Annual Report for Period: 06/2012 - 06/2013
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 primary research focus is developing for and using shared cyberinfrastructure for assembly, annotation and comparative analysis of new organism genomes. This fifth year effort has produced substantial results with *EvidentialGene*, introduced in prior reports. In conjunction with this software engineering, it has been used to produce well-annotated gene sets for several animal and plant species. An engineering "discovery" has been made and substantiated: that gene construction from mRNA-seq data is now surpassing genome-based gene predictions in biological accuracy. This has ramifications to many areas of biosciences and related health and agriculture fields that rely on accurate gene information from animals and plants. The software developed in this project is now being used by others to advance their gene discovery for a range of organisms. This work and itsß discoveries have relied on the substantial computational and data storage resources of the NSF-funded shared cyberinfrastructure of Teragrid/XSEDE and National Center for Genome Analysis Support.

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 show here with Crustacean, Insect and Tick examples. These methods are relatively simple, rapid and biologically valid; simpler, quicker and better than genome-based predictions. While not yet in general practice, these are recommended as they yeild large improvements to published mRNA assemblies or genome-based predictions. RNA assembly combined with genome-based modelling gives more complete answers, but gene-centric projects will benefit by allocating more effort to transcript sequencing.

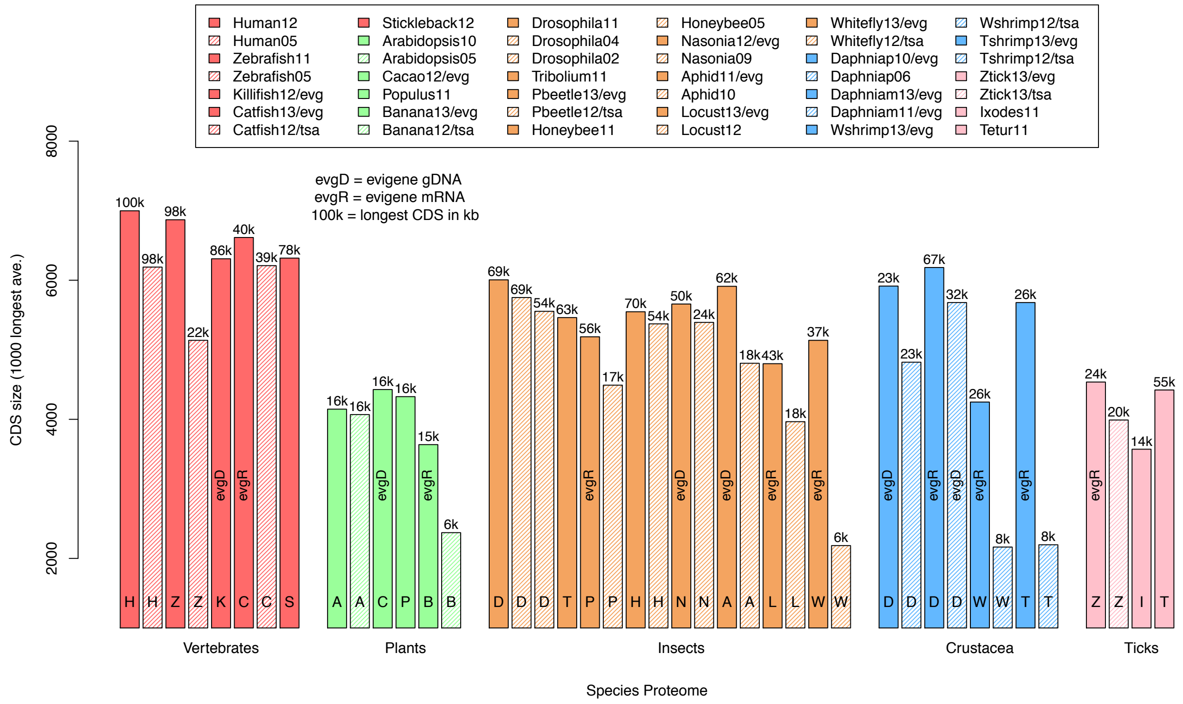

Figure 1. Gene-ome completeness of Animals and Plants.

The Figure 1 graph shows the average size of the longest 1000 proteins in several mature and new animal and plant gene sets. Versions of gene sets from present (2012-13, solid bars) past to 2005 or 2002 (hatched bars) are shown. The gene sets include those developed in the EvidentialGene project, marked as evgD, evgR (genomic or transcriptome only), and related reference or new species gene sets for comparison. This doesn't attempt a full display of species, but a fair set to compare to EvidentialGene sets. The EvidentialGene sets are at or above their peers in this size quality, and size correlates highly with homology to other species genes. Among the EvgR sets (mRNA assembly) are some that are the most complete of their clade (Ticks and Crustacea, with high quality mRNA gene sets for plant and fishes also).

Why mRNA-assembly is better than genome-gene prediction.

* Genome assembly gaps, mis-assemblies are common, and disrupt gene models.
* Transposons are a major problem for gene modelling, and abundant.
* Long introns, common w/ transposons, fragment genome-gene models
* Genome gene predictor training on valid species genes is essential and difficult, no training for mRNA assembly.
* Genome gene prediction remains an educated guess. mRNA-assembly is now engineering with reliably good results given accurate data and methods.
* Reference protein mapping needed for genome-genes, but introduces errors. mRNA-assembly without ref proteins can/does produce stronger homology genes with no confounding of evidence.

* mRNA-assembly is simple, quick, accurate, less cost, more complete relative to genome-gene prediction.

Examples
- *Daphnia magna* genome-assembly is missing 1/2 expected size to gaps, duplicates. Dmag mRNA genes are now are most complete of crustacea.
- *Ixodes* tick genome genes very fragmented with transposons, Zebra tick mRNA genes much more complete
- Pine beetle genome-gene predictions (Maker) are well below *Pogonus* beetle mRNA-genes in homology to reference genes, despite use of ref genes for predictions.
- Killifish genome assembly has scrambled many kilobase segments in gene regions (25% to 50% of long genes have poor genome assembly). This is detected with mRNA-assembled genes, validated with other fish species homology, to improve genome assembly for further population studies.

New or update Gene-ome gene sets built with EvidentialGene during 2012-2013:
Arthropods: Locust, *Locusta migratoria*; Beetle *Pogonus chalceus;* Whitefly *Bemisia tabaci*; Waterflea *Daphnia magna*; White shrimp *Litopenaeus vannamei;* Tiger shrimp *Penaeus monodon;* Zebra tick *Rhipicephalus pulchellus;* Plants: Banana *Musa acuminata*; Chocolate tree *Theobroma cacao*; Pine *Pinus taeda*; Vertebrates: Catfish *Ictalurus punctatus;* Killifish *Fudulus heteroclitus*;

EvidentialGene mRNA assembly pipeline software.

EvidentialGene tr2aacds.pl is the new software outcome, an easy to use pipeline for processing large piles of transcript assemblies into a biologically useful "best" set of mRNA, classified into primary and alternate transcripts. Classification is based primarily on CDS-dna local alignment identity. Transcripts at one locus share exon-sized or larger identities. Perfect fragment CDS are dropped, those with some CDS base differences are kept, with longest CDS as primary transcript. UTR identity is ignored (for now) because many of the mis-assemblies are from joined/mangled genes in UTR region.
Algorithm of tr2aacds:
0. collect input transcripts.tr, produce CDS and AA sequences, work mostly on CDS.
1. perfect redundant removal with **fastanrdb**
2. perfect fragment removal with **cd-hit-est**
3. **blastn**, basic local align high-identity subsequences for alternate tr.
4. classify main/alternate cds, okay & drop subsets by CDS-align, protein metrics.
5. output sequence sets from classifier: okay-main, okay-alts, drops.
See http://eugenes.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html

Other Evigene scripts for mRNA assembly
evigene/scripts/rnaseq/trformat.pl : regularize and unique IDs in transcript.fasta, adding prefixes for parameter sets.
evigene/prot/namegenes.pl : add gene function names from UniProt and Conserved Domains (CDD) with delta-blastp.

evigene/scripts/rnaseq/**asmrna_trimvec.pl** : process NCBI vector screen, and trim end gaps in transcripts.
evigene/scripts/evgmrna2tsa.pl : check mRNA, add annotation, create public IDs and sequence files, write Genbank TSA format for public submission.

Discussion.
EvidentialGenes results are not perfect, but this approach appears to be working. A major remaining need is in tuning for problem cases. Expert inspection combined with evidence rescoring reduces these, but the last 10% require effort similar to the first 90%.

The data explosion from low cost, high throughput sequencers now surpasses abilities of software and shared cyberinfrastructure to fully and properly analyze these data. Transcriptome assembly and gene construction is now engineering rather than art/science; the methods work when properly applied to produce biologically valid genes. However, current software for this, shared hardware for data storage, network transfers, memory requirements, are all stressed by the large size of these data sets. Software in this area is improving on a monthly basis. Cyberinfrastructure resources are available and improving. But analyses of these data sets now require much effort in slicing and reducing data in creative ways to enable software and hardware to process, often with more failed attempts than successful ones, or with partial results that do not fully use the available data.

Biologists are capable of analyzing genomes and gene sets now without a large effort from expert bioinformaticians. The availability of genomics software, in shared-computing environments (e.g XSEDE) is growing, along with easier to use, web-enabled interfaces (e.g. Galaxy). Documentation is well enough developed that biology graduate students can and do learn requisite genome informatics methods, with some script programming for data processing, to produce genome annotations as part of their research. There remain aspects that need improvement in the software and cyberinfrastructure so that these biology students can turn their data into biologically valid models.

References:

1. Zhao *et al*. 2011. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics, 12:S2. doi:10.1186/1471-2105-12-S14-S2
2. Schulz MH, DR Zerbino, M Vingron and E Birney, 2012. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012 Apr 15;28(8):1086-92 PMID:22368243
3. Martin, JA and Zhong Wang, 2011. Next-generation transcriptome assembly. Nature Reviews | Genetics, doi:10.1038/nrg3068
4. Trachana K, Larsson TA, Powell S, Chen W-H, Doerks T, Muller J, and Bork P. 2011. Orthology prediction methods: A quality assessment using curated protein families. Bioessays 33: 769–780. doi:10.1002/bies.201100062