XSEDE Award: MCB100147, Genome Informatics for Animals and Plants
Annual Report for Period: 2021-22 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

Recent public release of EvidentialGene software adds the Gnodes pipeline for measuring accuracy of genome assemblies of animals and plants. Gnodes is a Genome Depth Estimator that calculates genome sizes, and sizes and completeness of their components, based on DNA depth of coverage, using unique, conserved gene spans for a standard depth. Gnodes is very accurate, given accurate DNA samples, matching independent molecular measures of genome size (flow cytometry, Figure 1). This contrasts with popular measurers with larger error ranges.

This software has been developed using XSEDE resources over the past 2 years, and tested with a range of reference genomes, from model insects and plants to fish, chicken, pig and human genomes. In the last 6 months, a major memory use flaw was found for large genomes (human, pig). Refactoring the implementation resolved this, reducing memory use to low levels (under 120 Gb depending on data), with improved speed and analyses. A very large pine tree genome of 22 Gb was tested successfully with Gnodes.

A genome re-assembly of model plant *Arabidopsis* resolves a 20 year old large discrepancy (Bennett et al 2003). DNA samples contain the 157 megabases expected from molecular measures, while public assemblies have 120 Mb. The discrepancy according to Gnodes is missing duplications, including coding genes. As one reason to develop Gnodes, recent *Daphnia* waterflea assemblies have poorer quality than those of a decade earlier. Extensive gene duplication is a likely reason: 50% of *Daphnia* DNA aligns to genes sequence, much more than the 10-20% of measured insects and vertebrates, or 25% in measured plants.

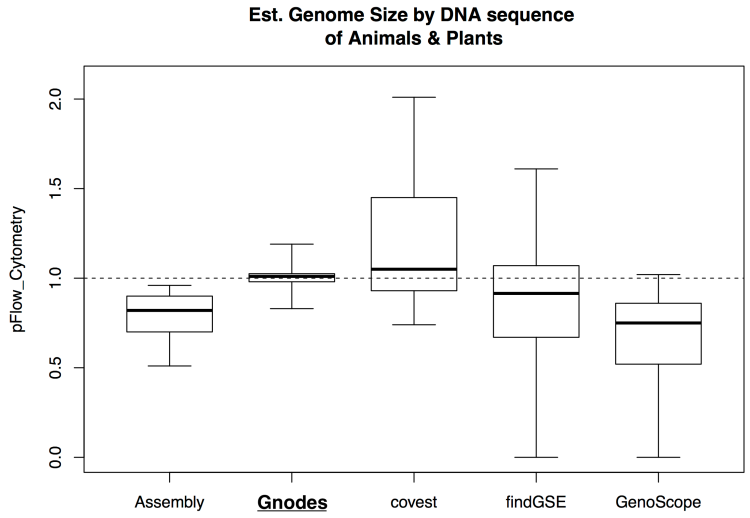

Figure 1. Boxplots (median, range) of estimators' equivalence to flow cytometry (FC) measured genome sizes. Gnodes is very accurate, whereas k-mer histogram methods (GenomeScope, covest, findGSE) are rather inaccurate, with a wide range of estimates. Assembly sizes are typically below FC measured sizes.

Evigene use for biomedical genomics including human disease studies is increasing. This includes a recent human disease study of a pathogenic amoeba (Phan *et al*. 2020) with Evigene led to discovery of therapeutic compounds targeting amoeba genes. A USDA resource for porcine genes, relevant to agricultural and biomedical uses, benefits from Evigene (Dawson *et al*. 2020). Evigene products are consistent with the USDA Blueprint for Animal Genome Research, and agricultural plant genomics. Farmed animal studies that use Evigene include pig, sheep, fishes, and others; crop plant studies include rice disease, corn, tomato, potato, fruits, walnut, trees, and others.

This project includes Jetstream objectives (Jetstream NSF Annual Report, 2018-2019). Project public services now on Jetstream virtual host include euGenes.org, with interactive animal and plant genome data services, software and data including EvidentialGene; wfleabase.org for Daphnia waterflea genomes; Bio.net with biology public news/discussion, archive-only now. Basic web statistics of public usage for Jetstream instance of 2019-2021, compared with prior use are given in Table 1.

Table 1. Project public Internet usage, 2012-2021, on Jetstream and before.

	euGenes		Daphnia		Bionet/IUBio	
Year	**Daily**	**UniqIP**	**Daily**	**UniqIP**	**Daily**	**UniqIP**
2021	10129	24146	17289	36505	13126	109983
2020	11500	56800	20100	101100	33500	716800
2016	--	--	--	--	87100	362900
2014	3100	19200	5100	3000	58100	471800
2013	3000	15500	5000	3700	72100	828400
2012	2600	12300	6600	4900	94500	885800

Year 2021 is 2020.Jul to 2021.July; 2020 is 2019.May to 2020.June at Jetstream instance;
euGenes.org = animal and plant genome web database; Daphnia = wfleabase.org, waterflea genomes;
Bionet = Bio.net biology news/discussion, archive-only; Daily = requests/day; UniqIP = distinct hosts;

Scientist-use increases of Daphnia and euGenes are noted. EvidentialGene project comprises about 15% of euGenes requests. Bionet news archive usage remains high, in part due to continued interest and value of these bioscience discussions. Bionet methods-and-reagents comprises 15% of accesses. New plant genomics research uses methods reported on Bionet. A Google Cloud trial for project services is not cost-effective: storage costs are high, and they appear ill-suited for long-term, low-funded data science projects.

Objectives for project renewal

Gnodes integration with EvidentialGene:

The two year effort on genome measurement with Genome Depth Estimator (Gnodes), has produced a valuable new tool for this field, useable and useful to many genome projects. There remain some detailed improvements to be added: expanding the range and facility of genome data use, encouraging use of this tool among the science community. A remaining objective is full integration of this DNA (chromosomes) tool with the RNA (genes) reconstruction by Evigene. This goal, to more accurately reconstruct duplicated genes, is fairly straight-forward, as Gnodes output contain the basic information needed. Evigene's reconstruction pipeline will incorporate that to produce validated duplicate gene annotations. The value of accurate duplicate gene sets is clear from many biological studies on the role of duplications in rapid evolution and adaptation to environmental, disease and other organismal needs. Current informatics for identifying duplicated genes are often inadequate. Evigene is well suited to accurate recovery of these, and is used in 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), and other difficult areas of gene reconstruction.

EvidentialGene improvements:

A primary objective is further improvements, documentation and simplifications, of Evigene gene reconstruction pipeline (Gilbert 2019a). Its suitability for a broad segment of genomics projects will be demonstrated. A second objective is examples of model and biomedical/agricultural important animals and plants. This project is relevant to a wide range of genome studies [see accompanying list of 260 publications using Evigene, 2014 - 2022]. Many current genomics recipes are still inaccurate, and it is common for groups lacking genome informatics experience to create their pipeline. Though a long-term, effortful project, it is bearing long-term results.

NIH ANVIL Genome Data Cloud to ACCESS:

AnVIL, sponsored by NIH's Genome Research Institute, is a federated cloud platform designed to manage and store genomics and related data, enable population-scale analysis, and facilitate collaboration through the sharing of data, code, and analysis results (Schatz et al 2022). This effort is clearly of interest for interoperability with NSF-sponsored ACCESS endeavors to enable science research in cloud computing. This investigator will endeavor to find a common path between these science cyberinfrastructures, to facilitate this Evigene project, and other genomic sciences outside of commercial cloud services (AWS and Google Cloud) that AnVIL appears tied to. Many science data contributors to the NIH public data sets, and consumers of NSF cyberinfrastructure for analyses of such public data, lack funds for these commercial resources. A solution is needed, which should be of interest to

members of these projects. Micheal Schatz, a genome informatician at Johns Hopkins (with an XSEDE resource, and of NIH AnVIL project) is a senior author of bioinformatic tools and genome assemblies, highlighted in this PI's current and planned work, and may be interested in an AnVIL-ACCESS integration. Robert L. Grossman, also with AnVIL project, has been involved in big-data projects with Globus, TeraGrid, XSEDE.

Data publication and updates via Jetstream:
The archiving and reactivation mechanism of Jetstream virtual hosts is very suited to preserving such active science data services. Investigation of such for EvidentialGene genome data sets, 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

Bennett, MD, IJ Leitch, HJPrice and JS Johnston (2003) Comparisons with Caenorhabditis (100Mb) and Drosophila (175Mb) using flow cytometry show genome size in Arabidopsis to be 157Mb and thus 25% larger than the Arabidopsis genome initiative estimate of 125Mb. Ann. Botany, 91, 547-557 doi:10.1093/aob/mcg057

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. (2022). Genes ruler for genomes, Gnodes, measures assembly accuracy in animals and plants. Draft document, http://eugenes.org/EvidentialGene/other/gnodes/gnodes_doc2draft.pdf

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

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

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

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

Rao T B, R Chopperla, R Methre, et al (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

Schatz MC, AA. Philippakis, E Afgan, E Banks, V J. Carey, R J. Carroll, A Culotti, K Ellrott, J Goecks, R L. Grossman, I M. Hall, K D. Hansen, J Lawson, J T. Leek, A O'Donnell Luria, S Mosher, M Morgan, A Nekrutenko, B D. O'Connor, K Osborn, B Paten, C Patterson, F J. Tan, C O Taylor, J Vessio, L Waldron, T Wang, K Wuichet, and AnVIL Team (2022). Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. Cell Genomics 2; doi: 10.1016/j.xgen.2021.100085

Gilbert, DG. (2022) Genes ruler for genomes, Gnodes, measures assembly accuracy in animals and plants.

Abstract: 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 less reliable K-mer histogram methods. The problem of half-sized assemblies of duplication-rich Daphnia is addressed. A 20-year old Arabidopsis genome discrepancy is resolved in favor of 157Mb as measured with flow-cytometry. Not all genome DNA samples contain a genome, examples and reasons for this are discussed. The T2T completed human genome assembly of 2022 is complete by Gnodes measures, with about 5% uncertainty. With full genome DNA, Gnodes measures within 10%, usually within 5%, of flow cytometry, indicating they are both measuring the same content. Public URL: http://eugenes.org/EvidentialGene/other/gnodes/