Annual Report for Period: 06/2013-10/2014
Principal Investigator: Gilbert, Donald G.
XSEDE Award: MCB10014, Genome Informatics for animals and plants
related NSF Award: 0640462, Shared genome database informatics and cyberinfrastructure

Progress Summary

The research focus is developing for and using shared cyberinfrastructure for assembly, annotation and comparative analysis of new eukaroyte genomes. This sixth year has produced substantial results with *EvidentialGene*: 1. A complete gene set and genome annotation for the environmental, population genomic killifish *Fundulus heteroclitus;* 2. A complete, finished gene set for the environmental genomic model water flea *Daphnia magna;* 3. Assistance to the loblolly pine genome project [1] with transcript assembly methods and computations; 4. Draft gene assemblies for honey bee (*Apis mel*.) and deer tick (*Ixodes scap*.) of significance for agricultural and health improvements. In conjunction with gene set production, new algorithms for merging methods of transcriptome and genome construction have been conceived, developed and implemented in the EvidentialGene code set for these projects. This work and its discoveries rely on substantial computational and data storage resources of the NSF-funded XSEDE and National Center for Genome Analysis Support. This includes cpu and datastore during methods test and refinement for repeated alignment and assembly of a millions of gene proteins and billions of RNA-sequence gene fragments.

Gene-ome construction with mRNA-seq

For the last 2 decades, complete gene sets have been predicted from gene signal statistics in genomic DNA. The advent of high quality, high volume transcript sequencing provides data suited to constructing genes without statistical guesses, from biological gene products. Informatics methods now have caught up to this data, to construct biologically accurate, measurably complete organism gene sets, or transcriptomes. Recently improved mRNA assembly methods of the EvidentialGene project (http://arthropods.eugenes.org/EvidentialGene/) are relatively simple, rapid and biologically valid; simpler, quicker and better than genome-based predictions. RNA assembly combined with genome-based modelling gives more complete answers.

Table 1 provides gene set completeness scores from orthology of plant species (scored with bi-directional best blastp and OrthoMCL gene family clustering), similarly Table 2 lists gene set completeness for fish species. Gene set completeness is scored by average protein size deviations from family, alignment bitscore, and family membership count. Bitscores and family counts are influenced by taxonomic distances as well as gene set qualities. The Cacao-Evigene, Pine-Evigene and Killifish sets are produced with this project's software, and have highest orthology completeness scores. A recent independent comparison of mRNA assembly methods found also that EvidentialGene produces highest quality transcriptome gene set for *Nicotiana* plant [2].

Table 1. Gene set completeness of plants.

Gene set	Size	Bits	Family	Tiny
Cacao- Evgn	13	544	15161	0.7%
Cacao-Cr	9	527	14897	1.5%
Pine- Evgn	56	434	11344	1.3%
Pine-Maker	-106	342	8761	30%
Poplar	0	512	15130	1.6%
Arabidopsis	-2	428	13345	1.0%
Soybean	-15	477	14559	2.7%
Amborella	-6	355	11766	4.1%

Table 2. Gene set completeness of fish.

Gene set	Size	Bits	Family	Tiny
Killifish *	46	585	17272	1.1%
Maylandia	37	595	16469	1.1%
Tilapia	12	568	14905	1.9%
Platyfish	-2	548	15305	4.7%
Zebrafish	-14	527	15190	4.8%
Spot. Gar	18	468	---	3.4%

Table 1,2 legend: Size = protein size difference from family median; Bits = bitscore from blastp for all families with 3+ plants; Families= number of orthologous gene families found; Tiny = count of tiny protein size outliers (-3sd below family median) . Ortho-completeness of gene sets computed with OrthoMCL, blastp, and gene family metrics. Table 1 geneset sources: Cacao-Evgn, Pine-Evgn = this project; Cacao-Cr = Cacao-criollo, Cirad collab.; Pine-Maker = Loblolly pine project genome-genes Maker product ; Arabidosis = model organism; Poplar, Soybean, Amborella = other plant genome projects. Table 2 gene set sources: Killifish * = this project; Maylandia = NCBI; Tilapia, Spotted Gar = Ensembl; Platyfish = Independent + Ensembl, Zebrafish = model organism project.

Examples from this project include

- *Daphnia magna* genome-assembly is missing 40% expected size to gaps and duplicate assembly problems. D. magna mRNA-assembled genes, in progress, appear to be the most orthology-complete of crustaceans.

- *Ixodes* deer tick genome genes are very fragmented, due to abundant transposons, whereas *Ixodes* and Zebra tick mRNA gene assemblies are more complete.

- Pine beetle genome-gene predictions from Maker [6] are much less complete than mRNA-assemblies of *Pogonus* beetle in homology to reference insect genes, despite use of those reference genes in Maker predictions.

- Loblolly Pine tree mRNA-assembled genes are much more orthology-complete than genome-gene predictions from Maker for this tree [1] (Table 1).

- Killifish genome assembly has scrambled many kilobase segments in gene regions: 25% of long genes have poor genome assembly, approx. 5000 fish genes map poorly or not at all to 2nd assembly, below the 1st assembly. This is detected with mRNA-assembled genes, validated with other fish species homology, to improve genome assembly for population studies.

It would not be possible to obtain full species gene orthology using only genome assembly mapped methods for several of these projects, despite their use of current high quality short-read sequencing and assembly methods. mRNA gene assembly combined with genome-based modeling can give more complete answers, and gene-centric projects will benefit by allocating more effort to transcript sequencing.

Solutions to gene set inaccuracy

Current and future genome projects are in need of improved methods for accurate gene set discovery and annotation, in ways that can be used with limited project informatics resources but relying on best practices supported through informatics centers and development projects such as this one. Better use of gene evidence, expression, protein homology, structure information (introns, sizes, protein completeness) applied to many alternative computed gene models can

resolve, per locus, the most accurate gene representation supported by evidence .. similar to how expert curators resolve gene evidence.

Rule-based classification of millions of gene models scored for evidence compliance, ranked by score/rule, exceptions evaluated by experts and re-rankings is found to produce highly accurate, complete gene sets. Evidence includes genome mapped structure, coding sequence metrics, protein homology with reference species genes, expression measures including EST/RNA sequence coverage, intron signal (genome map) agreement, artifact/aberration detection (gene-joins and fragments, UTR/CDS ratio).

A critical component of this approach to perfecting gene sets is the ability to select biologically valid models from a large superset that includes fragments, fusions and fabrications of the gene assembly and prediction components. EvidentialGene uses extensive evidence annotation and maximization of evidence support, minimization of errors. It is closest in approach to NCBI Eukaryote Genome Annotation Pipeline [3]. It differs from peer methods of Glean [4], EvidenceModeler [5], MAKER [6] and others for its deterministic evidence scoring, detailed per gene annotations, and single-best model/locus approach.

Problems to Resolve.

Merging of genome-mapped and genome-free scoring methods for gene evidence is a planned step. Reference-free and genome-based genes have errors unique to both, gene evidence and qualities in common and different. A general approach to this has been developed with the Killifish project, but needs integration for further general uses.

Recently evolved genes are the hardest to accurately assembly or model, with duplication repeat problems, weak or no orthology models, variable expression (i.e. lower than orthology genes for standard environs, but high for special environs). Such are subject to mis-assembly, gene-joins, poor genome mapping, gapping, and other gene construction problems.

Environmentally responsive genes are more often recently evolved, species-specific or new paralogs, with lesser known functions inferred from orthology. This finding turned up for *Daphnia pulex* [7], and appears with *Daphnia magna* and *Killifish*, as well as other gene x environ studies. Constructing accurate *Daphnia* gene sets are essential to resolving complex associations in environmentally responsive genes in this model of environ-genomics.

Broader Impacts from this Cyberinfrastructure Support

This project provides new developments in genome database informatics and cyber-infrastructure to several bioscience communities for discovering genomes and gene-environment interactions. New methods that determine high quality gene information have been developed and published for others to use. Resulting gene structures and descriptions of several animals and plants are placed into public databases that will continue for many years as a basis of new discoveries.

Societal impacts of this work today and in future include: environmental health impacts of man-made toxins and activities are being discovered now via environmental genetic responses in the aquatic sentinel water flea, made possible with outcomes of this project; improvements to sustainable agricultural through understanding plant disease resistances and other traits that aid farmers around the world, using outcomes of this project for genomics of chocolate trees, pine

trees and other plants; better understanding of genetic interactions of human disease vectors and agriculturally beneficial and pest bugs such as aphids, ticks, bees and wasps.

Other discipline impacts. Biomedical disciplines are aided by improved, more accurate gene and genomic information for human disease vectors and vertebrate animals. Environmental health and toxicology disciplines are aided by the improvements in knowledge of the environmental-genomics model organism of the water flea *Daphnia*. Sustainable agricultural disciplines are aided by genome knowledge improvements for breeding disease resistance and other traits in valuable plants.

Information infrastructure impacts. The outcomes of this project provide long-lasting value in their high level of biological accuracy, measured with objective standards above several other widely employed genome annotation methodologies. This project's genomic information will continue to support new discoveries for years to come.

References

1. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martínez-García PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu LS, Gilbert D, Marçais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JF, Lorenz WW, Whetten RW, Sederoff R, Wheeler N, McGuire PE, Main D, Loopstra CA, Mockaitis K, Dejong PJ, Yorke JA, Salzberg SL, Langley CH, 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 15(3):R59., PMID: 24647006.

2. Nakasugi K, Crowhurst R, Bally J, Waterhouse P (2014) Combining Transcriptome Assemblies from Multiple De Novo Assemblers in the Allo-Tetraploid Plant Nicotiana benthamiana. PLoS ONE 9(3): e91776. doi:10.1371/journal.pone.0091776

3. Thibaud-Nissen, Françoise, Alexander Souvorov, Terence Murphy, Michael DiCuccio, and Paul Kitts. 2013. Eukaryotic Genome Annotation Pipeline. The NCBI Handbook [Internet]. 2nd edition http://www.ncbi.nlm.nih.gov/books/NBK169439/

4. Mackey AJ, Pereira FCN, and Roos DS, 2006. GLEAN: improved eukaryotic gene prediction by statistical consensus of gene evidence. Authors' MS Draft.; and Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM: Creating a honey bee consensus gene set. Genome Biol 2007, 8:R13.

5. Brian J Haas, Steven L Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E Allen, Joshua Orvis, Owen White, C Robin Buell and Jennifer R Wortman, 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology 2008, 9:R7 doi:10.1186/gb-2008-9-1-r7

6. Holt C, Yandell M: MAKER2: an annotation pipeline and genome- database management tool for second-generation genome projects. BMC Bioinformatics 2011, 12:491

7. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Cáceres CE, Carmel L, Casola C, Choi JH, Detter JC, Dong Q, Dusheyko S, Eads BD, Fröhlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsveld J, Kriventseva EV, Kültz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzky C, Smith Z, Souvorov A, Sung W, Tang Z, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye Y, Shaw JR, Andrews J, Crease TJ, Tang H, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, Boore JL, 2011. The Ecoresponsive genome of Daphnia pulex, Science, 331:555, doi:10.1126/science.1197761; PMID: 21292972;