Annual Report for Period: 06/2014-04/2016
Principal Investigator: Gilbert, Donald G.
XSEDE Award: MCB100147, Genome Informatics for Animals and Plants

Abstract

Renewal of this **XSEDE Genome Informatics for Animals and Plants** project will facilitate the accurate discovery and reconstruction of animal and plant genes, in current and future genomics collaborations, including those by this author and those independently undertaken.

Precision genomics is essential in medicine, environmental health, sustainable agriculture, and biological research. Yet popular genome informatics methods lag behind the high levels of accuracy and completeness in gene construction that are attainable with current RNA-seq data.

EvidentialGene is a genome informatics pipeline for gene construction that has a measurably high accuracy and completeness rate, for insects, ticks and crustaceans to crop plants and trees, to fishes and other vertebrates. It uses big data from gene sequencers, generating bigger gene sets than alternate methods, then efficiently reduces those into accurate species gene sets using biological criteria of protein codes and orthology. EvidentialGene is in production use at shared cyberinfrastructure centers in USA, Sweden, Australia and elsewhere.

Recent examples with disease vector mosquitoes *Aedes* (yellow fever, Zika virus) and *Anopheles* (malaria), show EvidentialGene surpasses accuracy of published genes from popular genomics methods of MAKER, Trinity and Vectorbase. For fishes, Evigene surpasses those recently published from MAKER, Trinity and NCBI Eukaryote genome annotation pipelines.

Keywords: gene reconstruction, animal and plant genes, genomics for precision medicine, environmental health, sustainable agriculture, big data, bioinformatics pipeline, high performance computing, RNA-seq data, transcriptome assembly

Project URL: http://eugenes.org/EvidentialGene/

Current status of project

Notably this year has published outcomes to public databases and journals of accurate gene sets for the water flea *Daphnia magna* and Atlantic killifish (*Fundulus heteroclitus*) with collaborators. Project products listing (attached) has publications of years 2015, 2016 and in-progress results. The support of NSF DBI-0640462 and XSEDE-MCB100147 are acknowledged. These collaborative projects are now out of funds, but the PI continues work to complete these.

Compute services and genome data storage needs of this project been effectively handled at the XSEDE SDSC Comet and Gordon compute clusters and NSF-project Oasis storage. During 2015 project period, genome computes used 35,136 service units at Comet, and 27,450 SU at Gordon, with 5 to 8 TB persistent data storage, plus transient storage of 1-2 TB per project run. The project methods work on any similar compute clusters with preferably 128+ GB memory and a few Terabytes of parallel disk storage. The production methods will run on any Unix-based laptop or desktop computer for small data sets and test cases.

The code base of this project, as Perl language scripts and methods examples, is published at http://sourceforge.net/projects/evidentialgene/, and project home http://eugenes.org/EvidentialGene/ , and customers use these. The code set contains a broad range of genome informatics methods, as needed for gene set reconstruction and publication to databases.

The genome data base of this project is published via public genomics databases including NCBI, UniProt (see products listing), and via project web-site and SourceForge. However, multi-year storage of intermediate data and analyses is an important part of ongoing collaborations, as final deposition to public databases are often 5-year efforts, with re-analyses and re-annotations required to meet changing requirements. Current project genome data at SDSC is approx. 5 Terabytes of multi-year storage, plus short term use of a few TB. Backup of this long-term valuable genome data to IU Scholarly Storage System is done a few times/year.

These methods process "big data" sets of genomics, and have been developed with high efficiency as a core aspect. This includes the processing of sets of 10s of billions of RNA-seq read data, production of 10s of millions of gene assemblies via memory+disk intensive graph assembly of sequences, large sequence alignment tasks, orthology analyses, gene annotations, etc. The widely used EvidentialGene tr2aacds pipeline has an efficient 4 step redundancy reduction algorithm, stepwise reducing large over-produced gene sets to the most accurate subset. Perl pipeline scripts call to C/C++ compiled methods that are efficient, with PBS and related cluster multiprocessor methods. For instance, the recent reconstruction of 3 mosquito gene sets took roughly 3 days per species of effort for the primary over-assembly of genes by several assemblers, each assembly run using from 6 to 12 hours on 8+ cores with 128 GB of memory at comet.sdsc. Primary RNA-seq data is approx. 200 GB per species per source. Intermediate assembly storage is approx. 400 GB/run set, with 30 assembly run sets for 3 species. Further effort on orthology analyses, annotations, gene alignments, etc. are smaller compute tasks but add weeks of expert effort.

Comparisons with related methods show EvidentialGene's are more or as computationally efficient. E.g., Evigene's clusterized OrthoMCL for orthology analysis is comparably faster than "FastOrthoMCL, and may also surpass a optimized version at XSEDE-TACC. OrthoMCL is a relatively large and important genomics compute, based on all by all NCBI-BLASTp gene alignments of 10s of species (300,000s of genes, more as desired). This all by all BLASTp step is parallelized by data splitting and uses roughly 8-12 hrs on a current 32-core node. Following this, analyses with the original OrthoMCL would take 12+ hours, but required 1-cpu serial mode. Evigene's multi-processor version completes that step in 1-2 hrs with fewer cpu (1 per species). This task is used repeatedly during accurate gene set reconstruction as an essential validation step, so efficiency is very helpful. Some of the applications in these genomics pipelines are cluster MPI-aware (e.g. most gene assemblers). Another standard efficiency method is to embarrassingly parallelize the biology data by splitting to many subsets, computing in parallel, then collate results. This works well for most current genome informatics applications, and avoids effortful, expert-bound software refactoring to MPI methods.

Objectives planned for project renewal

Facilitate further use of EvidentialGene in other projects.

The author continues to seek venues and projects that will evaluate and use these products, among Galaxy bioinformatics cyberinfrastructure project and Generic Model Organism (GMOD) membership, XSEDE centers that support gene and genome assembly, and with projects such as VectorBase (genome database of mosquitoes and ticks), and other organism genome projects with significant informatics expertise. In-progress papers will aid in the recognition of project values. Use of EvidentialGene methods in other projects has been increasing since 2014 (see project products list). This includes installation and use at a few shared cyberinfrastructure centers around the world. Wider use of this accurate method for gene set reconstruction is being sought. This project can and is being used by small groups with limited informatics expertise. It is in need of independent validation by experts in the field to raise awareness of its values.

Improve complex and difficult areas of gene reconstruction.

Aspects of gene construction that are complex and need improvements in Evigene include the paralog/alternate problem (is it one or two loci?). Non-coding RNA gene classification methods are now ignored by Evigene as a separate problem needing distinct evidence methods, but should be added as their importance is well recognized. Addition of population- and clade-level gene set reconstruction and validation, with classification gene heterozygosity effects is underway. Improved measurement of complex and uncommon gene types is needed: trans-spliced and reversed-strand genes, stop-codon read-thru genes, biologically chimeric genes, weakly expressed gene assembly.

Highlight values of gene reconstruction as primary genome information, independent and intertwined with primary chromosome-assembly.

New methods that merge the best of chromosome-based modeling and mRNA assembly are needed to reach the highest level of accuracy. Initial efforts of a hybrid DNA/RNA gene reconstruction in Daphnia and Killifish projects show this approach is effortful, in part due to differing criteria of a "gene" (as a location on a chromosome, versus an assembled sequence), as well as weak genomics community standards for gene-first information (chromosome-first is the standard). Planned work will streamline and automate the publication of genes to databanks, addressing conflicts in gene data representation, as well as needs for screening, annotation and discrimination of biological versus artifactual or contaminant gene contents. Use of independently reconstructed genes for validation and improvement of chromosome assemblies will be encouraged. Steps toward this with restricted orthology gene sets have been made (CEGMA, BUSCO), but chromosome reconstruction with full gene sets as independent evidence can notably improve genome accuracy.

Significance

"*The quality of a gene set is dependent on the quality of the genome assembly*" is commonly accepted dogma in the field of genomics (quote from Ensembl gene builds). New evidence from expressed gene RNA sequencing show no this is longer a valid dogma. Gene assembly from RNA-seq is often more accurate and complete than genome-gene modeling by objective assessments of orthology and other gene-family tests. Accurate assembly of chromosomes is a harder problem, though it is now implicitly considered the primary data of a "genome", and today's chromosome assemblies remain inaccurate, incomplete and difficult to validate.

Gene re-construction done independently of chromosome assembly modeling avoids several error sources: genome mis-assemblies, transposon inserts and long intron difficulties, HMM predictor over-corrections, errors of reference proteins mapped to genome. How many gene-based studies have significant artifacts of quality? In a recent review of gene orthology, "genome annotation emerged as the largest single influencer, affecting up to 30%" of the discrepancies among orthology assessments (Trachana 2011). Accurate knowledge of our human genome is essential for medical uses, yet remains inaccurate with respect to genes and chromosome sections (Goldfeder 2016). Gene function, derived from orthology or by experiment, is sensitive to these imperfections. Differential expression measures are muddled on imperfect genes. Many biology studies that use genome-wide constructed genes hinge on the accuracy of gene sets.

Gene construction methods named EvidentialGene (Gilbert 2013), under development over several years, is at the point of producing highly accurate coding gene sets. This accuracy is measured with objective criteria of orthology completeness, the presence and fullness of protein coding genes that share homology with related species genes. This result is summarized for several animal and plant gene sets at http://eugenes.org/EvidentialGene/. These include accurate genes for a vertebrate fish (*Fundulus het.*), *Cacao* Chocolate bean tree, Loblolly pine tree, and Banana plant, arthropods of honey bee *Apis*, beetle *Tribolium*, deer tick *Ixodes*, the Lyme disease vector, and water flea *Daphnia*, an environmental health model organism.

This is a reproducible, objective assessment. These are in comparison to other good and widely used gene production methods: current NCBI Eukaryotic Genome Annotation Pipeline, ENSEMBL genome annotation pipeline, genome projects using AUGUSTUS, MAKER, GLEAN, EvidenceModeller, and others. Transcript assembly sets produced with EvidentialGene methods are superior to those of several widely used methods, including Trinity, SoapDenovo, Velvet/Oases, TransAbyss, though Evigene uses results of these.

EvidentialGene methods for gene assembly are orthology-blind, that is other species genes are not used in assembling a new gene set. This contrasts with genome gene-prediction methods, which typically use all available ortholog genes to model genes in a new genome assembly. From this, the Evigene sets are also expected to be as complete for novel or species-unique genes as they are for ortholog genes, and artifacts in other species gene sets are not transferred, or created when species have evolved genes. Use of ortholog evidence in predicting new genes leads to errors of both commission (mis-modelled orthologs) and omission (missed non-orthologs).

Recent 2016 results for mosquito gene sets of *Aedes* (vector of Zika virus and yellow fever) and *Anopheles* (vector of malaria), reconstructed with EvidentialGene (**Tables 1,2**), show their higher accuracy in comparison to recently published gene sets for the same species and same source data, using the now popular informatics methods of Trinity (RNA assembly, Grabherr et al 2011) and MAKER (genome gene prediction pipeline, Holt and Yandell 2011). Evigene produced 80% to 90% accurate genes, using rough-draft level effort, versus under 65% for the popular methods. These results for mosquito genes replicate those for prior gene sets of animals and plants noted above. Reasons for improved accuracy with multiple gene assemblers and options are well researched and identified (Zhao et al 2011; Haznedaroglu et al 2012; Schulz et al 2012; Peng et al 2013), thus the popularity of less accurate methods should be of concern in this field.

Genes reconstructed of the killifish *Fundulus* by this project have higher accuracy compared to MAKER and NCBI gene models, as indicated in **Table 4** for conserved ortholog genes. Additionally, non-ortholog genes of killifish contain a large set of putatively recent fish genes, findable in genomes of related fish but unreported as gene models (**Table 5**). These are reported in killifish due to orthology-blind methods of EvidentialGene, but are likely overlooked in others due to a reliance on orthology data by other modeling methods. For the pine tree *Pinus taeda*, Evigene's set aligns 87% to genes of *Arabidopsis* model plant, versus 71% for MAKER modeled genes (**Table 6**, Neale et al. 2014).

Explanation of EvidentialGene's higher accuracy than Trinity or any single gene assembler are: (a) recognize that accurate assembly of 20,000s of genes with different qualities is complex, that no single method, parameter set, nor data subset will properly construct all loci, (b) use a biological measure, the evolutionary conservation of **protein codes,** to identify accurate genes, and (c) recognize that de-novo assembly methods that work well for chromosomes and genes are similar. Gene assembly is simpler in important ways: no or few repetitive transposon spans, versus problematic large repetitive spans in chromosomes, and little assembly is required for genes that average 10 times longer than sequencer machine outputs, versus chromosomes that average over 100,000 times machine output sizes. Gene assembly needs adjusting for varying expression levels, alternate transcripts, and other differences from chromosomes. One critical adjustment is the kmer or read shred size, that has different optimal values for different loci [Zhao et al 2011; Haznedaroglu et al 2012; Schulz et al 2012; Peng et al 2013]. Gene assembly method effects on accuracy are summarize in **Table 3**. Based on these and similar results for other species, there is a flaw in using single, short kmer read shredding as in the Trinity assembly algorithm. Other gene assembly methods do better, but as with genome-gene modeling, a combination of methods provides a more accurate gene set.

The explanation for EvidentialGene's higher accuracy versus MAKER and other chromosome-based gene modeling methods is different, but rests on the mentioned aspects that both chromosome assemblies and gene modeling on those introduce different errors than gene assembly. MAKER and the first EvidentialGene version share the same approach of genome-based gene modeling with mature gene predictors (AUGUSTUS, fgenesh, SNAP, others), scoring many over-produced models per locus for evidence agreement, then selecting best locus-location representatives. An upper limit to accuracy of this approach was discovered in this project, where all models at some loci fail to recover gene evidence found in RNA assembly of those loci.

Of the Evigene improved genes summarized in Tables 1 and 2, compared to genome-located models, a notable subset are split-mapped over chromosome pieces, or partly mapped onto chromosome mis-assemblies. Another notable subset are properly mapped to chromosomes, but genome predictors have missed or fragmented models at those loci. The transposon and repetitive problem that is large in chromosomes contributes to inaccurate gene modeling, as well as to fragmented and mis-assembled chromosomes. Genome-gene predictors have limited ability with complex and unusual gene structures, as located on chromosomes, such as trans-splicing and antisense transcription, but much of this complexity is biologically removed from the transcribed mRNA sequences used for gene assembly.

There is an accuracy limit with gene RNA assembly also, related to low expression levels for some loci. Experimental studies need extra effort to measure expression at all loci, some of which are found in restricted environmental or developmental conditions. This project finds a relatively small 1% - 10% of loci fall below the expressed RNA levels needed for assembly, depending on study methods. A hybrid approach to gene set reconstruction, using both chromosome-located and RNA-assembled methods, is the obvious route to complete and accurate genes. Development of this hybrid approach, in other projects as well as this one, is ongoing, with indications that it is not a simple one: discrepancies in genes constructed each way are common and not readily resolved as sources of artifacts are not obvious. Biological exceptions to standard gene structures often resemble methodological artifacts from gene and genome assembly (e.g. biologically chimeric loci and artifactually joined loci).

Table 1. *Aedes aegypti* mosquito genes, orthology-completeness of 3 versions

Table 1A. ***Aedes_aegypti*** **x Highly Conserved REFERENCE (BUSCO, nr=3055)**

Statistic	Evigene	PubTrinVb3	Vecbase3
found	99.5%	98.6%	98.3%
align	91.3%	86.5%	85.1%
best	42.3%	5.2%	3.0%

equal 52%

Table 1B. ***Aedes_aegypti*** **x** ***Drosophila mel.*** **REFERENCE (nr=11146)**

Statistic	Evigene	PubTrinVb3	Vecbase3
found	99.0%	97.5%	97.1%
align	86.4%	82.4%	81.1%
best	44.0%	9.2%	6.1%

equal 47%

Table 1C. ***Aedes_aegypti*** **x** ***Anopheles gambia*** **REFERENCE (nr=14014)**

Statistic	Evigene	PubTrinVb3	Vecbase3
found	99.1%	97.3%	96.6%
align	94.3%	89.7%	87.2%
best	44.3%	10.4%	8.5%

equal 45%

Table 2. *Anopheles funestus* and *Ano. albimanus* mosquito genes, orthology-completeness of 3 versions

Table 2A. Highly conserved REFERENCE (BUSCO, nr=3041)

	Anopheles_funestus			***Anopheles_albimanus***		
Statistic	**Evigene**	**MAKER**	**Trinity**	**Evigene**	**MAKER**	**Trinity**
found	99.8%	98.9%	98.7%	98.6%	98.7%	97.4%
align	89.0%	85.1%	83.7%	87.2%	84.9%	82.4%
best	33.4%	6.9%	3.1%	39.7%	11.3%	4.6%

equal 60% 49%

Table 2B. ***Drosophila mel.*** **model REFERENCE (nr=11043)**

	Anopheles_funestus			***Anopheles_albimanus***		
Statistic	**Evigene**	**MAKER**	**Trinity**	**Evigene**	**MAKER**	**Trinity**
found	98.8%	97.8%	97.2%	96.5%	97.8%	95.6%
align	83.9%	80.5%	79.3%	81.3%	80.6%	78.5%
best	38.6%	10.8%	4.0%	40.3%	17.4%	5.7%

equal 50% 42%

Table 2C. ***Anopheles gambia*** **REFERENCE (tr total=14870, locus total=12994)**

	Anopheles_funestus			***Anopheles_albimanus***		
Statistic	**Evigene**	**MAKER**	**Trinity**	**Evigene**	**MAKER**	**Trinity**
found	98.9%	98.6%	97.7%	96.4%	98.2%	96.0%
align	96.9%	93.1%	90.2%	91.0%	91.2%	85.0%
best	39.9%	12.4%	3.7%	41.5%	19.5%	6.4%

equal 48% 39%

Footnotes to Tables 1,2

Statistics: found = % reference proteins with significant alignment to test gene sets; align = % alignment of target proteins sets to reference proteins; best = % pairwise count of best alignment of two target gene sets to reference.

Reference gene sets are of the fruitfly model organism (*Drosophila mel.*) curated by FlyBase, and *Anopheles gambia* mosquito curated by VectorBase, where *nr* indicates total gene locus count found among all gene sets (reference set total is larger by a few hundred). The BUSCO (Benchmarking Universal Single-Copy Orthologs, Simao 2015) set is the highly conserved subset of Drosophila genes.

Gene set and source data publications

Aedes Vecbase3 geneset is Aedes-aegypti-Liverpool_PEPTIDES_AaegL3.3 of vectorbase.org
Aedes PubTrinVb3 geneset is of Matthews et al. 2016; doi:10.1186/s12864-015-2239-0 . PubTrinVb3 uses Trinity denovo RNA-assembler, Cufflinks genome RNA-assembler, and PASA EST-gene pipeline.
Anopheles gene sets are of Neafsy et al 2015, doi:10.1126/science.1258522. This published MAKER gene source and RNA-seq data source. This report used Trinity but did not publish these assemblies, I reran Trinity assembly.
Anopheles Evigene set uses only subset of PubTrinVb3 data of Matthews et al. 2016
Aedes Evigene set merges evg2aedes (new data) with evg1aedes, updated 2016.04.05
Aedes_aegypti RNA-seq SRA accessions used for Evigene:
evg1aedes = SRP037535 (male+fem, 10 of 68 SRX read sets) of PubTrinVb3
evg2aedes = SRP047470 (male+fem, 4 of 6 SRX) and SRP046160 (embryo), of Hall et al 2015.
evg1aedes alone is better than PubTrinVb3, by ~33%, using only subset of same PubTrinVb3 data. evg2aedes alone is less complete than evg1, using less effort/data, but contains ~3000 better gene loci (some replace evg1, some are unfound in evg1).

Table 3. Gene Assembler Methods for Accurate Genes

Table 3A. Gene Assemblers Used To Reconstruct Mosquito Genes

Assembler	**Version**	**Code source**
Velvet/Oases	v1.2.10 2013	https://www.ebi.ac.uk/~zerbino/oases/
idba_tran	v.1.1.1 2013	http://www.cs.hku.hk/~alse/idba_tran/
SOAP-Trans	v.1.03 2013	http://soap.genomics.org.cn/SOAPdenovo-Trans.html
TrinityRnaseq	r20140717 (v2.1.1)	https://github.com/trinityrnaseq/trinityrnaseq/wiki

Table 3B. Gene Assembler Methods for Accurate Longest 10K Genes

Anopheles_funestus			*Anopheles_albimanus*		
Count	**Unique**	**Method**	**Count**	**Unique**	**Method**
		assembler			**assembler**
4092 40.9%	2450 24.5%	idba_tran	4622 46.2%	1464 14.6%	idba_tran
2059 20.6%	682 6.8%	soap-trans	2900 29.0%	352 3.5%	soap-trans
1754 17.5%	561 5.6%	trinity	2408 24.1%	305 3.1%	trinity
6122 61.2%	4505 45.1%	velvet/oases	7636 76.4%	4492 44.9%	velvet/oases
		kmer			**kmer**
1495 15.0%	263 2.6%	k05	2219 22.2%	80 0.8%	k05
2785 27.9%	1156 11.6%	k25	3903 39.0%	811 8.1%	k25
4047 40.5%	2077 20.8%	k35	5130 51.3%	1255 12.6%	k35
3053 30.5%	1112 11.1%	k45	4897 49.0%	771 7.7%	k45
2831 28.3%	983 9.8%	k55	4553 45.5%	492 4.9%	k55
2173 21.7%	680 6.8%	k65	4764 47.6%	779 7.8%	k65
1520 15.2%	399 4.0%	k75	4341 43.4%	544 5.4%	k75
1117 11.2%	378 3.8%	k85	3920 39.2%	375 3.8%	k85
719 7.2%	213 2.1%	k95	3460 34.6%	168 1.7%	k95

Table 3C. Gene Assembler Methods for Accurate Highly Conserved Genes (BUSCO)

Anopheles_funestus			*Anopheles_albimanus*		
Count	**Unique**	**Method**	**Count**	**Unique**	**Method**
		assembler			**assembler**
1269 47.9%	700 26.4%	idba_tran	1082 42.2%	309 12.1%	idba_tran
686 25.9%	178 6.7%	soap-trans	692 27.0%	75 2.9%	soap-trans
515 19.4%	90 3.4%	trinity	569 22.2%	50 2.0%	trinity
1655 62.5%	1054 39.8%	velvet/oases	2089 81.6%	1285 50.2%	velvet/oases
		kmer			**kmer**
494 18.7%	107 4.0%	k05	458 17.9%	28 1.1%	k05
822 31.0%	261 9.9%	k25	957 37.4%	174 6.8%	k25
1089 41.1%	465 17.6%	k35	1177 46.0%	251 9.8%	k35
925 34.9%	293 11.1%	k45	1169 45.6%	200 7.8%	k45
883 33.3%	245 9.3%	k55	1085 42.4%	133 5.2%	k55
731 27.6%	165 6.2%	k65	1203 47.0%	266 10.4%	k65

540 20.4%	103 3.9%	k75	1070 41.8%	192 7.5%	k75
411 15.5%	119 4.5%	k85	950 37.1%	136 5.3%	k85
240 9.1%	68 2.6%	k95	787 30.7%	70 2.7%	k95

Footnotes to Tables 3

Statistics: **Count** is number of gene loci classified as most accurate, including perfect duplicate assemblies, and percent of total gene loci. **Unique** is that subset of accurate assemblies produced by a single method (assembler or kmer size). **Method** is the assembler or kmer setting, where kmer options were used spanning the read size of 100 bp for 3 of the methods. Trinity has a single kmer setting of 25 (originally, now adjustable up to 32 maximum).

Velvet/Oases remains the single best gene assembler, reconstructing 60% to 82% of accurate genes, but note that each assembler contributes some uniquely best genes. The majority of genes are most accurately assembled with kmer (read shred size) at or above 1/2 read length of 100 bp. Trinity is less capable in part due to its restricted kmer choice, and lack of scaffolding with read pairs. There are various comparison papers, contradicting each other, on how to choose an accurate gene assembler. One reason for those contradictions is that some comparisons use only 1 kmer setting, which isn't good, or use error-prone ways of merging multiple gene assemblies. A second disagreement is in the proper measures of gene accuracy.

The Evigene approach is to produce and assess millions of gene assemblies for coding sequence qualities, selecting the most complete genes from among the large collection of incomplete or inaccurate assemblies. Many gene assembler comparison papers focus on technical measures like "N50" length of transcripts, or "reads-mapped-back" counts of gene fragments recovered. These are not primary biological accuracy measures, as they don't assess the protein code, nor reflect the fact that genes have a proper size, often not much longer than the sequenced fragments. Extending genes beyond true size to improve a length or read map count creates artifacts.

A simple, meaningful gene set quality statistic is the average length of 1000 longest proteins*, which has biological maxima, is quick and easy to calculate, and will usefully compare gene sets of same and related species. The most precise and meaningful measure of coding gene accuracy is via homology assessment to other species. Protein size has a strong positive correlation with conserved homology to other species, so it serves as a secondary measure for recently evolved or undiscovered genes that lack strong species homology.

* Longest 1K proteins statistic, http://eugenes.org/EvidentialGene/about/EvidentialGene_quality.html

Table 4. Quality effect of gene set construction methods for vertebrate fish, comparing methods of NCBI Eukaryote Annotation (nc), EvidentialGene (evg), MAKER2 (mk), EvidenceModeller (em), within and across species, from Killifish genome paper-in-review. Subset of highly conserved genes (BUSCO vertebrate) is given. Species are kfish = *Fundulus heteroclitus*, atlantic killifish; amolly = *Poecilia formosa*, amazon molly; notfur = *Nothobranchius furzeri*, african turquoise killifish; pike = *Esox lucius*, northern pike.

Geneset	nFound	%Found	%Align	%Tiny	%Big
kfish.evg	4045	98.7	91.5	0.3	0.5
kfish.nc	4031	98.4	89.5	1.3	0.7
notfur.em	3996	97.5	87.5	2.9	1.1
notfur.mk	3726	90.9	76.6	5.2	2.2
pike.nc	4060	99.1	93.2	0.6	0.6
pike.mk	3114	76.1	56.5	7.8	0.2
amolly.nc	4050	98.9	92.2	0.8	0.8

Table 5. *Fundulus* coding sequence loci (Funhe), found in related fish using blastn. Related fish are amolly, guppy (*Poecilia reticulata*), notfur (notfur.em set), and zfish (*Danio rerio*, distant relative). Ortholog loci share 1:1 clustering among genes of 2+ fish. Non-ortholog loci do not cluster with other fish gene models, but include some with or without protein homology. Inparalogs are excluded from both subsets as determining uniqueness is complex. See http://eugenes.org/EvidentialGene/killifish/Genes/inotherfish/

Table 5A. Ortholog Funhe loci (n=21099) found in 4 related fish

N_Fish	Genome_Assembly	Gene_Models
1+	20973, 99%	19979, 94%
2+	20896, 99%	19337, 91%
3+	20462, 96%	17534, 83%
4	18129, 85%	na
0	126, 0%	1120, 5%

Table 5B. Non-ortholog Funhe loci (n=10169) found in 4 related fish

N_Fish	Genome_Assembly	Gene_Models
1+	7897, 77%	1899, 18%
2+	7182, 70%	871, 8%
3+	5881, 57%	352, 3%
4	3670, 36%	na
0	2272, 22%	8270, 81%

Table 6. *Pinus taeda* Loblolly pine tree x Arabidopsis REFERENCE (nr=15812), gene sets of Neale et al. 2014

Statistic	Evigene	MAKER
found	98.5%	91.5%
align	87.4%	70.8%
best	64.6%	18.6%
equal	17%	

References

Gilbert, DG. 2013 Gene-omes built from mRNA seq not genome DNA. 7th annual arthropod genomics symposium. Notre Dame. http://eugenes.org/EvidentialGene/about/EvigeneRNA2013poster.pdf and http://globalhealth.nd.edu/7th-annual-arthropod-genomics-symposium/

Goldfeder, et al. 2016. Medical implications of technical accuracy in genome sequencing. Genome Medicine. http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0269-0

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. 2011 Full- length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29: 644–652

Hall A.B. et al. 2015. A male-determining factor in the mosquito Aedes aegypti. Science vv; doi: 10.1126/science.aaa2850

Haznedaroglu et al. 2012. Optimization of de novo transcriptome assembly from high-throughput short read sequencing data improves functional annotation for non-model organisms. BMC Bioinformatics, 13:170. doi:10.1186/1471-2105-13-170

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

Neafsy D, Waterhouse RM, et al. 2015 Highly evolvable malaria vectors:the genomes of 16 Anopheles mosquitoes. Science 347:6217; doi:10.1126/science.1258522;

Matthews et al. 2016. The neurotranscriptome of the Aedes_aegypti mosquito. BMC Genomics 17:32; doi:10.1186/s12864-015-2239-0;

Neale et al. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology 15:R59. doi:10.1186/gb-2014-15-3-r59

Peng Y, Leung HC, Yiu S-M, Lv M-J, Zhu X-G, Chin FY. 2013. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics, 29:i326–i334; doi:10.1093/bioinformatics/btt219

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics, 28, 1086–1092.

Simao, FA, Waterhouse, RM, Ioannidis, P, Kriventseva, EV & Zdobnov, EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210-3212 .

Trachana K, et al. 2011. Orthology prediction methods: a quality assessment using curated protein families. doi: 10.1002/bies.201100062

Xie Y, Wu G, Tang J, Luo R, Patterson J, et al. 2013 SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. arXiv:13056760. ; Bioinformatics

Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, et al. 2011 Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 12: S2.

Products Listing

Recent publications and products from this project

Daphnia magna transcriptome by RNA-seq across 12 environmental stressors, **Scientific Data**, 2016 accepted; by Orsini L, DG Gilbert, R Podicheti et al.; **NCBI Transcript Assembly Archive** annotated gene set, Bioproject PRJNA284518, GenBank Accessions GDIQ00000000 and GDIP00000000; *Acknowledged funding*: National Science Foundation (DBI-0640462).

Daphnia magna annotated genome, **Scientific Data**, 2016 in prep. by Colbourne J, M Pfrender, DG Gilbert, et al.; **NCBI Genbank** annotated genome, Bioproject PRJNA298946, GenBank Accession: LRGB00000000; *Acknowledged funding*: National Science Foundation (DBI-0640462 and XSEDE- MCB100147 to DGG). Project products also at http://eugenes.org/EvidentialGene/daphnia/daphnia_magna/

The Atlantic killifish (*Fundulus heteroclitus*) genome and the landscape of genome variation within a population. 2016, In review; by N M. Reid, C E. Jackson, DG Gilbert, et al. ; **NCBI Transcript Assembly Archive** annotated gene set, Bioproject PRJNA269174, Accession GCES00000000; **NCBI Genbank** annotated genome, Bioproject PRJNA177717 (in progress); *Acknowledged funding*: National Science Foundation (DBI-0640462 and XSEDE-MCB100147). Project products also at http://eugenes.org/EvidentialGene/vertebrates/killifish/

Genome Re-Annotation of the Jewel Wasp *Nasonia vitripennnis*, **BMC Genomics**, 2016, In review; by A Rago, DG Gilbert, JH Choi, et al. *Acknowledged funding*: National Science Foundation (DBI-0640462 to DGG); Project products at http://eugenes.org/EvidentialGene/arthropods/nasoniawasp/

Gulia-Nuss M, Nuss AB, Meyer JM, et al. [includes DGG] (2016). Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. **Nature Comm**. Feb 9;7:10507. doi: 10.1038/ncomms10507. This project products at http://eugenes.org/EvidentialGene/arthropods/deertick/

Gilbert DG (2016). Accurate and complete reconstruction of mosquito gene sets of *Aedes* (yellow fever, Zika virus disease vector) and *Anopheles* (malaria disease vector) with EvidentialGene; http://eugenes.org/EvidentialGene/arthropods/mosquito/

Gilbert DG (2015). EvidentialGene public software repository. Products at https://sourceforge.net/projects/evidentialgene/ and http://eugenes.org/EvidentialGene/

Recent publications and services using products of this project

Alabama Supercomputer Authority, www.asc.edu/supercomputing/software.shtml, Supported software includes EvidentialGene.

Swedish Bioinformatics Support, www.biosupport.se, doi: 10.1186/2047-217X-2-9, Supported software includes EvidentialGene.

Scott, A. D., Stenz, N. W. M., Ingvarsson, P. K. and Baum, D. A. (2016), Whole genome duplication in coast redwood (Sequoia sempervirens) and its implications for explaining the

rarity of polyploidy in conifers. New Phytol. doi:10.1111/nph.13930

Luo H, Xiao S, Ye H, Zhang Z, Lv C, Zheng S, et al. (2016) Identification of Immune-Related Genes and Development of SSR/SNP Markers from the Spleen Transcriptome of Schizothorax prenanti [fish]. PLoS ONE 11(3): e0152572. doi:10.1371/journal.pone.0152572

Franta Z, Vogel H, Lehmann R, Rupp O, Goesmann A, and Vilcinskas A. (2016). Next Generation Sequencing Identifies Five Major Classes of Potentially Therapeutic Enzymes Secreted by Lucilia sericata Medical Maggots. BioMed Research International, vol.2016, ID 8285428, 27p. doi:10.1155/2016/8285428

Lynch JA (2015). The Expanding Genetic Toolbox of the Wasp Nasonia vitripennis and Its Relatives. Genetics, Vol. 199, 897-904, doi:10.1534/genetics.112.147512

Visser EA, Wegrzyn JL, Steenkmap ET, Myburg AA, Naidoo S (2015). Combined de novo and genome guided assembly and annotation of the Pinus patula juvenile shoot transcriptome. BMC Genomics,16:1057; doi:10.1186/s12864-015-2277-7

Chen, S; McElroy, J. S; Dane, F; Goertzen, L R. (2015). Transcriptome Assembly and Comparison of an Allotetraploid Weed Species, Annual Bluegrass, with its Two Diploid Progenitor Species, Poa supina Schrad and Poa infirma Kunth. The Plant Genome; doi: 10.3835/plantgenome2015.06.0050

Faddeeva A, Studer RA, Kraaijeveld K, Sie D, Ylstra B, Mariën J, et al. (2015) Collembolan Transcriptomes Highlight Molecular Evolution of Hexapods and Provide Clues on the Adaptation to Terrestrial Life. PLoS ONE 10(6): e0130600. doi:10.1371/journal.pone. 0130600

Postnikova OA, Hult M, Shao J, Skantar A, Nemchinov LG (2015) Transcriptome Analysis of Resistant and Susceptible Alfalfa Cultivars Infected With Root-Knot Nematode Meloidogyne incognita. PLoS ONE 10(2): e0118269. doi:10.1371/journal.pone.0118269

Horn F, Uzum Z, Mobius N, Guthke R, Linde J, Hertweck C. (2015). Draft genome sequences of symbiotic and nonsymbiotic Rhizopus microsporus strains CBS 344.29 and ATCC 62417. Genome Announc. 3(1):e01370-14. doi:10.1128/genomeA.01370-14.

Liu Y, Lin-Wang K, Deng C, Warran B, Wang L, Yu B, et al. (2015) Comparative Transcriptome Analysis of White and Purple Potato to Identify Genes Involved in Anthocyanin Biosynthesis. PLoS ONE 10(6): e0129148. doi:10.1371/journal. pone.0129148

McTaggart SJ, Hannah T, Bridgett S, Garbutt JS, Kaur G, Boots M. (2015) Novel insights into the insect trancriptome response to a natural DNA virus. BMC Genomics. Apr 17;16(1):310. doi: 10.1186/s12864-015-1499-z .

Chen, S; McElroy, J. S; Dane, F, and Peatman E (2014) Optimizing Transcriptome Assemblies for Eleusine indica Leaf and Seedling by Combining Multiple Assemblies from Three De Novo Assemblers. The Plant Genome, v8; doi:10.3835/plantgenome2014.10.0064;

Duncan, R. P., Husnik, F., Van Leuven, J. T., Gilbert, D. G., Davalos, L. M., McCutcheon, J. P. and Wilson, A. C. C. (2014), Dynamic recruitment of amino acid transporters to the insect/ symbiont interface. Molecular Ecology, 23: 1608-1623. doi: 10.1111/mec.12627