XSEDE Award: MCB100147, Genome Informatics for Animals and Plants
Annual Report for Period: 2019-2020   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 high levels of accuracy and completeness in gene reconstruction that are attainable with today's genome data. The EvidentialGene project develops a reconstruction pipeline with such high accuracy and completeness, using an algorithm that is simple and robust. It uses big data from gene sequencers, then efficiently reduces that to accurate genes with biological criteria. EvidentialGene is in production use around the world, including at XSEDE's JetStream.

Current status of project

Science community use of this project's work highlight needs for (a) gene paralog and alternate transcript resolution; (b) more accurate gene reconstruction for non-model genomes, often with deficient or unvalidated chromosome-assembly-only and gene modeling methods; (c) reliable and inexpensive automated reconstruction of animal and plant gene sets; (d) validation of genome informatics recipe tools, a number of which are popular but have serious flaws. This project is increasingly used in these areas, and increasingly cited, for example relative to CyVerse, a bio-centric XSEDE-collaborative cyber-infrastructure (27% in 2020 vs 22% in 2019, Google Scholar counts of EvidentialGene/CyVerse, 2020-June: 46/171; 2019: 64/293).

Evigene version 4, a major update and software release, has been accomplished. In particular, reconstruction of gene alternate transcripts and paralogs is improved. Duplicated genes (paralogs) and extensive alternate transcripts of genes are of importance, and have been subject to poor reconstruction. Evigene is particularly suited to accurate recovery of these, and is used for studies of often-duplicated venom genes (Modahl *et al*. 2020; Hanf *et al*. 2020), plant disease resistance genes (Rao *et al*. 2019), animal receptor and other duplicated genes (Hearn *et al*. 2020; Gazda *et al*. 2020), and other difficult areas of gene reconstruction. The SRA2Genes pipeline has been updated in all 12 steps, with extensive tests on animal and plant species using XSEDE shared cyberinfrastructure. Important improvement areas are protein ORF computation, alternate/paralog classification, non-coding gene classification, gene alternate exon and strand resolution, foreign contaminant and vector cleaning, data annotation and publication.

A critical review of related gene reconstruction methods, and report on Evigene's accuracy in relation to these, is pre-published (Gilbert 2019b), and has seen wide reading and citations. Among the important findings are that other popular reconstruction methods fail to recover many valid gene alternate transcripts and paralogs, summarized in Figure 1 for human and arabidopsis plant reconstructions.

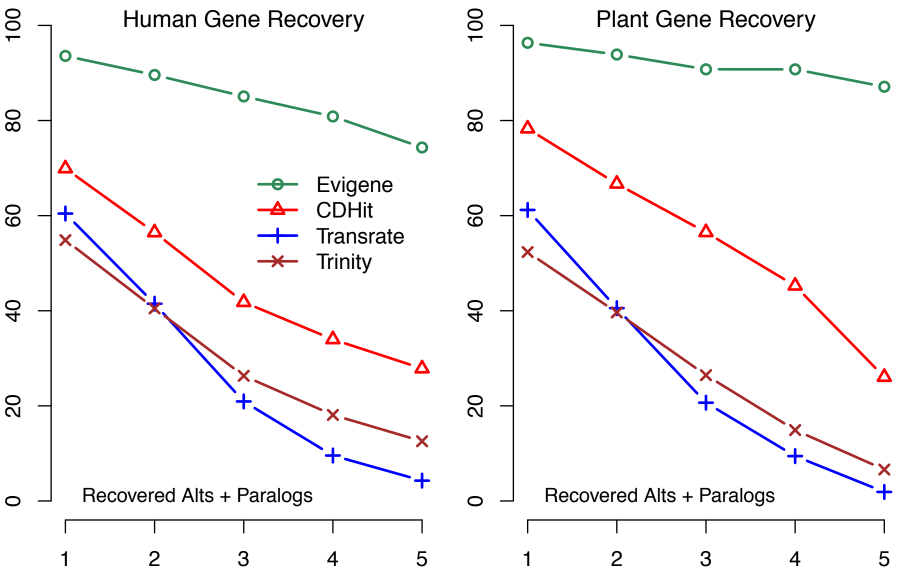

Figure 1. Gene reconstruction method recovery of Human and Plant gene paralogs and alternate transcripts. Number of recovered reference transcripts per gene-family is the X-axis, and the percent recovery is the Y-axis. Methods of reconstruction are Evigene, CD-Hit, TransRate and Trinity (from Gilbert 2019b).

Genes of the deer tick *Ixodes scapularis* are now reconstructed at a higher level of accuracy than available elsewhere, including recent NCBI genome annotations. Especially improved are tick **clade-specific genes** that, as with *Daphnia* waterflea clade-specific genes (Hearn *et al*. 2020; Kvist *et al*. 2020), are among those often overlooked by other gene reconstruction methods. This gene set is published at NCBI TSA, awaiting finishing corrections. Evigene improvements from this effort include addition of a **heterozygosity** option, as the tick RNA samples are all from heterozygous bio-samples, this organism being unsuited to isogenic methods. Other software improvements of note from this tick gene effort are updates to the full SRA2Genes pipeline for publishing gene sets to NCBI TSA, including gene contamination screening.

Evigene outcomes are consistent with the USDA Blueprint for Animal Genome Research 2018–2027 (USDA 2018), as well as agricultural plant genomics. Farmed animal studies of 2019-20 that use Evigene include pig, sheep, fishes, shrimp and clam. Crop plant studies of 2019-20 that use Evigene include rice blight disease, corn, tomato, potato, berry fruit, pear, walnut, faba bean, pine trees, eggplant, fungus phytopathogens.

Evigene use for biomedical genomics including human disease studies is increasing. This includes a recent NIH human disease study of genes of a pathogenic amoeba (Phan *et al*. 2020). Amoeba gene reconstruction with Evigene was a cost-effective method that led to discovery of 59 therapeutic compounds targeting amoeba genes. A USDA data-resource for porcine cytokine genes, relevant to agricultural and biomedical uses, benefits from Evigene genes (Dawson *et al*. 2020 drawing on Gilbert 2019a). Pig immune system and cytokine genes are relevant to human disease, as SARS-COV-2 infection can cause excessive production of cytokines.

This project utilizes these Jetstream objectives (Jetstream NSF Annual Report, 2018-2019): "Self-serve academic cloud services [for] personalized research computing; Hosting of persistent VMs to provide services beyond the command line interface for science gateways and other science services; New modes of sharing computations, data, and reproducibility."

Project public services now on Jetstream virtual host include euGenes.org, with animal and plant genome interactive data services, software and data resources including EvidentialGene; wfleabase.org for Daphnia waterflea genome data services; Bio.net with biology public news/discussion, archive-only now; and others. Basic web statistics of public usage for this past 2019-2020 period, compared with prior use to 2010 are given in Table 1.

Table 1. Project public Internet usage, 2010-2020, on Jetstream and before.

	euGenes		Daphnia		Bionet/IUBio		Biomirror	
Year	**Daily**	**UniqIP**	**Daily**	**UniqIP**	**Daily**	**UniqIP**	**GB/Day**	**UniqIP**
2020	11500	56800	20100	101100	33500	716800	--	--
2016	--	--	--	--	87100	362900	--	--
2014	3100	19200	5100	3000	58100	471800	100	3100
2013	3000	15500	5000	3700	72100	828400	100	6900
2012	2600	12300	6600	4900	94500	885800	100	6600
2011	4200	30700	10200	4900	148200	964400	200	8100
2010	4100	42100	--	--	156700	1662200	200	6800

Year 2020 is 2019.May to 2020.June at Jetstream instance; Year 2014 is partial, 2014.Jan to July; euGenes.org = animal and plant genome data (web database, software, data, docs); Daphnia = wfleabase.org, waterflea genome data; Bionet = Bio.net biology public news/discussion, archive-only on 2019; Biomirror = bio-mirror.net, biology public data share service, turned off 2019; Daily = successful requests/day; UniqIP = distinct hosts served;

Statistics are measured in same way from web use logs, but for Jetstream they include many more robots, mostly malicious, which were blocked at prior IUBio instance. Costs prevented this at move to Jetstream.org. At a guess, 1/4 of 2020 usage are people or benign robots. Scientist-use increases of Daphnia and euGenes have been noted by the author. EvidentialGene project comprises about 15% of euGenes requests. Daphnia requests since 2014 have a large increase, due in part to increasing scientific interest. Another part is due to genome data appearing like foreign phone numbers accessed by "smart" phones.

Bionet news archive usage remains high, in part due to continued interest and value of the many years old bioscience discussions on this venue. E.g. Bionet methods-and-reagents, with many bio-methods recipes and useful discussions, comprises 15% of accesses. Discussion on bionet.immunology of AIDS epidemic research are now popular. New plant genomics research uses methods reported via Bionet in prior 2 decades. Bionet is now read-only archive of1990-2019 discussions, the active Email/Usenet portion was too effortful to resurrect.

A Google Cloud trial as an alternate to Jetstream for project public services is not cost-effective. Data storage costs on commercial clouds are about 100 times that of commodity disk prices. With project needs for persistent, on-line storage increasing into the 20+ TB range, commercial data farms are not suited for long-term, low-funded data science projects.

Objectives for project renewal

EvidentialGene improvements and examples:

A primary objective is further improvements, esp. documentation and simplifications, of Evigene software, focused on the SRA2Genes automated gene reconstruction pipeline (Gilbert 2019a). This now is working well, and its suitability for a broad segment of genomics projects will be better demonstrated. Thus a second primary objective is further examples, with model and biomedical/agricultural important animals and plants. With completion of deer tick gene set, and improvements in paralog reconstruction, several other Evigene sets will be updated. This includes the model zebrafish, that was worked on during the 2019 period, but became stuck on extensive paralog resolution problems. Also to be updated are a set of insects, including disease vector mosquitos *Aedes* and *Anopheles*, and a set of plants, including corn, which has extensive gene duplications from a whole genome duplication events.

Relevance of this project to a range of genome studies is evident. Many current genomics recipes are still inaccurate, and it is common practice for individual labs to put together their own genome informatics pipeline, based more on literature reports than on experience in genome informatics. For example, a recent tree swallow study (Bentz *et al*. 2019) used Evigene but combined it with other methods that this author finds flawed (Gilbert 2019b). Thirty years of this author's experience in this field are funneled into this project, including extensive objective measures that find errors and limitations of methods not disclosed in literature reports. Though a long-term, effortful project, it is bearing long-term results. The Evigene approach of genes-first genomics offers an alternate validation for genome study errors.

Data publication and data workstation updates via Jetstream:

Data-intensive work in this project needs a replacement data workstation connected to high-speed Internet, as a scientist's workbook for less cpu-intensive, but data-rich, analyses not suited to shared compute clusters. A new virtual host is needed for this, to replace those lost in 2019. Public genome data web services are interactive, with term searches, genome map and data report displays. The archiving and reactivation mechanism of Jetstream virtual hosts is very suited to preserving such active science data services. Investigation of such for euGenes, wFleabase and Bionet of this project will be part of this coming project period.

Acknowledgements

XSEDE, and prior TeraGrid, have provided shared computational resources for over 15 years of development and genome information production, for this Genome Informatics for Animals and Plants, Award# MCB100147.

References

Bentz AB, GWC Thomas, DB Rusch, KA Rosvall (2019). Tissue-specific expression profiles and positive selection analysis in the tree swallow (*Tachycineta bicolor*) using a de novo transcriptome assembly. Scientific Reports, 9:15849; doi: 10.1038/s41598-019-52312-4

Dawson HD, Y Sang, JK. Lunney (2020). Porcine cytokines, chemokines and growth factors: 2019 update. Res. Veterinary Science, ISSN 0034-5288; doi: 10.1016/j.rvsc.2020.04.022

Gilbert, DG. (2019a). Genes of the Pig, *Sus scrofa*, reconstructed with EvidentialGene. PeerJ 7:e6374; doi: 10.7717/peerj.6374 ; NCBI GenBank genes entry DQIR00000000, full data at IUScholarWorks doi: 10.5967/K8DZ06G3

Gilbert, DG. (2019b). Longest protein, longest transcript or most expression, for accurate gene reconstruction of transcriptomes? bioRxiv 829184; doi: 10.1101/829184

Hanf, Z. R., & Chavez, A. S. (2020). A comprehensive multi-omic approach reveals a relatively simple venom in a diet generalist, the northern short-tailed shrew, *Blarina brevicauda*. Genome Bio. and Evo. doi: 10.1093/gbe/evaa115

Hearn J, J Clark, PJ. Wilson, and TJ. Little (2020). *Daphnia magna* modifies its gene expression extensively in response to caloric restriction revealing a novel effect on haemoglobin isoform preference. bioRxiv; doi: 10.1101/2020.05.24.113381

Kvist, J., C.G. Athanasio, M.E. Pfrender, J.B. Brown, J.K. Colbourne & L. Mirbahai (2020). A comprehensive epigenomic analysis of phenotypically distinguishable, genetically identical female and male *Daphnia pulex*. BMC Genomics 21, 17; doi:10.1186/s12864-019-6415-5

Modahl C.M., Durban J., Mackessy S.P. (2020). Exploring toxin evolution: venom protein transcript sequencing and transcriptome-guided high-throughput proteomics. In: Priel A. (ed) Snake and Spider Toxins. Methods in Molecular Biology, vol 2068. Humana, New York, NY; doi: 10.1007/978-1-4939-9845-6_6

Rao T B, R Chopperla, R Methre, E. Punniakotti, V. Venkatesh, B. Sailaja, M. R Reddy, A Yugander, G. S. Laha, M. S Madhav, R. M. Sundaram, D. Ladhalakshmi, S. M. Balachandran, S K. Mangrauthia (2019). Pectin induced transcriptome of a *Rhizoctonia solani* strain causing sheath blight disease in rice reveals insights on key genes and RNAi machinery for development of pathogen derived resistance. Plant Molecular Biology; doi: 10.1007/s11103-019-00843-9

Phan, I Q, C A Rice, J Craig, R E Noorai, J McDonald, S Subramanian, L Tillery, L K Barrett, V Shankar, J C Morris, W C Van Voorhis, D E Kyle, P J Myler (2020). The transcriptome of *Balamuthia mandrillaris* [pathogenic amoeba] trophozoites for structure-based drug design. bioRxiv doi:10.1101/2020.06.29.178905; NIH dataset, doi: 10.35092/yhjc.12478733.v1

USDA Blueprint for Animal Genome Research 2018–2027; doi: 10.3389/fgene.2019.00327

Independent research using products of this project

DOI of 43 Reports using EvidentialGene software in 2019:
10.1101/520650, Sweetpotato plant; 10.3389/fpls.2019.00101, Arabidopsis plant; 10.1038/s41598-018-37701-5, Lupin plant; 10.1093/aob/mcz013, Orchid plant; 10.1007/s11103-019-00843-9, Rice blight fungus; 10.1038/s41598-019-39860-5, Green alga; 10.1186/s12864-019-5578-4, Daphnia waterflea; 10.1016/j.dib.2019.103751, Teleost fish; 10.1101/383877, Drosophila fruitfly immune genes; 10.1093/aobpla/plz019, Melastoma plant; 10.1534/g3.119.400214, Eleusine plant; 10.1101/612085, Tomato plant; 10.30473/CB.2019.42829.1754, Coriandrum sativum plant; 10.3389/fpls.2019.00654, Hypericum perforatum plant; 10.1016/j.envres.2019.05.038, Woodlice insect; 10.1371/journal.pone.0210358, Sea anemone invertebrate; 10.1093/molbev/msz115, Frog amphibian; 10.1111/tpj.14276., Mixotrophic plants; 10.1016/j.margen.2019.05.007, Red sea urchins invertebrate; 10.1093/gigascience/giz138, Pear plant; 10.1038/s41598-019-46492-2, Octopus invertebrate; 10.1111/imb.12599, Insect genomes; 10.1038/s41598-019-47985-w, Eggplant plant; 10.1126/science.aav9314, Hydra invertebrate; 10.1534/g3.119.400357, Pine tree plant; 10.22092/ijmapr.2019.125294.2510, Citrullus medical plant; 10.1093/eep/dvz016, Daphnia waterflea; 10.1186/s12864-019-6024-3, Lake amphipods; 10.1101/766444, Maize plant; 10.1101/775841, Trifolium plant; 10.21203/rs.2.14504/v1, Trichoplax; 10.1101/845818, Potato plant; 10.1186/s12915-019-0696-7, Nasonia wasp; 10.1534/g3.119.400529, Tomato plant; 10.1186/s12864-019-6157-4, Shrimp arthropod; 10.1038/s41598-019-52312-4, Tree swallow bird; 10.1186/s12864-019-6177-0, Clam invertebrate; 10.1016/j.margen.2019.04.001, Sea coral; 10.1002/ece3.5646, Nonmodel species; 10.1038/s41467-019-13596-2, Insect; 10.1186/s12864-019-6183-2, Berry fruit plants; 10.1038/s41598-019-55734-2, Fungus phytopathogens; 10.1017/wsc.2019.56, Crabgrass plant;

DOI of 34 Reports using EvidentialGene software in 2020 through June:
10.1007/978-1-4939-9845-6_6, Venom genes of snakes and spiders; 10.1186/s12864-019-6415-5, Daphnia pulex waterflea; 10.1186/s12864-019-6444-0, Pine tree plant; 10.1016/j.cbd.2020.100705, Cuttlefish mollusc; 10.1093/gigascience/giz152, Giant squid invertebrate; 10.1101/2020.01.24.918201, Ragwort plants; 10.1016/j.dib.2020.105166, Crayfish invertebrate; 10.1016/j.margen.2020.100753, sea anemone invertebrate; 10.1101/2020.02.26.966523, Faba bean plant; 10.1002/mrd.23332, Guppy fish; 10.1371/journal.pone.0230266, Four crustaceans; 10.1007/978-3-030-41769-7_6, Daphnia waterflea; 10.1101/2020.03.18.996108, Myrtle rust fungus; 10.1038/s41598-020-62408-x, Newt amphibian; 10.1101/2020.04.21.054320, Cryptomeria japonica; 10.1371/journal.pone.0232005, Persian walnut plant; 10.1016/j.rvsc.2020.04.022, Pig mammal; 10.1093/gbe/evaa075, Beetle insect; 10.1016/j.csbj.2020.05.010, Transcriptome bioinformatics; 10.1016/j.ygeno.2020.01.026, Catfish fish; 10.1093/gbe/evaa101, Sea urchin invertebrate; 10.1101/2020.05.24.113381, Daphnia waterflea; 10.1101/2020.02.12.946319, African lungfish fish; 10.1111/mec.15466, Sira poison frog; 10.1038/s41437-020-0325-9, Butterfly insect; 10.1126/science.aba0803, Birds; 10.1101/2020.06.08.139964, Transcriptome bioinformatics; 10.1101/2020.06.29.178905, pathogenic amoeba; 10.1093/gbe/evaa115, Short-Tailed Shrew mammal; 10.1002/csc2.20014, Turfgrass plants; 10.1101/2020.06.26.173617, Orchid plants; 10.1093/gbe/evaa069, Prairie dog mammal; 10.1093/gbe/evaa124, Snow sheep mammal; 10.1111/mec.15439, Coral invertebrates.
See also http://eugenes.org/EvidentialGene/evigene/docs/evigene-cites.txt