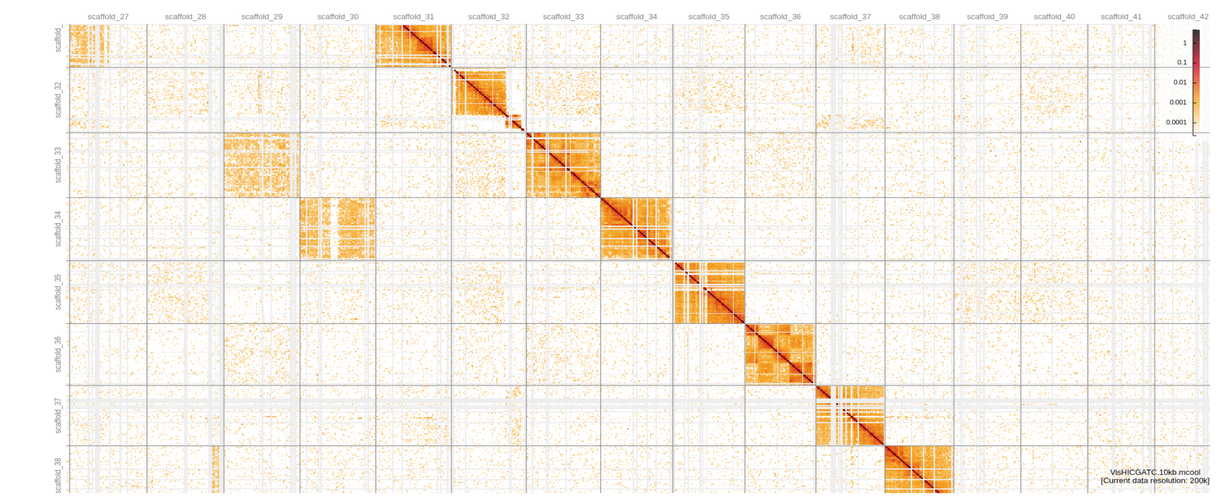

Figure 4. Visualization of Scaffolding with a Probable Mis-join (scaffold 32). Within scaffold 32, there is not a smooth transition in interaction frequency. The beginning of scaffold 32 has almost no interaction to the end of scaffold 32. This represents a probable mis-join of contigs. The algorithm used to scaffold made an error here. Mis-joins are incredibly detrimental to further scaffolding and downstream steps like gene annotation. Mis-joins can be avoided by visual inspection, checks of read coverage, and more than likely coding an algorithm to verify results.

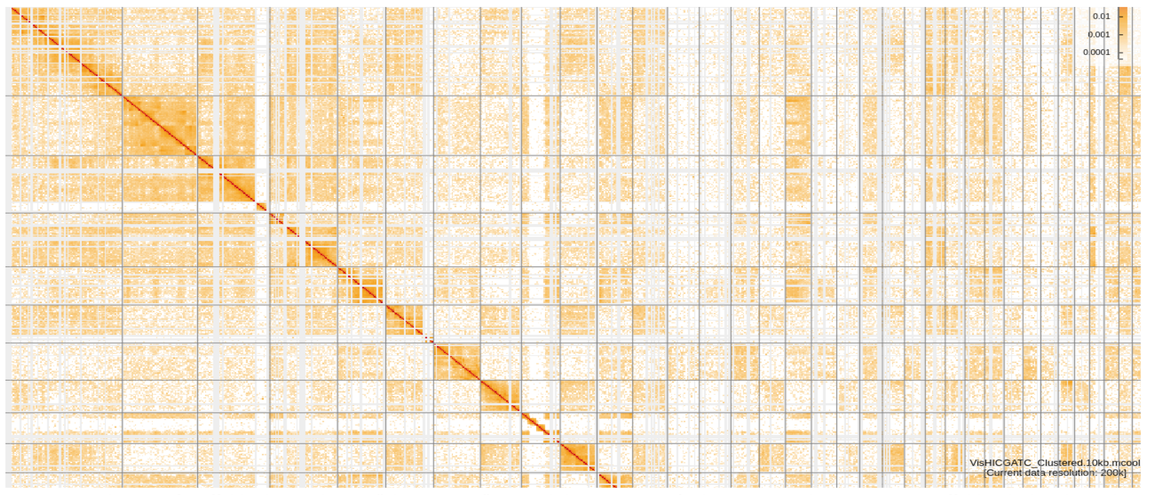

Figure 5. Visualization of Scaffolds Clustered into a "Chromosome" Spectral clustering n = 20). The interaction between some scaffolds is higher than others, yet they have not been joined together. The algorithm used to scaffold (SALSA v 2.0 [7]) only checks for interactions a set distance from the end of contigs/scaffolds. As a result, a large number of connections between large pieces are ignored. A basic way to make use of this information is to create an adjacency matrix where entries are the number of interactions between scaffolds. We can then perform clustering on the adjacency matrix with the number of expected chromosomes. One resulting "chromosome" is shown in figure 5. The interaction frequency between pieces is much higher than the "random" pieces shown in figure 3 and figure 4.

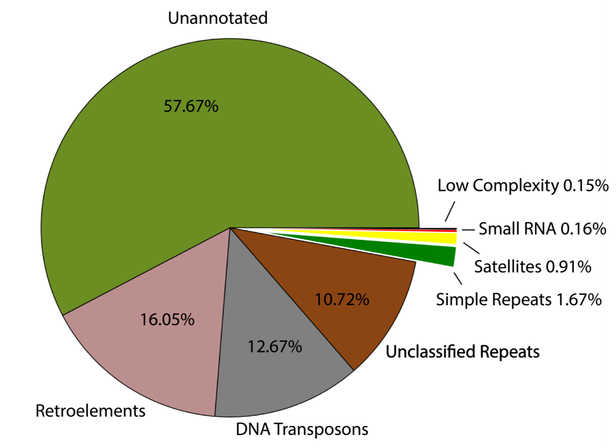

Figure 6. Annotation of Genome Repeats using Repeat Masker[8]. After Genome Assembly the next step is to annotate the resulting DNA sequence. What is contained within this new genome? How does it compare to other species? Where are the genes? The first step to answering these questions is "masking" repetitive sequences. Figure 6 is a pie chart showing how much of the assembled genome was labeled a type of repetitive sequence.

Discussion and Future Work

The results of the genome assembly are promising, but there are a few areas that should be improved. First, based on visual evidence we know there are mis-joins in our assembly. These are damaging for any downstream applications and must eventually be removed. Second, the assembly contains spurious gene duplications. The number of duplicated genes in the BUSCO score is 7.0%; its possible these genes are actually duplicated, but odds are high heterozygosity has created two gene copies one from each parent instead of collapsing them into one gene. This is very damaging to downstream applications like scaffolding itself. We have removed as many of these duplications as possible using open-source software. Removing the remaining duplications will require coding and creating an algorithm ourselves. Third, we know that the scaffolding algorithm used omits information that could be used to group and order large scaffolds. It is probable that we could order our scaffolds into chromosomes by creating/coding our own algorithm.

As it stands, 70% of the *P. feriarum* assembly is in scaffolds of size at least ~1 million bp. There is plenty of information to draw biological conclusions from; the genome is ready for analysis! Currently we are running MAKER v 2.0 [9] to produce gene annotations using previous work on the transcriptome of *P. feriarum* [10]. It remains to be seen if the genome assembly is of high enough quality to produce accurate gene annotations and conduct a robust genome wide association study to uncover the genes driving incipient speciation in P. *feriarum*.

Improving the assembly seems to be a never-ending iterative process. The pipeline used for assembly covers dozens of steps, and at each step the choice of program and parameters is nontrivial. Many papers promise programs which will provide optimal results for certain conditions, but experience has shown the best way to find which program and parameters work best is through testing. Testing out every different option is computationally unfeasible for a large vertebrate genome. The first assembly step with Canu v 1.9 [11] took 15 Tb of space, 100s of Gb of ram, and two and a half months on a 64 core computer. Only educated guesses can be made about the best way to improve the assembly which programs, parameters, type/amount of sequencing, or even what stage of the assembly should be improved.

Its probable the next chapter of my dissertation will involve improving this assembly. Currently Dr. Lemmon and I are working on a project which will involve using sequence capture to vastly improve the on-target information provided by HiC sequencing [12]. This will require writing an algorithm which will scaffold this new data. Even without new sequencing, I currently believe we could improve the assembly by correctly clustering scaffolds into chromosomes and then writing an algorithm to order the scaffolds.

References

[1] Lemmon, E. M. (2009). Diversification of conspecific signals in sympatry: geographic overlap drives multidimensional reproductive character displacement in frogs. Evolution, 63(5), 1155-1170.
[2] Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., ... & Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature, 592(7856), 737-746.
[3] Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., ... & Hedges, S. B. (2022). TimeTree 5: An expanded resource for species divergence times. Molecular Biology and Evolution, 39(8), msac174.
[4] Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210-3212.
[5] Edwards, R. J., Tuipulotu, D. E., Amos, T. G., O'Meally, D., Richardson, M. F., Russell, T. L., ... & White, P. A. (2018). Draft genome assembly of the invasive cane toad, Rhinella marina. Gigascience, 7(9), giy095.
[6] Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., ... & Gehlenborg, N. (2018). HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome biology, 19(1), 1-12.
[7] Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C. S. (2017). Scaffolding of long read assemblies using long range contact information. BMC genomics, 18(1), 1-11.
[8] Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0.
2013-2015 <http://www.repeatmasker.org>
[9] Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics, 12(1), 1-14.
[10] Ospina, O. E., Lemmon, A. R., Dye, M., Zdyrski, C., Holland, S., Stribling, D., ... & Lemmon, E. M. (2021). Neurogenomic divergence during speciation by reinforcement of mating behaviors in chorus frogs (Pseudacris). BMC genomics, 22, 1-23.
[11] 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.
[12] Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B. M., ... & Fraser, P. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome research, 25(4), 582-597.
[13] Gregory, T.R. (CURRENT YEAR). Animal Genome Size Database. http://www.genomesize.com.
[14] Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature communications, 11(1), 1432.
[15] 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.