Accurate & Complete Gene Construction with EvidentialGene

eugenes.org/EvidentialGene/

Don Gilbert 2016 June gilbertd@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**

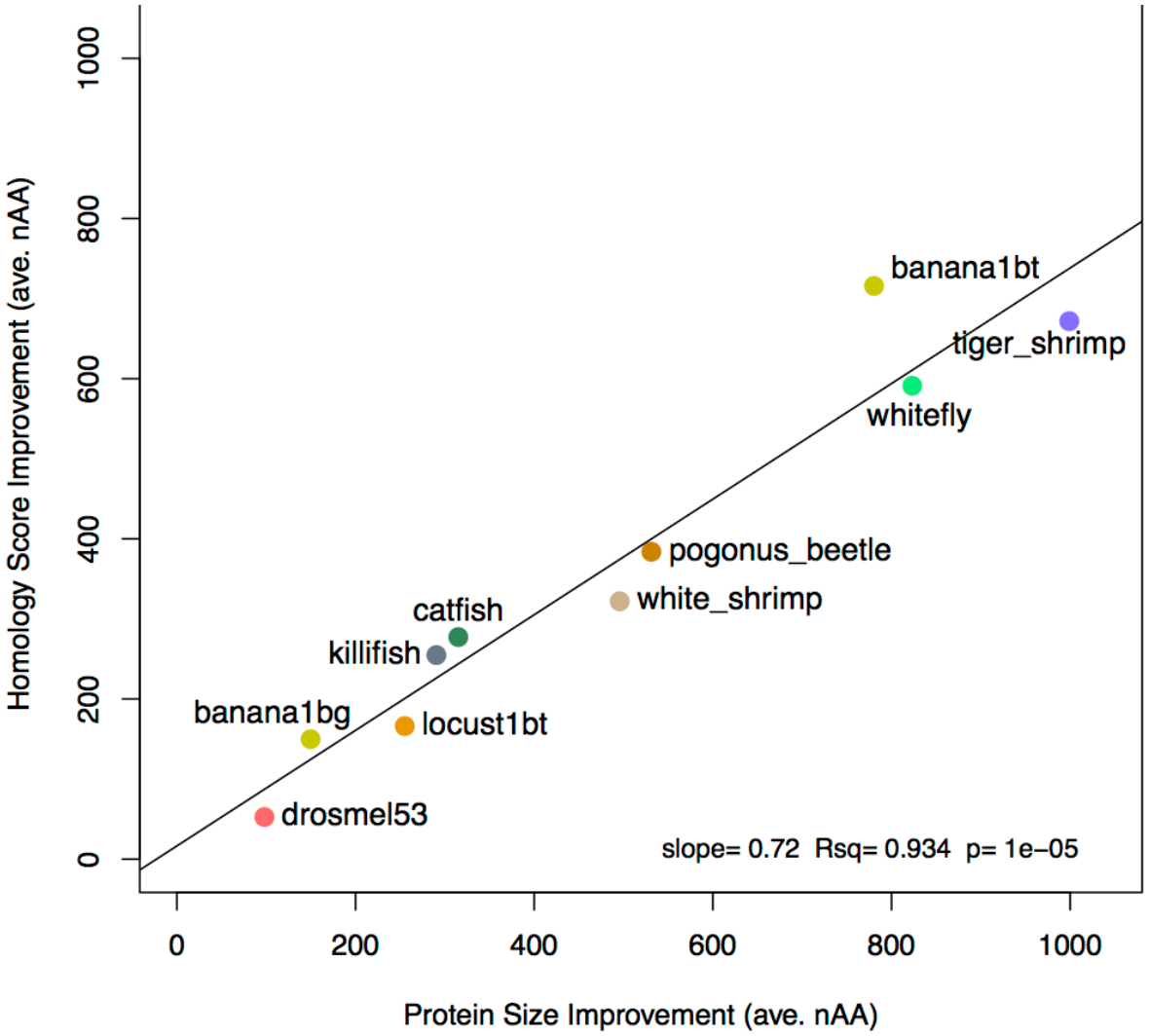
  Evigene**D** = genes modeled on chromosome DNA sequences,

  Evigene**R** = genes assembled from RNA sequences,

  Evigene**H** = hybrid methods of gene reconstruction (*in development*)

Coding Sequence Ruler

- Protein sizes correlate well with homology, across gene sets of a species
- Coding sequences discriminate paralogs with same protein
- Simple to calculate, a ***coarse ruler*** for 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

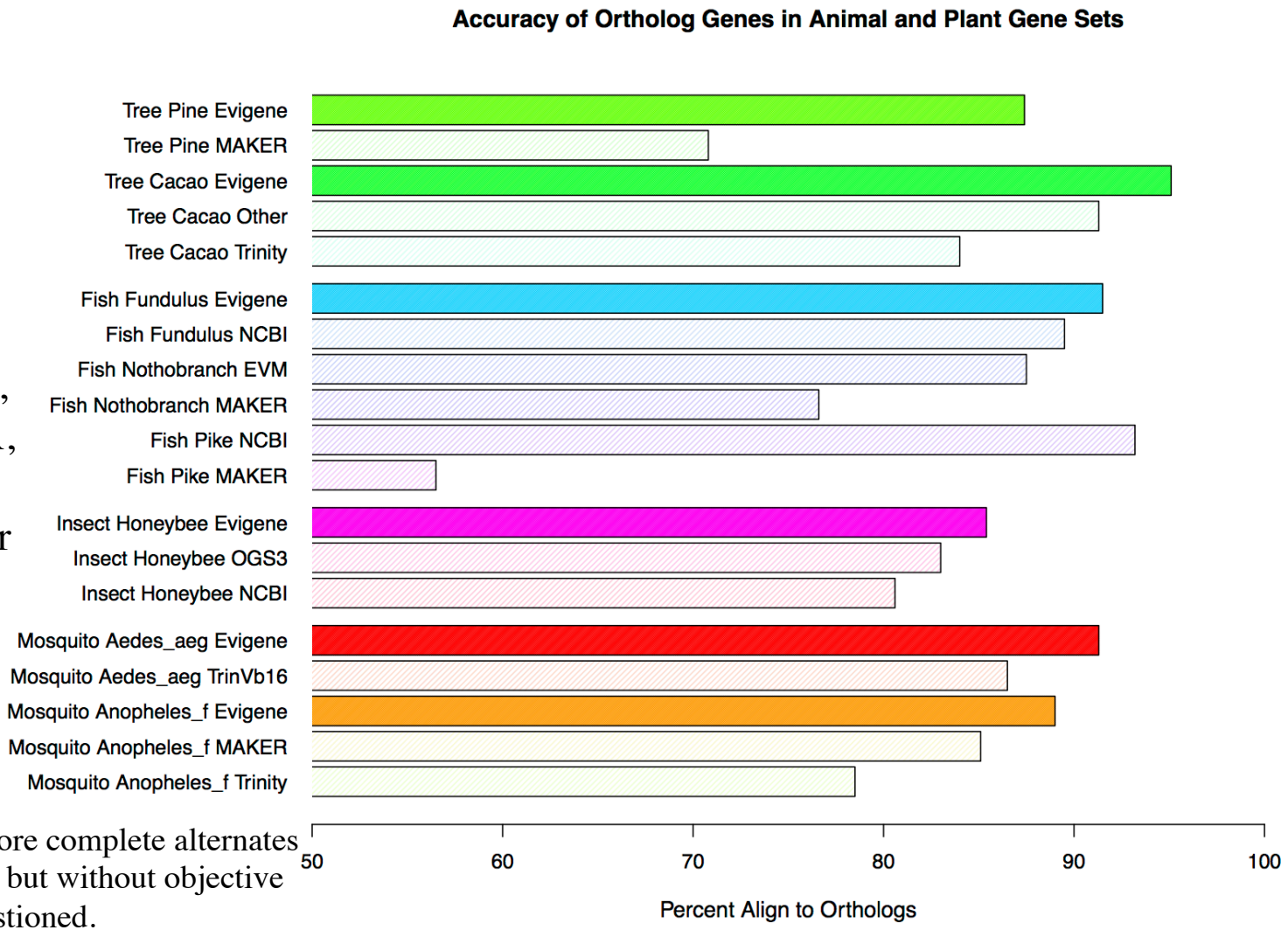
Gene assembly is better than genome-gene prediction

- 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]

Ortholog Accuracy in Animals & Plants

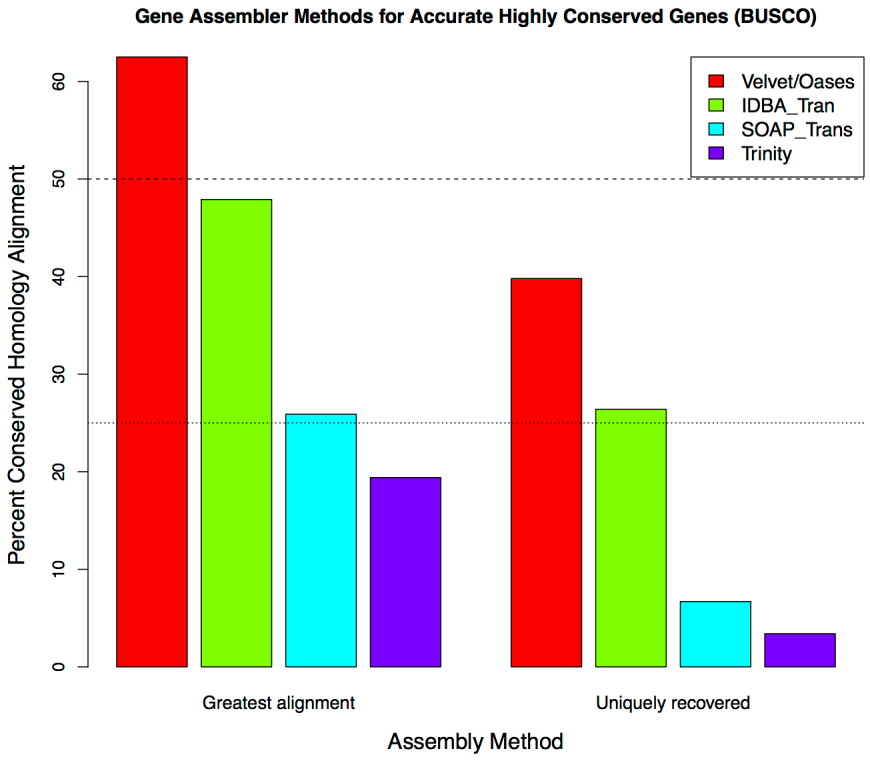
EvidentialGene reconstructions have more complete alignment to reference genes, versus MAKER, Trinity, NCBI EGAP and other methods, for plants, fish, arthropods.

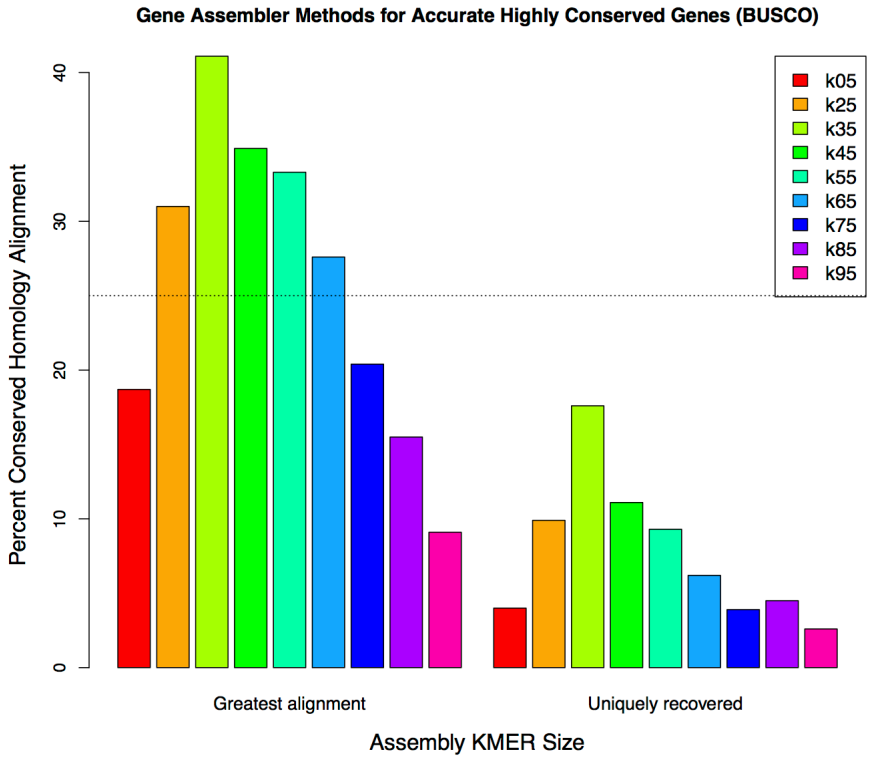


Evigene also builds more complete alternates and non-ortholog sets, but without objective ruler those can be questioned.

Assembly Results for Mosquito Genes

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





Velvet/Oases, 1.2.10 2013, https://www.ebi.ac.uk/~zerbino/oases/ **Trinity**, 2.1.1 2014, https://github.com/trinityrnaseq/trinityrnaseq/wiki

idba-tran, 1.1.1 2013, http://www.cs.hku.hk/~alse/idba_tran/ **SOAP-Tr**, 1.03 2013, http://soap.genomics.org.cn/SOAPdenovo-Trans.html

How does Evigene work? (EvigeneR)

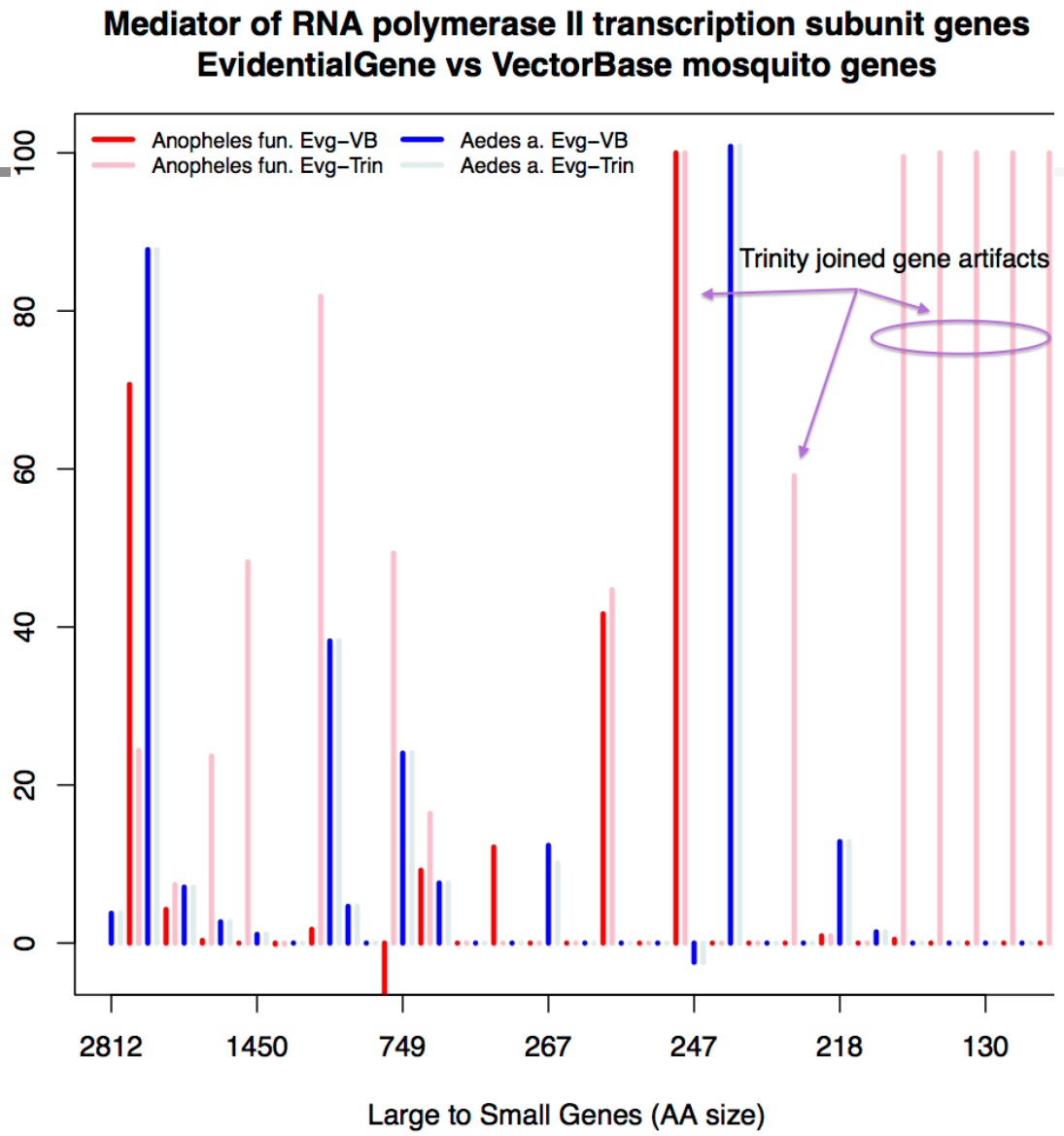
- Over-assemble RNA (100 assemblies/10 M transcripts from 200 M to 10 Bln accurate Illumina paired reads, with de-novo & chr-align methods, several data slices, kmer sizes, options)
- Find coding sequences (smart ORF/CDS finder)
- Remove redundancy (stepwise efficient: fastanrdb > cdhit)
- Classify gene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

Example Mosquito Gene Accuracy

25 well-conserved Mediator subunit genes of RNA polymerase II transcription, ranging from 2800 aa to 120 aa in size.

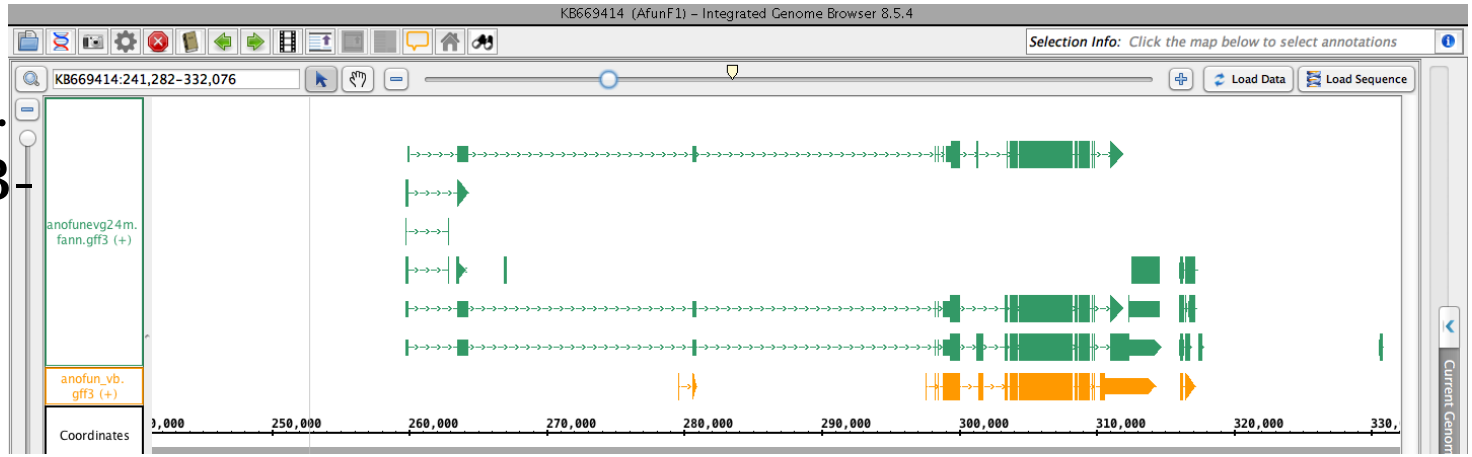
Evigene reconstructions are compared to 2016 public sets of VectorBase using MAKER (*Anopheles*) and independent using Trinity (*Aedes*).

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.

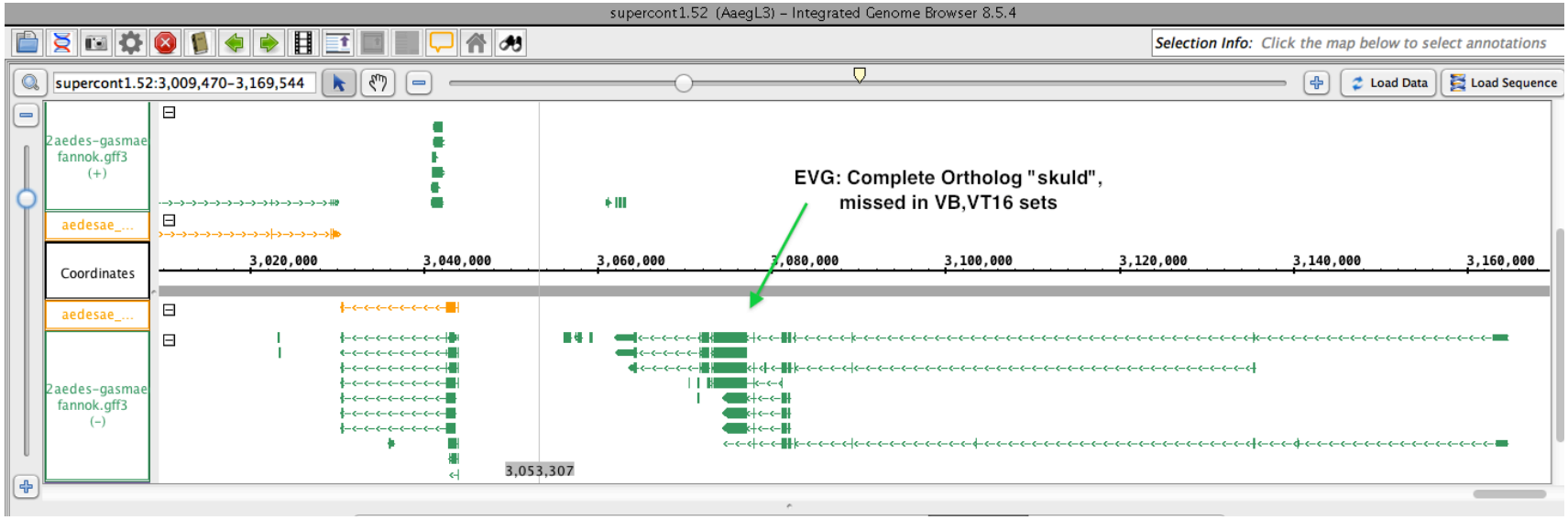


Evigene correction: Fragment (*Anoph*) & Miss (*Aedes*) of Mediator subunit gene.

Anopheles fun.
Evigene vs VB-Maker



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, but weeks to assess & explain their value to others.

I will help and collect Galaxy-ready scripts for Evigene

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

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

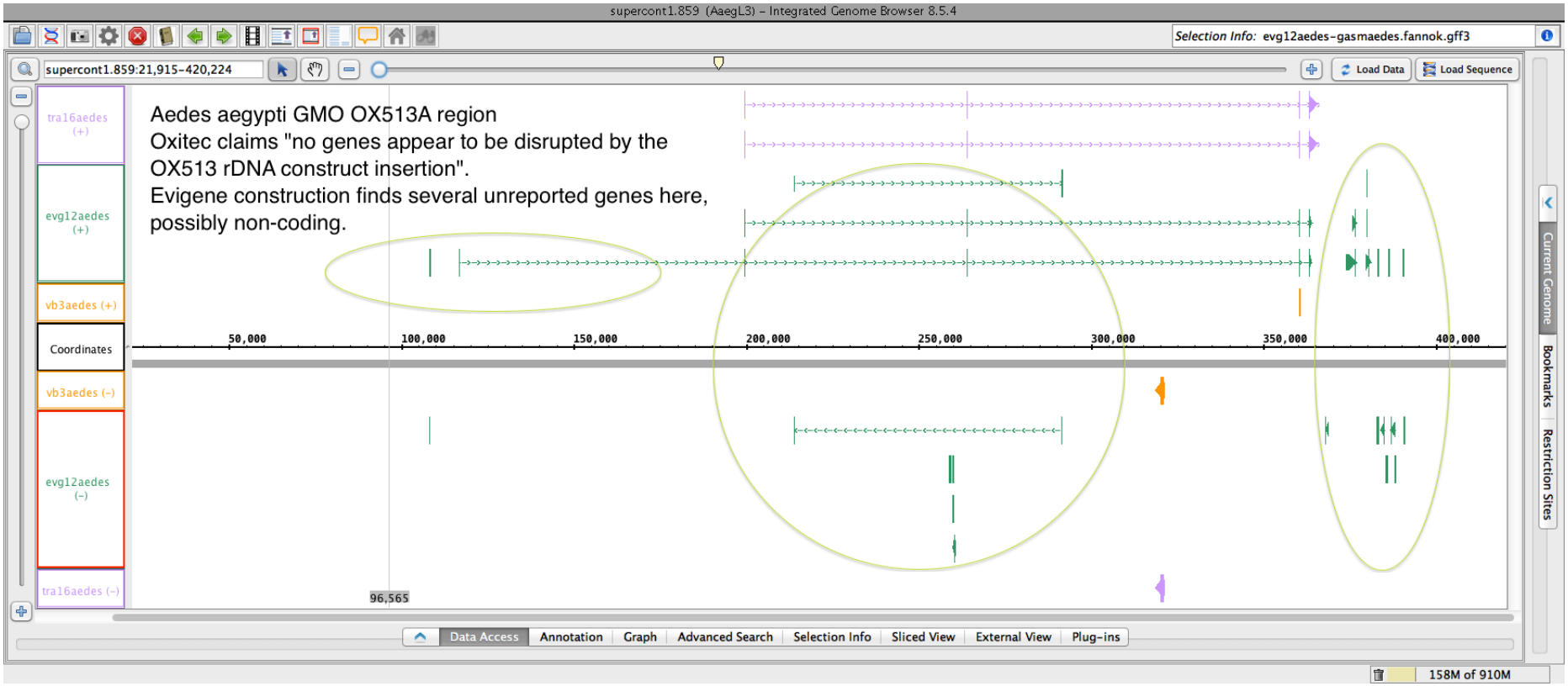
- No-assembly required

  Single molecule, long read PacBio .. better? is future here? .. very few publ

- Corn Genes Test Case

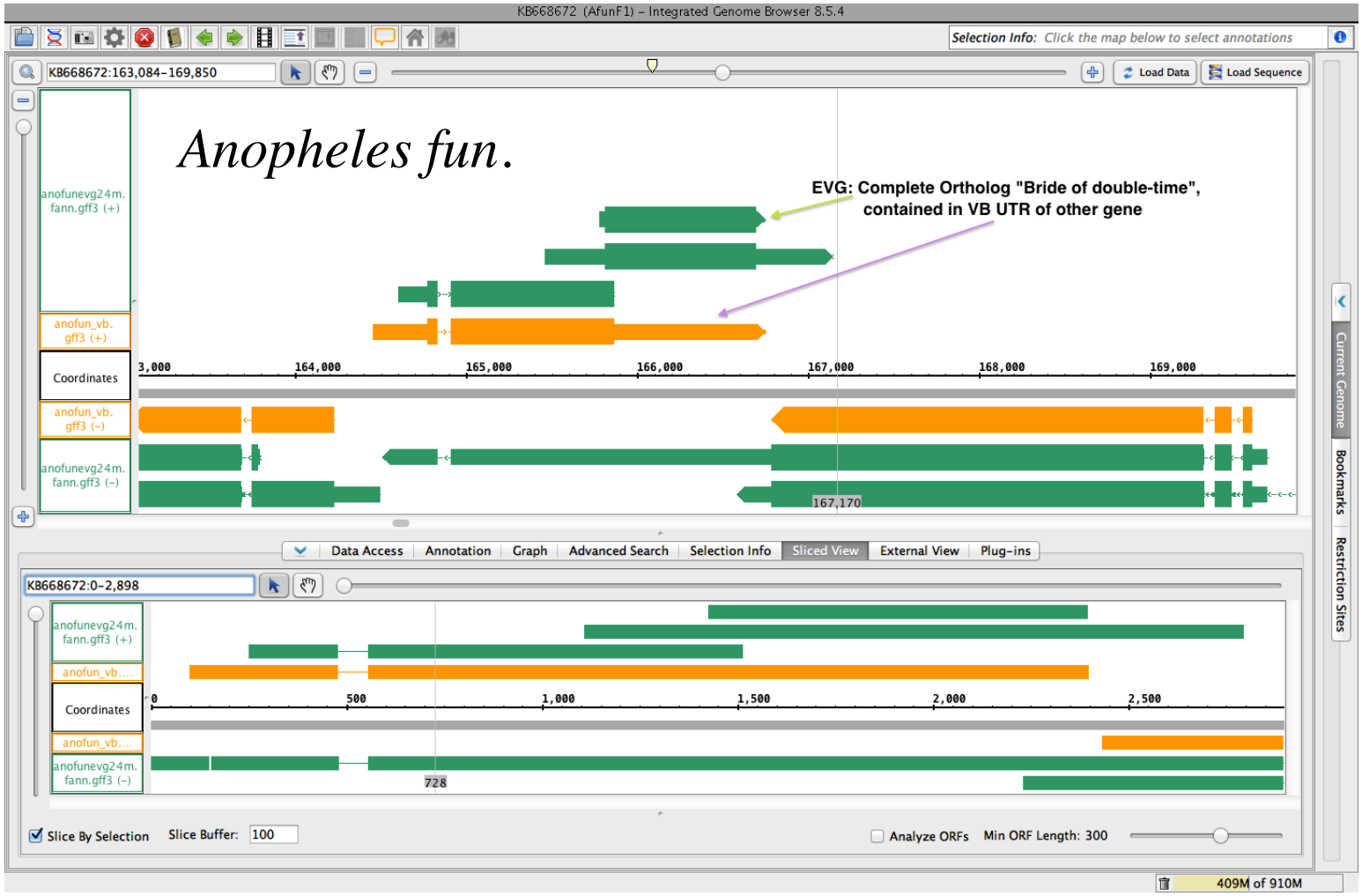
  PacBio long no-assembly + Gramene, Wang et al. (Doreen Ware lab) 2016, doi:10.1038/ncomms11708
  vs Illumina-short over-assembly, JGI w/ Rnnotator, Martin et al. 2014, doi:0.1038/srep04519
  vs Evigene with Illumina-short over-assembly (same RNA), D. Gilbert, in progress

Evigene: *Aedes aegypti* GMO region



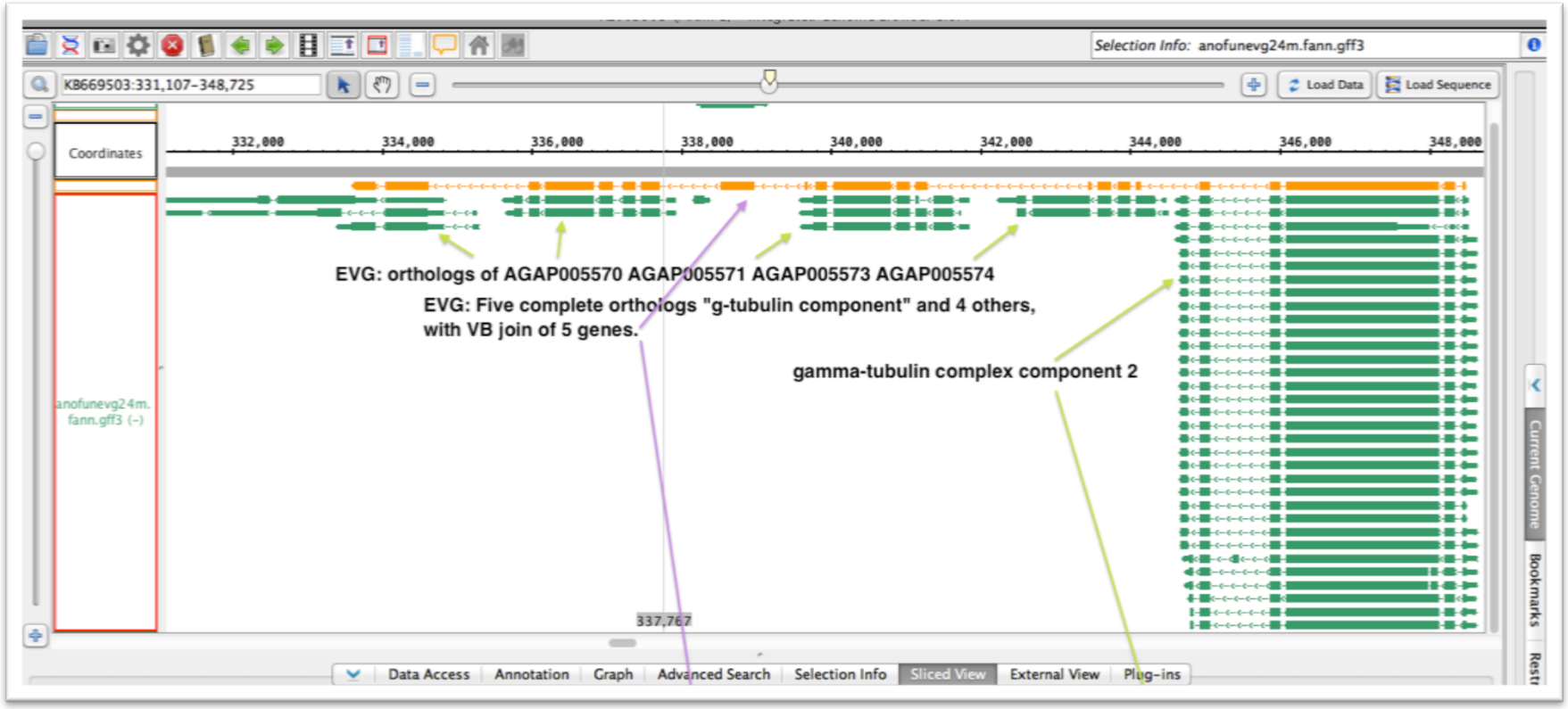
US.FDA, 2016, Oxitec Mosquito: GMO *Aedes aegypti* OX513A
http://www.fda.gov/AnimalVeterinary/DevelopmentApprovalProcess/GeneticEngineering/GeneticallyEngineeredAnimals/ucm446529.htm
Sect 9.2.4 "The combined flanking sequence was compared with the relatively poorly annotated Ae.aegypti genome sequence, 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



Evigene correction: Join of Large Gene

Anopheles funestes



Accurate Genes with Inaccurate Genomes: GMOD topics from EvidentialGene

eugenes.org/EvidentialGene/

Don Gilbert 2016 June gilbertd@indiana.edu

How best to use Evigene with genome projects?

- **Less expertise needed for reliably accurate genes**

EvigeneR recipe requires less expertise in set up, training, than gene modeling.

Accuracy does **not depend** on quality of chromosome assembly, gene predictor training, nor availability of accurate reference species genes.

It does depend on accurate and complete transcribed gene RNA data, but not as much as some think. I recover all but 1% to 5% orthologs, some as fragments, depending on data.

Organismal expression from many tissues, stages, stresses, sequenced with current Illumina paired-end, preferably strand-specific, at amounts of 75-100 Mil pairs/sample. Billions paired reads for more complete genes (rare alternates turn up w/ more data).

- **Gene assembly before chromosome assembly**

validate gene assembly with reference orthology, then validate/correct chromosome assembly with gene assembly

Genome Annotation Workflow, Olive tree

2016-June, lab of Toni Gabaldon, doi:10.1186/s13742-016-0134-5

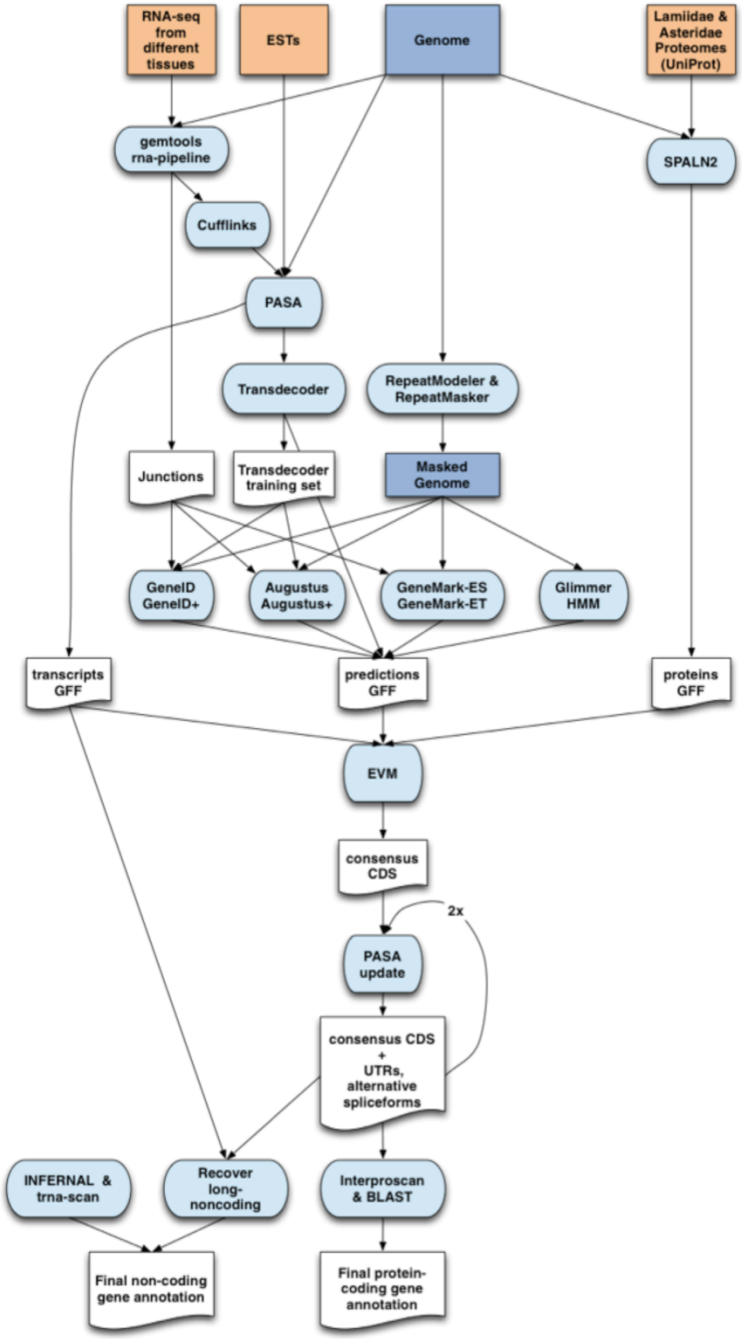


Fig. 5 Overview of the annotation pipeline. Input data for annotation are shown at the top of the flow chart. Computational steps are shown in light blue and intermediate data are shown in white

Evigene suggested Gene Annotation Workflow

Simplicity has value in tracing errors in gene reconstruction. Final checks (CDS orthology, ncRNA) trigger iterative refinements to raise accuracy.

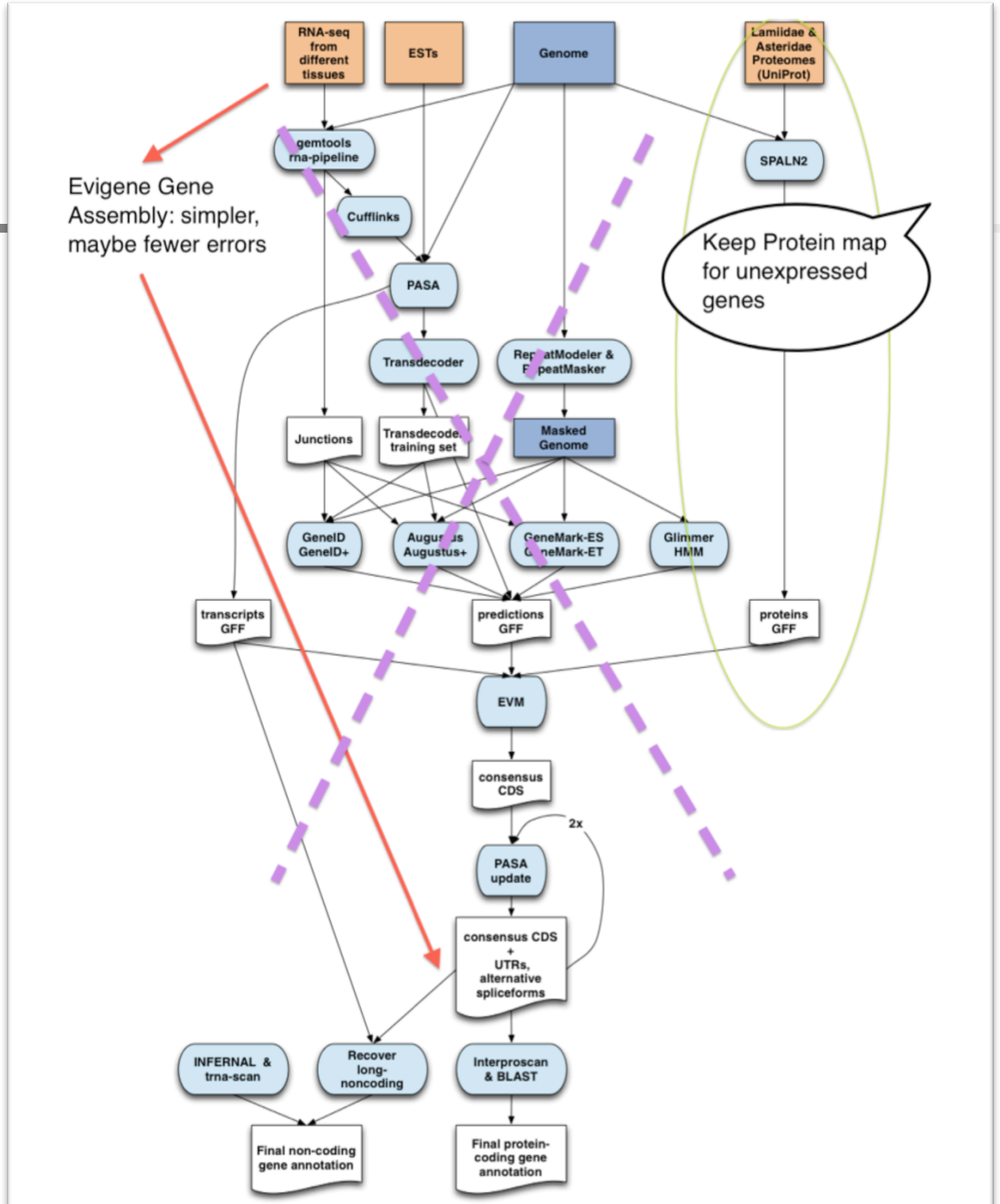


Fig. 5 Overview of the annotation pipeline. Input data for annotation are shown at the top of the flow chart. Computational steps are shown in light blue and intermediate data are shown in white

What changes to genome projects do Evigene results suggest?

Changes to genome projects to add gene assemblies that don't always match chromosome assembly:

Match, compare, and merge gene assembly with genome-predicted gene models. Trivial for many loci, hard for some.

Add more gene-focused results (annotated transcripts, gene tables)

Semi-independence of gene and chromosome data

Store outside of GFF/location-based data.

Add unlocated transcript-sequence as pseudo-genome scaffolds

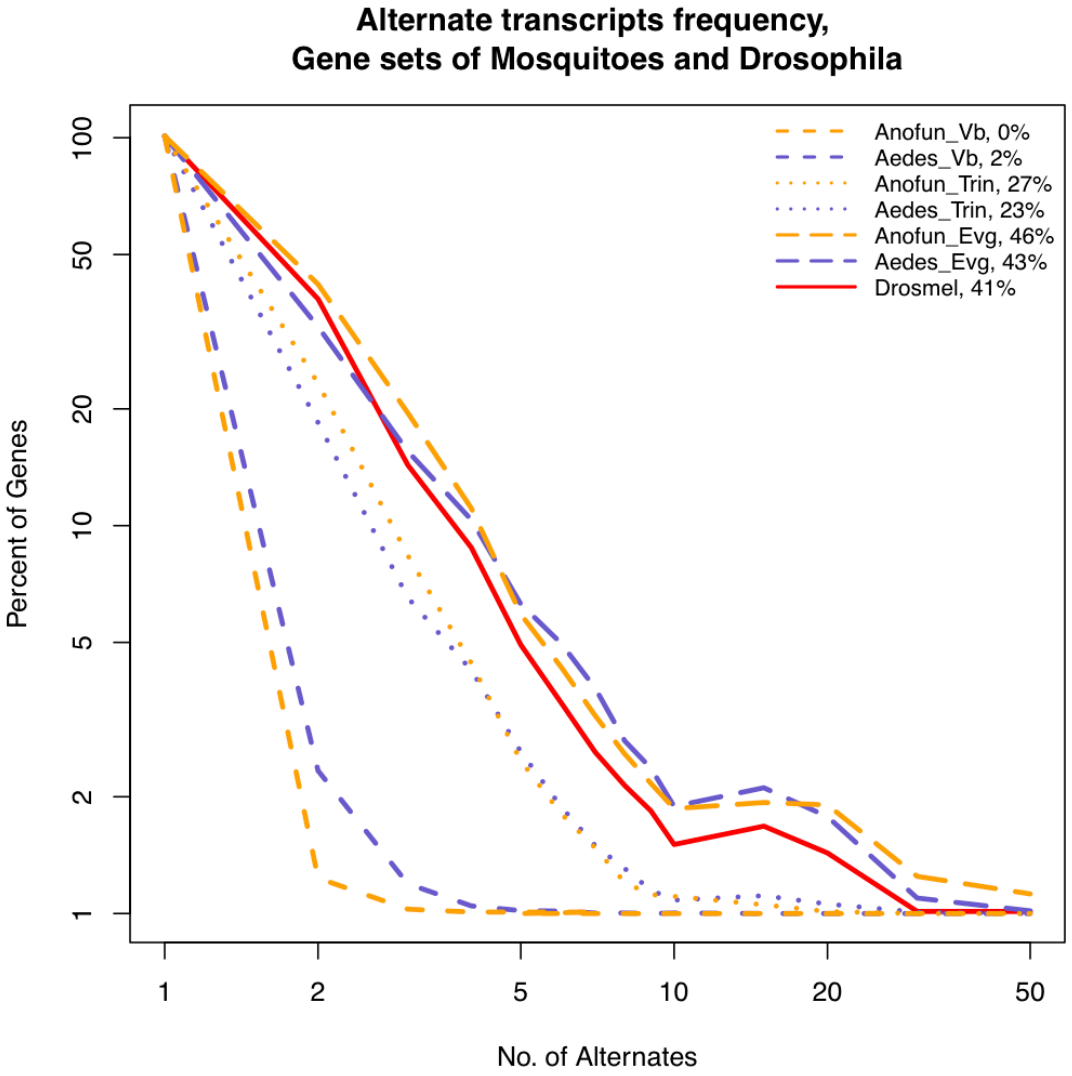
Alternate transcripts are ..

* Alternate transcripts are plentiful, in biology and gene assemblies, more than many projects now report.

* Alts are fully reconstructed by Evigene, in proportions similar to Fruit fly model organism, versus lower for Trinity, much lower for VectorBase-MAKER, in mosquitoes.

* Alts are sometimes confused with paralogs, w/ or w/o chromosome locations.

* Alts are needing more analyses: primary vs alternate is artificial distinction.



What of many Non-ortholog genes?

What to do with Non-ortholog genes?

"too many genes”? Many throw these away, without positive reason

population/related coding conservation (killifish example),

differential expression analyses find non-ortho response

long non-coding and maybe-coding genes (no coding conservation)

How does Evigene fit into genome projects?

How to assess gene set accuracy and completeness?

- Ortholog subsets CEGMA, BUSCO, and full assessment with OrthoMCL
- Expression analyses (read recovery, differential expression, ..)

Genes reconstructed from transcribed RNA avoid many problems of genome-structure: introns, transposons, alternate exons, trans-splicing, anti-sense splicing are DNA-gene problems.

Low expression is an transcribed-gene problem, and assemblers do put wrong pieces together (Evigene removes many).

Exceptions in Gene biology and Construction artifacts often **look the same**.. partial duplications/fragment models, fusions/gene-joins

Problems of hybrid gene sets

Hybrid gene set where genes are more accurate than chromosomes (or disagree in places).

a. Don't fit dogma nor existing public databases

b. Discrepancies can be hard to resolve

c. Genes != genome locations

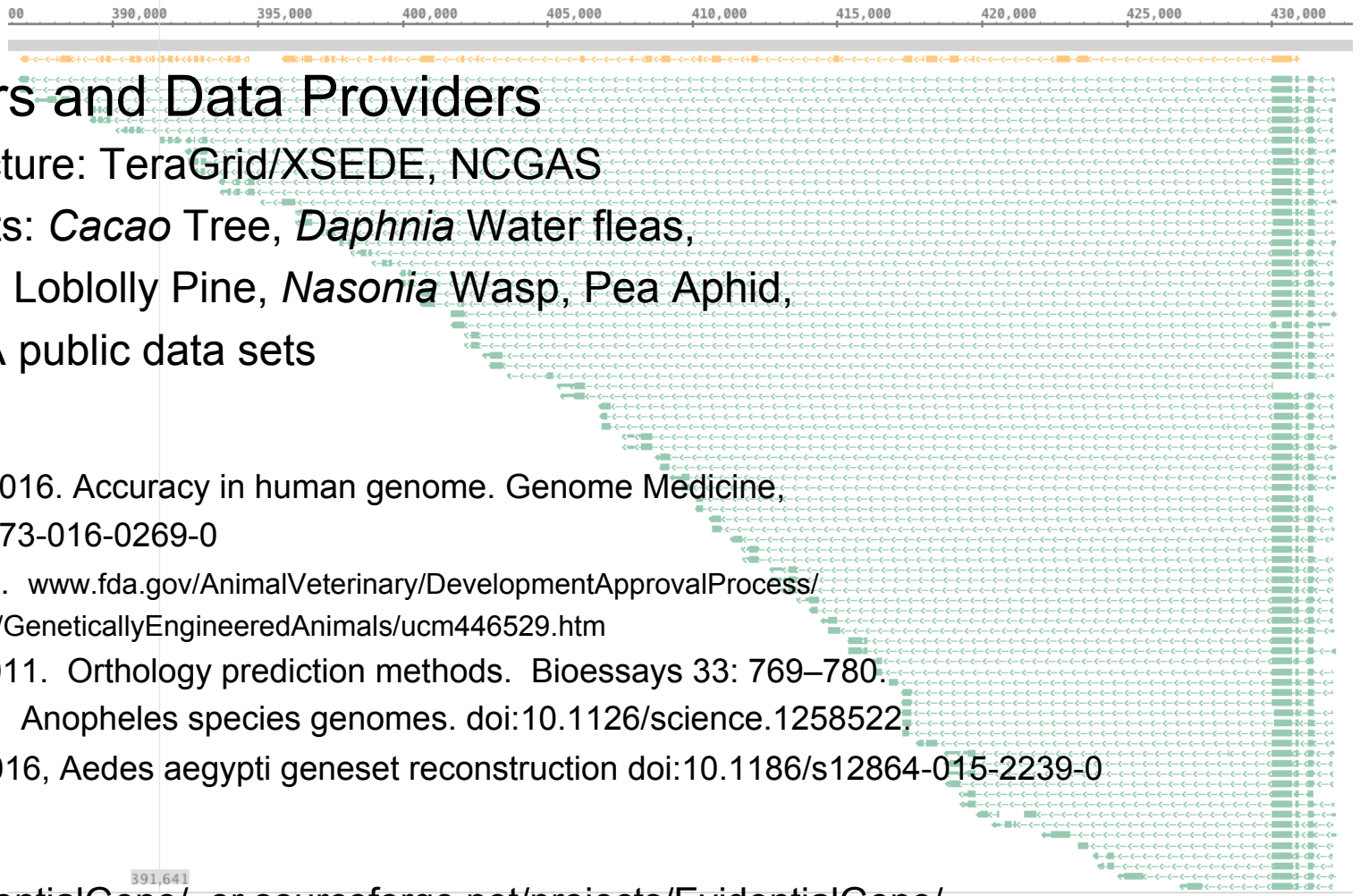
GFF, genome views **fail** to describe genes

Expert annotation via Apollo, etc., gm-views need adjust

Minimal update: publish gene transcript, & protein, cds sequences. Keep semi-independent annotations on genes.

End note

gilbertd@indiana.edu



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

1. Goldfeder et al., 2016. Accuracy in human genome. Genome Medicine, doi:10.1186/s13073-016-0269-0
2. GMO Aedes 2016. www.fda.gov/AnimalVeterinary/DevelopmentApprovalProcess/GeneticEngineering/GeneticallyEngineeredAnimals/ucm446529.htm
3. Trachana et al. 2011. Orthology prediction methods. Bioessays 33: 769–780.
4. Neafsy et al 2015, Anopheles species genomes. doi:10.1126/science.1258522.
5. Matthews et al. 2016, Aedes aegypti geneset reconstruction doi:10.1186/s12864-015-2239-0

This Project

eugenes.org/EvidentialGene/ or sourceforge.net/projects/EvidentialGene/