Accurate & Complete Gene Construction with EvidentialGene

eugen.es.org/EvidentialGene/

Don Gilbert  2016 June  gilberetd@indiana.edu
What is EvidentialGene?

• **Classifier of gene models**
  Class = good, alternate, bad, redundant, coding, non-coding

• **Recipe for gene set reconstruction**
  Over-Assemble genes, from RNA-seq or/and gene predictions
  Use **Coding-Sequence Ruler** to select locus representatives
  Score **Homolog alignments to many reference proteins**
  Clean up and publish, via NCBI TSA, other routes

• **Variants**
  EvigeneD = genes modeled on chromosome DNA sequences,
  EvigeneR = genes assembled from RNA sequences,
  EvigeneH = hybrid methods of gene reconstruction (*in development*)
- Protein sizes correlate well with homology, across gene sets of a species
- Coding sequences discriminate paralogs with the same protein
- CDS is simple to calculate, a coarse ruler of over-assembly set
- Homology (BLASTp) is fine ruler for reduced set
Why is Evigene Useful?

accurate, objective, simple  gene set reconstruction

Human protein-coding genome regions are unreliable for 20% of disease genes [1]
GMO with inaccurate genes can be a problem, eg. GMO Aedes mosquito [2]
Orthology measures are strongly affected by gene inaccuracies [3] (express functions also)

Over-assembly is better than one assembly

Each locus has own expression values, alternates, neighbor genes that affect assembly
Weaknesses of one method compensated by others

Gene assembly is better than genome-gene prediction

Transposons and long introns are major problems for genome-gene modeling
Chromosome assembly gaps, splits, mis-assemblies are common
Predictors make poor guesses, even with clean data, un-clean data are common, species training is needed
Complex loci, trans-splicing, odd structures on genome are poorly handled, but removed biologically from transcripts.

Mosquito examples of Evigene vs MAKER & Trinity of 2015-2016 publs [5,6]
EvidentialGene reconstructions have more complete alignment to reference genes, versus MAKER, Trinity, NCBI EGAP and other methods, for plants, fish, arthropods.

Evigene also builds more complete alternates and non-ortholog sets, but without objective ruler those can be questioned.
Assembly Results for Mosquito Genes

Method and KMER effects on best assembly of Highly Conserved Genes of *Anopheles funestes*

Gene Assembler Methods for Accurate Highly Conserved Genes (BUSCO)

- **Velvet/Oases**, 1.2.10 2013, https://www.ebi.ac.uk/~zerbino/oases/
How does Evigene work? (EvigeneR)

- **Over-assemble RNA** (100 assemblies/10 M transcripts from 200 M to 10 Bln accurate Illumina paired reads, with de-novo & chr-align methods, several data slices, kmer sizes, options)
- **Find coding sequences** (smart ORF/CDS finder)
- **Remove redundancy** (stepwise efficient: fastanrdb > cdhit)
- **Classify gene loci** (BLASTn, align exons, classify by overlap)
- **Assess orthology** (BLASTp reference genes, OrthoMCL)
- **Annotate public sequence products**
- **NCBI submission processing**
25 well-conserved Mediator subunit genes of RNA polymerase II transcription, ranging from 2800 aa to 120 aa in size.

Evigene reconstructions are compared to 2016 public sets of VectorBase using MAKER (*Anopheles*) and independent using Trinity (*Aedes*).

Evigene improvements in percent alignment are shown, above zero means improvement, tall bars mean other gene set missed it. Alignment is to *Drosophila* reference genes.
Evigene correction: Fragment \((\text{Anoph})\) & Miss \((\text{Aedes})\) of Mediator subunit gene.

\textbf{Anopheles fun.}

Evigene vs VB-Maker

\textbf{Aedes aegypti}
Details on how EvigeneR works

Algorithm of gene classifier, tr2aacds.pl

1. input transcripts, calculate CDS and AA sequences, work mostly on CDS.
2. perfect redundant removal with fastanrdb (fast)
3. perfect fragment removal with cd-hit-est
4. high-identity align of transcript exons with blastn, to match alternates of each locus.
5. classify main, alternate, redundant transcripts / locus with CDS-align, protein metrics.
6. output classified sequence sets: okay-main, okay-alts, drop (redundant).

Other components of gene assembly

rnaseq/trformat.pl: regularize and unique IDs for transcript.fasta.
omcl/orthomcl_evg.pl, orthomcl_tabulate.pl: protein orthology with BLASTp & OrthoMCL
prot/namegenes.pl: gene function names from UniProt and Conserved Domains (CDD).
rnaseq/asmrna_trimvec.pl: vector, gap and contaminant screening (suited for NCBI).
evgmrna2tsa.pl: check and annotate mRNA, public IDs and sequence files, write Genbank TSA format for public submission.

http://eugenes.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html
Challenge to Galaxy Jockeys

- Install & test drive Evigene in Galaxy

Others use it now at compute centers and in projects, without my help.

Evigene needs work on ease of use by non-computists (also computists).

“One button” recipe or script for automated gene assembly is possible, can better link components.

- Compare to other gene reconstruction method(s)

~3 days/mosquito to build over-assembly and best genes, weeks to assess & explain value to others.

I will help and collect Galaxy-ready scripts for Evigene

more details http://arthropods.eugenes.org/EvidentialGene/about/ProjectReports/

Let’s Make Genomes Great Again!
Collaborators and Data Providers

Cyber-infrastructure: TeraGrid/XSEDE, NCGAS

Genome projects: Cacao Tree, Daphnia Water fleas, Fundulus Fish, Loblolly Pine, Nasonia Wasp, Pea Aphid, and NCBI SRA public data sets

References


This Project
eugenews.org/EvidentialGene/ or sourceforge.net/projects/EvidentialGene/
Evigene Compared to ...

- Related gene assembly methods
  Velvet/O Merge, CAP/extra-assembly, .. don’t help much, add errors
  Other Gene Annotation pipelines with RNA assembly .. see best ortho charts
  Rnnotator of JGI: similar Over-assembly, Reduce redundancy, NO Coding Ruler, fewer gene classing metrics

- No-assembly required
  Single molecule, long read PacBio .. better? is future here? .. very few publ

- Corn Genes Test Case
  PacBio long no-assembly + Gramene, Wang et al. (Doreen Ware lab) 2016, doi:10.1038/ncomms11708
  vs Illumina-short over-assembly, JGI w/ Rnnotator, Martin et al. 2014, doi:0.1038/srep04519
  vs Evigene with Illumina-short over-assembly (same RNA), D. Gilbert, in progress
Evigene: *Aedes aegypti* GMO region

US.FDA, 2016, Oxitec Mosquito: GMO *Aedes aegypti* OX513A
Sect 9.2.4 “The combined flanking sequence was compared with the relatively poorly annotated Ae.aegypti genome sequence, transcript and EST databases. The flanking sequence shows 94.6% identity to a single genome sequence contig (1.859). No new open reading frames were found inferring that **no genes appear to be disrupted by the #OX513 rDNA construct insertion** and no new genes are created.”
Evigene correction: Join of Small Gene

Anopheles fun.
Evigene correction: Join of Large Gene

Anopheles funestes