Annual Report for Period: 04/2016-04/2017
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 research in biological sciences. Yet the popular genome informatics methods lag behind the high levels of accuracy and completeness in gene construction that are attainable with today's accurate RNA-seq data.

EvidentialGene is a genome informatics pipeline for gene construction that has a measurably high accuracy and completeness rate, for the range of animals and plants. This pipeline algorithm is simple and robust, compared to gene modeling pipelines, and often outperforms their gene reconstructions. 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 by others, as reported in several recent publications, and is installed at cyberinfrastructure centers around the world.

Recent gene reconstruction comparisons for plants *Arabidopsis*, a model organism, *Zea mays* corn crop plant, and *Pinus* pine trees demonstrate that this project is out-performing long-read gene sequencing with **Pacific Biosciences** methods, as well as popular short-read methods. Recent gene reconstructions for these plants and for insect white fly *Bemisia*, a cotton/crop pest, and crustacean water flea *Daphnia*, enviromental test species, are objectively superior to those published in 2016-2017 with popular informatics methods of **Trinity** gene assembly, **MAKER** gene prediction, and NCBI Eukaryote Genome Annotation pipelines.

In the coming project period, improvements and additional methods will be incorporated into the EvidentialGene pipeline: simplified gene data publication to databases at NIH-NCBI and EBI, non-coding gene classification, merging of methods for chromosome-free gene assembly and chromosome-based gene modeling. New gene set reconstructions for model (zebrafish, frog and mouse) and bio-medically valuable animals are planned.

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

EvidentialGene is in production use by others, as reported in several recent publications, and is installed at cyberinfrastructure centers around the world. In this recent year, the project package has been installed in countries uk, de, au, jp, fr, be, se, dk, tw, ca, es, and others, at research institutions, goverment and commercial venues including usda.gov, usgs.gov, nersc.gov, tacc.utexas, ucdavis.edu, jetstream-cloud.org, xsede.org, amazonaws.com compute clusters.

Genes reconstructed during the 2016-2017 project period for plants *Arabidopsis*, a model organism, *Zea mays* corn crop plant demonstrate that this project is out-performing long-read gene sequencing with Pacific Biosciences methods, as well as popular short-read methods. Recent gene reconstructions for these plants and for insect white fly *Bemisia*, a cotton/crop pest, and crustacean water flea *Daphnia*, enviromental test species are objectively superior to those published in 2016-2017 from popular informatics methods for the same species and RNA expression data sets, as summarized below in Tables E1-4 (Gilbert 2017a), including Trinity gene assembly, MAKER gene prediction, and NCBI Eukaryote Genome Annotation pipelines.

There are now a few public Pac-Bio RNA gene sets, and publications suggesting genes from single-molecule sequencing may be more accurate than genes from Illumina short reads. Such data for Arabidopsis model plant, Zea mays corn, and pine trees, provide an objective comparison with different results: fully assembled Illumina RNA produces more accurate sets, including for loci where both methods recover some transcripts, and for alternate and paralog transcript reconstruction.

A recent publication (Hoang et al 2017) is an independent comparison of Pac-Bio RNA versus Illumina RNA over-assemblies, using Evigene for Illumina gene data reduction. However, the authors have altered Evigene's gene data processing pipeline for Illumina assemblies only, to include filtering by longest-transcript clustering, which is known to select for mis-assemblies that are longer than true coding genes, and documented as part of Evigene project to reduce accuracy by up to 30% (Gilbert 2013, 2017b). It is fair to call this a "BIG DATA" problem in genomics: apparently logical changes to automated processing of millions of data entities (gene reconstructions here) can reduce accuracy by large percentages, in ways the authors are unaware of without extensive testing. It is one of the reasons this project's goal of producing an easy-to-use implementation of the full Evigene reconstruction methodology will be of value to genome sciences.

Objectives planned for project renewal

In the coming project period, improvements and additional methods will be incorporated into the EvidentialGene pipeline: simplified gene data publication to databases at NIH-NCBI and EBI, non-coding gene classification, merging of methods for chromosome-free gene assembly and chromosome-based gene modeling. New gene set reconstructions for model (zebrafish, frog and mouse) and bio-medically valuable animals are planned. Groups that can benefit from use these products include bioinformatics projects such as Generic Model Organism (GMOD), Galaxy, and JetStream.org and centers that support gene and genome annotation (Gilbert 2016).

Goal1: reimplement full Evigene pipeline (Fig 1B, Table 1) for Galaxy installations, JetStream.org
Goal2: Evigene-H hybrid of RNA and Chromosome gene reconstruction pipelines.

Goal3: Evigene-N non-coding gene classifier and annotations.
Goal4: improve gene data publishing to NCBI/EBI/DDBJ gene databases

Gene re-construction done independently of chromosomes avoids several error sources: genome mis-assemblies, transposon inserts and long intron difficulties, HMM predictor over-corrections, errors of reference proteins mapped to genome. Two forms of gene reconstruction and annotation work-flow paradigms are diagrammed in Figure 1, A. genes modeled on chromosomes, B. genes assembled from RNA. Paradigm B is simpler. They differ in evidence sources and methods of gene reconstruction, but share components of gene annotation. The components of Evigene-G (modeled on genome) and Evigene-R (reconstructed from RNA) are proxies for similar components of peer methods.

Figure 1. Gene reconstruction work flows, A. modeled on chromosomes, B. assembled from RNA.

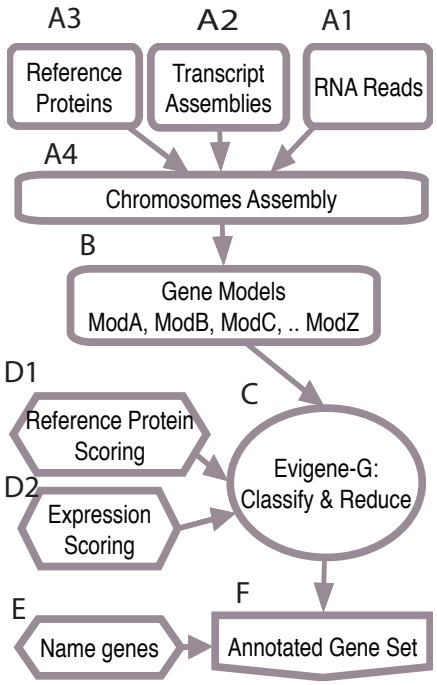

Figure 1A.

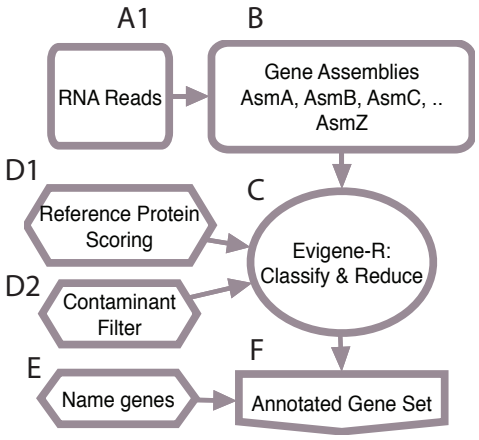

Figure 1B.

Paradigm A confounds use of reference proteins, key gene evidence, as both gene-model input and gene quality assessment (need the answer to get the answer), which B avoids. Paradigm A relies on accurate chromosome assemblies, often lacking, for accurate gene models. Paradigm A uses statistical modeling of gene structures, dependent on training (need the answer to get the answer), with difficulties in complex and non-standard gene structures. A subtle problem of complex paradigms that use semi-independent components is increased difficulty in validating gene reconstruction through all steps in a pipeline. If Paradigm B is simpler, and reconstructs genes as or more accurately than A, independent of confounding factors, it warrants further use in digital gene curation.

Current methodology works in steps, as components are developed, tested and refined separately. This requires expert effort to assess intermediate results and finish the serial pipeline process. Table 1 is a detailed list of Evigene components that correspond to gene reconstruction of Figure 1B.

Table 1. Evigene-R pipeline components of Figure 1B, and hybrid extensions (H1,N1)

A1. RNA data selection

B1. RNA data pre-processing: quality filters, formatting for assemblers, data normalization option.
B2. Assembler selection: Velvet/Oases, idba_tran, SOAPdenovo_trans, Trinity, others
B3. Assembler configurations: software choices, for Illumina pairs, other data, read-size and kmer steps, max alternates, other options
B4. Assembler runs, run failure handling
C1. Pre-process transcripts: **cdna_proteins, aaqual, trformat** for unique IDs, protein quality check
C2. Collected transcripts as input to Evigene tr2aacds, configure (minimal)
C3. Classify and reduce over-assembly set with **tr2aacds pipeline** to prefinished gene set, primary+alternates per locus
C4. Process tr2aacds output: **evgmrna2tsa** reformat, public IDs, other annotations
C5. Alternate transcript reclassification: **altbest, asmrna_altreclass, trclass2mainalt**
D1. Protein homology assessment: **orthomcl_evg**, **orthomcl_tabulate**, conserved domains from CDD, quality scores
D1. Evaluations: compare to other gene sets of same species for homology, location,
D2. Vector, contaminant screening of transcripts: **asmrna_trimvec**, **evgmrna2tsa**
E1. Protein function naming: **namegenes**
F1. Annotated transcript data submission: **evgmrna2tsa**, **geneattr** for tbl2asn formatting, gene names, scores
H1. Chromosome mapping: CDS-exon, intron locations, intron-chains to refine gene locus classifications. Resolve discrepancies between RNA-defined and chromosome-defined loci
N1. Non-coding gene classifications

Aspects of gene construction that are complex and need improvements in Evigene include the paralog/alternate problem. 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. 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 assemblies.

Basic theses of Evigene are (1) that any of a large set of models for a given locus can be deterministically measured and classified as biologically most accurate with gene evidence (this is in essence how expert annotators work), (2) that many different modeling programs/parameters are needed to produce among them the best gene model for each locus (we know this from years of genome projects), and (3) that models at each locus can and should be independently assessed for evidence (with gene-neighborhood metrics for joins, overlaps and such).

Evigene-H (hybrid of genome-modeled and RNA assembled). Combining mRNA assembled and genome-gene modeled genes is a valuable goal. The simple approach (Figure 1A model) of using mRNA assemblies as transcripts mapped onto chromosomes, as evidence for gene modeling is useful and works now. A careful approach to resolving discrepancies between mRNA assembled and genome modeled genes is not trivial, as there are several error sources, but is one that will benefit many genome projects. A hybrid of algorithms needs to merge the parts of Fig. 1A (A4 chr-map, B gene models) and Fig. 1B (B gene assembly), where parts C, D are extended to handle both chr-map data and gene alignment data, scoring accuracies both ways and resolving disagreements in those.

Results from 2016-17 project period have used chromosome-gene models x transcriptome merging, finding obvious errors from both sources: chromosome gaps and mis-assemblies, poor gene predictions, poor mRNA assemblies. Model errors can be greater than experience with only genome-modeled genes suggests, in part because chromosome assemblies with high mis-assembly rates are common.

This aim's approach to merging genome and transcriptome models of involves assigning quality scores to each locus model, derived from chromosome mapping quality, orthology measures, consensus

among population transcriptomes when available, and other, with allowance for errors in each measurement and weighting by attribute reliability (e.g. orthology to reference genes is more reliable than alignment to a chromosome assembly of unknown quality). Full automation of EvigeneH requires further work to add as a reliable pipeline. The primary work for this aim is to convert the genome-mapping requirement of Evigene-G to genome-map qualities comparable with Evigene-M and transcript assessment methods (Fig. 1A,B). Information engineering for this involves making tables with model IDs (i.e. object-oriented database), locus alignment and quality scores. Design of appropriate classifier(s) of these tables will classify by qualities of gene models. This differs from other gene informatics projects that rely on one or the other separate workflows show in Fig 1.

Evigene-N. Non-coding RNA genes are now discarded by Evigene-R (the mRNA classifier), but of course they are genes important to organisms, often are strongly expressed, and should be classifiable from RNA-seq assemblies or transcripts. Recent work indicates a range of ambiguity in coding versus non-coding genes, so that assessment of both will improve results of each form. This aim is in early development stage, with simple but unreliable non-coding measures. It requires investigation of methods to validate non-coding constructs in absence of distant-species sequence homology. Methods to test as reported in literature include population and related species consensus, expressed read back-mapping quality.

Significance

Reconstruction from RNA only provides independent gene evidence, free of errors and biases from chromosome assemblies and other species gene sets. Not only are the easy, well known ortholog genes reconstructed well, but harder gene problems of alternate transcripts, paralogs, and complex structured genes are usually more complete from Evigene methods.

Who should consider EvidentialGene for gene reconstruction?
* genomicists who want accurate, complete and objectively reconstructed genes, including those of you who may not believe my claims, but will look at objective results on this.
* model and well-supported genome projects, where curators can use these to improve precision of high value gene information.
* new species genomes, use as a primary gene set, with alternate transcripts, and/or assess gene predictions, chromosome assemblies for accuracy.
* gene/genome improvement projects, to add alternate transcripts, un-discovered and fragmented gene models.
* transcriptome and expression projects for more accurate genes.

A goal of this project is to reconstruct, and facilitate others to do so, many high-value (model, otherwise) animal and plant gene sets in coming years. This wants an easier to use, full implementation of Evigene methods in public platforms such as Galaxy and JetStream cyberinfrastructure projects, and through collaborations. This methodology is highly automatable (in a BIG DATA way), but still wants improvements. Species genes built with Evigene by independent authors include a range of plants and animals, and several of these papers provide independent reviews of Evigene versus other methods.

References

Gilbert D. (2013) How to get Best mRNA Transcript assemblies.
http://arthropods.eugenes.org/EvidentialGene/evigene/docs/perfect-mrna-assembly-2013jan.txt

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 D. (2017a). Animal and Plant gene set reconstructions with EvidentialGene. http://arthropods.eugenes.org/EvidentialGene/about/evigene_plantsanimals_2017sum.html
Gilbert D. (2017b). Error of using cd-hit-est longest-transcript-filter for gene assembly reduction versus using CDS quality filter as by EvidentialGene. http://arthropods.eugenes.org/EvidentialGene/evigene/docs/cdhiterr-arabidopsis-example.txt
Hoang NV, A Furtado, PJ Mason, A Marquardt, L Kasirajan, PP Thirugnanasambandam, FC Botha and RJ Henry (2017). A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. BMC Genomics (2017) 18:395; DOI 10.1186/s12864-017-3757-8

Project products

Recent publications and products from this project

Reid NM, CE Jackson, Don Gilbert, P Minx, MJ Montague, TH Hampton, LW Helfrich, et al, A.Whitehead (2017). The Landscape of Extreme Genomic Variation in the Highly Adaptable Atlantic Killifish. Genome Biol Evol 2017; 9 (3): 659-676. doi: 10.1093/gbe/evx023

Independent research publications and services using products of this project

9 from 2016, DOI: 10.1101/076067, 10.1111/nph.13930, 10.1371/journal.pone.0152572, 10.1155/2016/8285428, 10.1371/journal.pone.0157783, 10.1534/g3.116.028639, 10.1093/gbe/evw221, 10.1111/mec.13945, and PhD thesis http://escholarship.org/uc/item/1ww2p6mk

10 from 2017 (thru June), DOI: 10.1371/journal.pone.0171908, 10.1016/j.tplants.2017.03.001, 10.1186/s12864-017-3623-8, 10.1101/126177, 10.1371/journal.pone.0176956, 10.1111/jipb.12538, 10.1093/jxb/erx014, 10.1093/gbe/evx082, 10.1186/s13068-017-0828-7, 10.1186/s12864-017-3757-8 See also a recent NCGAS cyberinfrastructure use of Evigene: S. Sanders and M. Pfrender, 2017, http://hdl.handle.net/2022/21599 reconstructing salamander genes.

Tables from Gilbert (2017a)

E1. Plant model *Arabidopsis thal.* gene reconstructions compared

	AtAraport genes		Cacao genes		Introns
Geneset	Found%	AlignT%	Found%	AlignF%	Found%
AtAraport	--	--	88.7	70.6	88.1
AtEvigene	95.4	95	89.8	70.2	87.6
AtOases	90	91.2	na	na	81.1
AtIDBAtr	89.5	89.1	na	na	80.7
AtSOAPtr	88.9	87	na	na	79.1
AtTrinity	88.4	84.1	na	na	81.4
AtPacBio	58.1	48.2	64.2	60.5	56.3

E2. Corn *Zea mays* gene reconstructions

	Sorghum genes		Introns
Geneset	Found%	AlignT%	Found%
ZmEvigene	82.9	91.1	68.7
ZmGramene	81.9	90.3	68.1
ZmNCBI	81.3	89.6	na
ZmPacBio	78	82.4	68.2
ZmJgi4	77.6	81.2	68.9

E3. Whitefly *Bemisia tabaci* gene reconstructions compared

	Reference species				RNA
	Pea aphid		Fruit fly		Introns
Geneset	Found%	AlnT%	Found%	AlnT%	Found%
BtEvigene	81.2	88	74.1	74.9	68.5
BtNCBI	79.7	82.3	73.4	71.6	69.4
BtMaker	77.4	73.8	72.1	66	57.7
BtTrinity	73.5	59.2	68	53.2	50.5

E4. Water flea *Daphnia pulex* gene reconstructions compared

	Reference species				RNA
	Daphnia magna		Fruit fly		Introns
Geneset	Found%	AlnT%	Found%	AlnT%	Found%
DpEvigene	72	88.6	67.9	80.3	66.6
DpMaker	58.9	69.9	64.3	74.5	46.7

Arabidopsis gene set versions
AtAraport = public gene set of 2016 of Arabidopsis thal. from Araport.org
AtEvigene= Evigene classification/reduction of Illumina RNA assemblies
http://arthropods.eugenes.org/EvidentialGene/plants/arabidopsis/evigene2017_arabidopsis/
AtOases = Velvet/oases assembly of Illumina RNA,
AtIDBAtr = idba_tran asm of Ill. RNA,
AtSOAPtr = SOAP-Trans asm of Ill. RNA,
AtTrinity = Trinity asm of Ill. RNA,

AtPacBio = Pacific Biosciences SMRTAnalysis software assembly of Pac-Bio RNA data

Corn gene sets
ZmEvig = Evigene Zeamay5fEVm 2016 assembly of Illumina RNA-seq, public at
http://arthropods.eugenes.org/EvidentialGene/plants/corn/evg5corn/
ZmGram = Ensembl/Gramene 2016.09 Zm000nnnn,
ZmPacb = CSHL/Gramene PacBio gene assemblies of 2016 as SRA entries SRR3147024..054,
ZmNCBI = NCBI 2014 refgen zeamay
ZmJgi4 = JGI Rnnotator assembly set of Illumina RNA-Seq , 2014

Bemisia tabaci gene sets compared
BtEvig = Evigene gene assembly, 2016 update (vers 3), available at
http://arthropods.eugenes.org/EvidentialGene/arthropods/whitefly/whitefly3evigene/
BtNCBI = NCBI RefSeq gene models, 2016
BtMakr = Whitefly genome project genes modeled with MAKER, 2016, whiteflygenomics.org
BtTrin = TSA.GBII gene assembly 2015, Trinity of Illumina

Daphnia pulex gene sets
DpEvig7 Evigene genes of 2017 from
http://arthropods.eugenes.org/EvidentialGene/daphnia/daphnia_pulex/daphnia_pulex_genes2017/
DpMaker7 genes of 2017 from report of doi:10.1534/g3.116.038638

Measures
Genes Found% = percent of reference genes with significant alignment to gene sets (BLASTp/n of proteins or CDS),
Genes AlnT% = percent of aligned bases of reference gene bases
Introns Found% = percent of evidence introns aligned to gene set exons,
intron evidence from Illumina RNA-seq mapped to chromosome assemblies