XSEDE Award: MCB100147, Genome Informatics for Animals and Plants
Annual Report for Period: 2017-2018 Principal Investigator: D.G. Gilbert.

Precision genomics is essential in medicine, environmental health, sustainable agriculture, and biological sciences. Yet popular genome informatics methods lag behind the high levels of accuracy and completeness in gene reconstruction that are attainable with today's accurate RNA-seq data. EvidentialGene is a gene set reconstruction pipeline that has a measurably high accuracy and completeness. Recently reconstructed genes for pig and zebrafish model animals demonstrate this accuracy. This pipeline algorithm is simple and robust, compared to gene modeling pipelines, and often outperforms them. It uses big data from gene sequencers, generating bigger sets than other methods, then efficiently reduces those to accurate genes using biological criteria of protein codes and orthology. EvidentialGene is in production use by others, and is installed in cyberinfrastructure around the world, including at XSEDE's JetStream.

Current status of project

Genes of the pig, *Sus scrofa*, and the zebrafish, *Danio rerio*, model bio-medical and agricultural animals, have been reconstructed with EvidentialGene, in conjunction with SRA2Genes software developments of this project. These gene sets improves upon other methods, including finding human orthologs missing from current NCBI and Ensembl reference pig and zebrafish gene sets, additional alternate transcripts, and other improvements.

Methodology for accurate and complete gene set reconstruction from RNA is the newly developed automated SRA2Genes pipeline. SRA2Genes ties together several components that have been developed in past years, for a complete pipeline from raw RNA-seq data to gene transcript assemblies published to GenBank. EvidentialGene methods use several gene modeling and assembly components, annotates their results with evidence, then classifies and reduces this over-assembly to a set of loci that best recovers the gene evidence. Each component method has qualities that others lack, and produces models with better gene evidence recovery. For RNA-only assembly, this paradigm is refined to introduce a coding-sequence classifier (Gilbert 2013), which reduces large over-assembly sets (e.g., 10 million models of 100,000 biological transcripts) efficiently, using only the self-referential evidence of coding sequence metrics (protein length and completeness, UTR excess).

SRA2Genes software+data pipeline produces more accurate and complete gene sets than NCBI RefSeq and EBI Ensembl standards, for animals and plants of biomedical and agricultural importance. This project period has included SRA2Genes development, and uses in producing accurate gene sets for model and non-model organisms. The Pig gene set is published (Gilbert 2018), with gene sequences at GenBank, data description paper at PeerJ, and full data archived at IUScholarWorks. As well, 28 animal and plant gene sets constructed with EvidentialGene from 2010 to 2018 are published at **Scholarworks.iu.edu** public persistent archive, for continued use. Several of these will be submitted to NCBI transcript archive in their required data form in the coming year.

Table 1. Mis-modeled conserved genes in model animals, by three methods, for 2586 highly conserved vertebrate genes (Gilbert 2018, Table 4). Achievable value is near zero, as for Human.

Gene set	Pig	Cow	Mouse	Rat	Fish	Human
Evigene	18	—	—	—	14	—
NCBI	19	22	9	24	25	1
Ensembl	34	58	5	32	79	1

Animal gene set reconstructions by Ensembl and NCBI RefSeq are widely used and considered high quality, though recent published comparisons of these two are uncommon. The NCBI methods now commonly surpass those of Ensembl, as is found in these pig gene sets. Table 1 indicates this for reconstruction of conserved genes in five model animals, all with errors above the human set that are correctable (e.g., for pig, NCBI+Evigene contain all conserved genes). Neither NCBI nor Ensembl produce de-novo assemblies of RNA-seq. These projects however

can and do use assembled transcripts from their public databases. Evigene's de-novo assembled genes can thus improve these other widely used gene sets.

The computational cost is dependent on the size of raw RNA data sets, and desired effort in merging and finishing from many data sets. Intermediate data storage is destined to be preserved indefinitely (2-5 years) for reanalyses as needed, short-term storage is needed for 1-2 weeks during assembly computations, and published gene data sets have long term value like reference books. Computational costs incurred during reconstruction of the pig and fish gene sets in 2018, in XSEDE resource units, including development, revisions and testing, are as follows: **Pig** used 25K SU compute time with the various software engineering, revisions, 2.1 TB of intermediate persistent data storage, 1 TB short term storage per data slice, and 20 GB of published gene data. **Zebrafish** used 25K SU compute time with the various software engineering, revisions, 1.6 TB of intermediate persistent data storage, 1 TB short term storage per data slice, and 20 GB of published gene data. Estimated costs for production uses without engineering development effort, per organism with similar RNA data sets, are 5-10K SU compute time, 0.5 to 1 TB intermediate storage, and 20 GB publication storage. This methodology is near fully automated, and can be used to measurably improve in a cost-effective way, 100s to 1000s of biomedical and agricultural gene sets, extrapolating from Table 1 and related results.

Objectives for project renewal

Last project period (2017-18), objectives of this project were largely met: (a) DONE partly by NCGAS members: implementation of full pipeline in easy-to-use containers for biologists, such as Galaxy and JetStream, (b) DONE partly: non-coding gene classification (e.g. for zebrafish, corn, whitefly), (c) DONE partly: merging of methods for chromosome-free gene assembly and chromosome-based gene modeling (e.g. for daphnia water flea, others), (d) DONE with SRA2Genes: simplified gene data publication to databases at NIH-NCBI and EBI. New gene set reconstructions for model and bio-medically valuable animals, DONE: zebrafish, pig instead of frog and mouse.

Objectives for continuing project period are

(a) Continue gene data publications for organisms with existing EvidentialGene sets, which involves some data-centric computation to update with current gene evidences (sequences to GenBank, data description papers to PeerJ or other, software method description papers).

(b) Locate replacement data work station(s) for the PI, to replace current high-speed networked data stations that will be disconnected by new IU cybersecurity policy. This may include Jetstream based persistent virtual hosts, if needed multi-terabyte data store can be attached, or commercial resources (costly at present for terabyte data and high speed networking).

(c) Finish partly completed software components of EvidentialGene so that others can use as the author does. These include non-coding gene validation, alternate transcript and paralog gene validations, merge of chromosome-free gene assembly and chromosome-based gene modeling methods. This effort will include adding gene reconstructions for model and bio-medically valuable animals as test cases for the software improvements.

(d) Reapply to NIH and NSF for further funding of this work. Prior submission reviews had difficulty assessing significance of objective results of this project, which is a widespread problem in reviewing data-centric projects. Despite current lack of funding, this project's works are increasingly used and understood by biology and computing peers, as providing a useful framework for improvements to accuracy in genomic information.

Acknowledgements

XSEDE, and prior TeraGrid, have provided shared computational resources for a decade of development and genome information production, for this Genome Informatics for Animals and Plants, Award# MCB100147. IUScholarWorks staff, including Richard Higgins, have provided a permanent open-access repository of EvidentialGene animal and plant gene sets. NCBI staff have provided review effort for EvidentialGene data sets submitted to GenBank.

Independent research using products of this project

Adoption and revisions/improvements to EvidentialGene software in 2018:

- Public gene assembly pipeline at National Center for Genome Analysis Support and JetStream, https://scholarworks.iu.edu/dspace/bitstream/handle/2022/21599/Evolution%20Poster%202017b.pdf

- Venturini L, S Caim, G G Kaithakottil, D L Mapleson and D Swarbreck. (2018). Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience, 2018, 0(0):1-13; doi: 10.1093/gigascience/giy093/5057872 -- Mikado is built upon EvidentialGene methods
- Voshall Adam, 2018 May, "Consensus Ensemble Approaches Improve De Novo Transcriptome Assemblies", PhD Thesis. https://digitalcommons.unl.edu/computerscidiss/145 -- ConSembl is built upon EvidentalGene methods

Use of EvidentialGene methods of gene set reconstruction in 2018:
- Sea Urchin genome database, Evigene produced a more accurate gene assembly than Maker, at http://www.echinobase.org/Echinobase/Lv_other_assemblies
- Phytozome genome of *Anacardium occidentale* (Cashew tree), 2018, at https://phytozome.jgi.doe.gov/
- Evigene is used for gene construction in 30 papers from 2018, with digital object identifiers 10.1016/j.heliyon.2018.e00583 (sugarcane), 10.1016/j.dib.2018.03.013i (Atlantic chub mackerel), 10.1038/s41598-018-20262-y (Ants), 10.1186/s12864-018-4451-0 (New Zealand leafroller moths), 10.1093/gigascience/gix137 (clownfish), 10.1093/gigascience/gix125 (Neotropical timber forest tree), 10.1111/nph.14949 (plant arabidopsis), 10.1093/gbe/evy003/4827693 (sea anemone), 10.3389/fmicb.2018.01784 (grapevine pathogen), 10.1038/s41467-018-03384-9, 10.1534/genetics.118.300478 (peanut), 10.1016/j.margen.2018.03.007 (pacific bivalve), 10.1007/s11295-018-1248-y (conifer tree), 10.1007/s13258-018-0697-x (fish), 10.1155/2018/8084032 (raspberry), 10.1021/acschembio.8b00335 (skate fish), 10.1101/313395 (water flea), 10.1093/pcp/pcy058 (fir tree), 10.1007/s10126-018-9836-2 (red crab), 10.1093/gigascience/gix114 (amphibian), 10415/6253 (Eleusine plant), 10.1186/s12864-018-5015-0 (pine tree), 10.1016/j.cub.2018.06.0749 (animal), 10.1038/s41598-018-31148-4 (tiger shrimp), 10.1016/j.cub.2018.07.061 (octopus), 10.1101/269530 (grape plant), 10.1101/461483 (Italian truffle), 10.1016/j.cub.2018.10.035 (insect), 10.1371/journal.pone.0206695 (Japanese cedar tree), 10.1016/j.margen.2018.11.001 (barnacles), 10.1016/j.margen.2018.08.003 (marine coral)

References

Gilbert D. 2013. Gene-omes built from mRNA seq not genome DNA. 7th annual arthropod genomics symposium. Notre Dame. F1000Research (poster), doi: 10.7490/f1000research.1112594.1

Gilbert D. 2016. Accurate & complete gene construction with EvidentialGene. Talk at Galaxy Community Conference 2016, Bloomington IN. F1000Research, 5:1567 (slide set). doi:10.7490/f1000research.1112467.1

Gilbert, DG. 2018. Genes of the Pig, *Sus scrofa*, reconstructed with EvidentialGene. PeerJ accepted 2018-Dec, preprint doi: 10.1101/412130; NCBI GenBank gene sequence entry DQIR00000000, full gene data set at IUScholarWorks with doi: 10.5967/K8DZ06G3