Extension request for 2021 July-Dec, XSEDE Award MCB100147, Genome Informatics for Animals and Plants
Principal Investigator: D. G. Gilbert

Summary of work, 2020 to 2021 July

1. The major product, EvidentialGene, with 10 years of research effort is supported by XSEDE allocations.
Public URL: http://eugenes.org/EvidentialGene/

This is genome informatics software, including a pipeline for cyberinfrastructure to accurately reconstruct animal and plant gene sets. This is in wide use now among bioscientists, with about 118 citing publications in the 2020-2021 period from Google Scholar search. Compare this to CyVerse, a different but known, larger bioscience cyberinfrastructure project, with 442 citing publications at Google Scholar.

2. EvidentialGene Gnodes/Genome Depth Estimator, in progress since June 2020.
Public URL: http://eugenes.org/EvidentialGene/other/gnodes/

Gnodes is a Genome Depth Estimator for animal and plant genomes, also a genome size estimator. It calculates genome sizes based on DNA coverage of assemblies, using unique, conserved gene spans for its standard depth. Results of this tool match the independent measures from flow cytometry of genome size quite well in tests with plants and animals. Tests on a range of model and non-model animal and plant genome assemblies give reliable and accurate results, in contrast to unreliable K-mer histogram methods.

Gnodes produces detailed assessments of genome assemblies to aid improvments, including measures of major components of genes, transposons, duplicated and unique genome regions, and an assembly accuracy assessment of all genes in the genome.

Genome reconstruction is a Goldilocks problem: answers are often too hot, or too cold; the just-right solution takes effort to discriminate among these outcomes. Gnodes provides a measuring stick for too hot and too cold genome assemblies. When used to compare several assemblies of one organism, it spots over- and under-assembled portions, relative to its unique gene DNA depth measure. It can be used to estimate genome size from only gene coding sequences mapped with genomic DNA, and these tests show it is reliable for that. Gnodes is now a component of the EvidentialGene package.

Gnodes resolves a few discrepancies, such as Daphnia water flea genome assemblies that are only 1/2 size of flow cytometry measured size, and the well-known 40 megabase discrepancy in model plant Arabidopsis (Bennett et al 2003). Applied to the model fruit fly genome assembly, it suggests a 10% deficiency in this assembly that includes missed duplicates in several gene families.

Extensive gene coding sequence duplication is a likely reason that assemblies of Daphnia genomes have faltered at half-size. Half of Daphnia genomic DNA aligns to genes coding sequence, much more than the 10-20% of measured insects and vertebrates, or 25% in measured plants.

3. Genome re-assembly of Daphnia water flea species, in progress since June 2020.
   Public URL: http://wfleabase.org/genome/Daphnia_species_genomes/ and
   http://eugenes.org/EvidentialGene/daphnia/Daphnia_species_genomes/

As the impetus to the Gnodes genome measuring tool, recent Daphnia genome projects have produced poorer quality results than the initial works of a decade back. This should not be! Science is supposed to advance, not retreat, when public sharing of experimental and descriptive results can be buit upon. But Daphnia genomes were found originally to be much different, with extensive gene duplications, from known genomes of insect relatives.

In this effort to resolve the failures of new genomics data and software to recover all of the Daphnia genomes, I examined existing software, found much of it is flawed w/ respect to measuring duplicated regions of genomes. Then I worked out a more accurate approach, using experimental DNA sequences to measure gene and chromosome coverage, and depth. In the course of this I have also re-assembled genomic DNA from several experiments, for 3 species of Daphnia, using several assembly software methods. This consumes a large amount of cyber resources, relative to usual genome projects, as testing several methods concurrently, and sourcing several DNA experiments for this, is the underpining to accurate measurement for a Goldilocks problem such as this. One must look at several too-hot and too-cold results to approach the just-right answer.

This Daphnia genomics work has produced roughly 20 TB of intermediate data sets, both at XSEDE SDSC and at IU NCGAS cyber resources. Some of these data are still required to complete this project, as recalculations on these data with software updates is required to clean out problems. I've reduced these data to roughly 9 TB for the on-going needs to finish this Daphnia and Gnodes genomics work.

An outcome of this nearly complete work are several new Daphnia species genome assemblies that more accurately recover all the species gene and other duplications, and match external evidence of flow cytometry genome size estimates.