Accuracy and completeness of an Arthro-Tripod of gene-omes

Deer Tick, Honey Bee, and Water Flea, from mRNA-assembled and genome-modelled genes

2014 June

EvidentialGene   http://arthropods.eugenes.org/EvidentialGene/

Don Gilbert, Biology Dept., Indiana University, Bloomington, IN 47405, gilbertd@indiana.edu

Gene-ome construction for Arthro-Tripod

Gene sets of three species spanning the arthropod phylum are examined with recent mRNA assemblies and genome gene predictions. Apis mellifera, Daphnia magna, and Ixodes scapularis now have genes discovered from both approaches. Each approach has values and problems for biologically accurate knowledge of genes.

mRNA assemblies can produce species-complete sets of orthologs, unique genes, and alternate transcripts from direct gene evidence, without artifacts from other sources. Such assemblies surpass genome gene models in accuracy for new species projects, but with trade-offs in detail. This is evident for transposon-fragmented genomes like Ixodes, and paralog-rich genomes of Daphnia, versus more complete mRNA-gene assemblies. For the more mature Apis genome assembly and gene set, mRNA-assembly adds to accuracy of genes.

RNA-seq assembly struggles with too little or too much data for some genes, may err in discriminating alternates from high identity paralogs, and have other problems due to missing genome locations. Accurate genome assembly remains a harder problem, with modeling inaccuracies and ortholog biases adding to gene inaccuracy. Combining both approaches is best, when errors of both are addressed. Find informatics software and recent arthropod gene sets at http://arthropods.eugenes.org/EvidentialGene/

Changing the dogma of genome discovery projects is warranted, to first discover genes via mRNA assembly, adding genome assembly where cost-effective. Standards of quality for mRNA assembly can be improved, with better methods information, along with better assessment of errors from both approaches, to ensure high quality gene sets.

How to build perfect mRNA transcript assemblies Does & Don’ts

Don’t: Use one assembly program, default options. Do: Use several assembly programs and options, as each will provide some better transcripts/genes.

Don’t: Make many assemblies then pick one as a best gene set. Do: make millions of transcript assemblies using multiple kmers, programs and other options. Use all assemblies to pick best mRNA per locus. No single assembly is better for all loci than others, because of the large variation in expression levels, gene sizes, types of read errors, etc., that require different options and methods. Use multiple kmer sizes, from read-size down to 25.

Do: Use good RNA de-novo assemblers, even if genome assembly is available. Velvet/Oases, SoapDenovo-Trans, and Trinity produce good assemblies, in that order in my work, and each produces some best assemblies the others miss. Don’t: Use Cufflinks only for genome-mapped RNA. Cufflinks underperforms in assembly versus de-novo assemblers, and has more errors of commision (joins) and omission (missing).

Don’t: Select longest transcript as best mRNA, as this selects for errors in assembly. Many methods do this implicitly. Do: Select best mRNA with coding-sequence metrics (longest ORF, complete if possible). Longest proteins correlate strongly with strongest homology to other species.

Do: Use at least 200 Million reads of 100bp or longer, mate-paired, to get a complete transcriptome, from current Illimina sequencers. Don’t: Use longer 454 reads, due to high error rate. I've not tested very long reads of new machines, but their errors may cause similar problems.

Don’t: expect your species/data set to assemble in same way as others have reported. Don't rely on older software without testing newer, and don't expect newer versions to be better (but often they are).

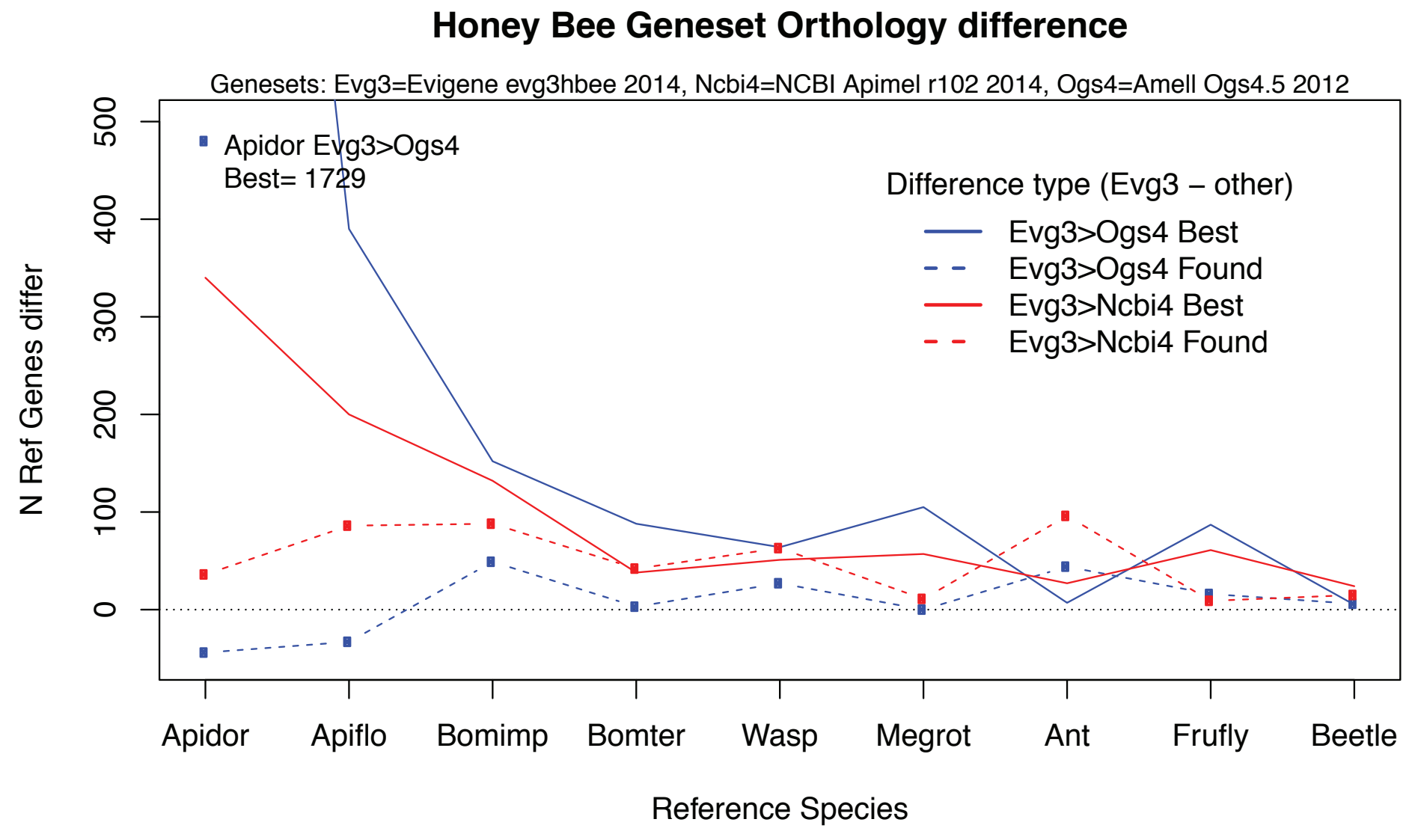
Do: use EvidentialGene scripts and methods for mRNA transcript assembly. The current EvidentialGene_trassembly_pipe is useable by others.
http://eugenes.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html

RNA assembly can outstrip computing resources. Large memory and cpu clusters are available as shared resources (NSF-XSEDE, others). Digital normalization and genome-mapped partitioning to assemble very large data sets in parts can help.

Apis mellifera Evigene set

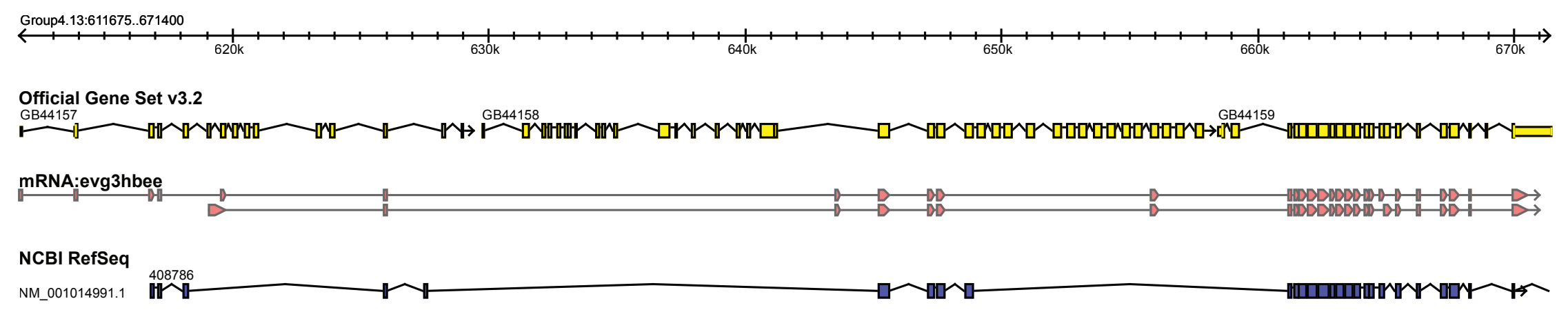
http://arthropods.eugenes.org/EvidentialGene/arthropods/honeybee/

```
#t2ac: EvidentialGene tr2aacds.pl VERSION 2014.05.25
#t2ac: bestorf_cds= evg3hbee.cds nrec= 6156631
#t2ac: nonredundant_cds= evg3hbeenr.cds nrec= 2257631
#t2ac: nofragments_cds= evg3hbeenrcd1.cds nrec= 1353185
Class Table for evg3hbee publicset
class       okay%   drop%  okay    drop
-----------------------------------------
althi       9.1     41     123465  549793
altmid      0.6     0.6    5297    12294   # paralogs here
altmfrag    0.4     0.4    2488    9534
main        4.3     4.7    59018   72091
noclass     1.0     9.2    13374   123991
fragments   0       27     0       375852
-----------------------------------------
total       15%    85%    203643  1149489  # ok = 5% of input.tr
=========================================
AA-quality for okay set of evg3hbee
okay.top  n=1000; average=2024; median=1725; min,max=1362,16948;
```


Honey Bee Geneset Orthology difference

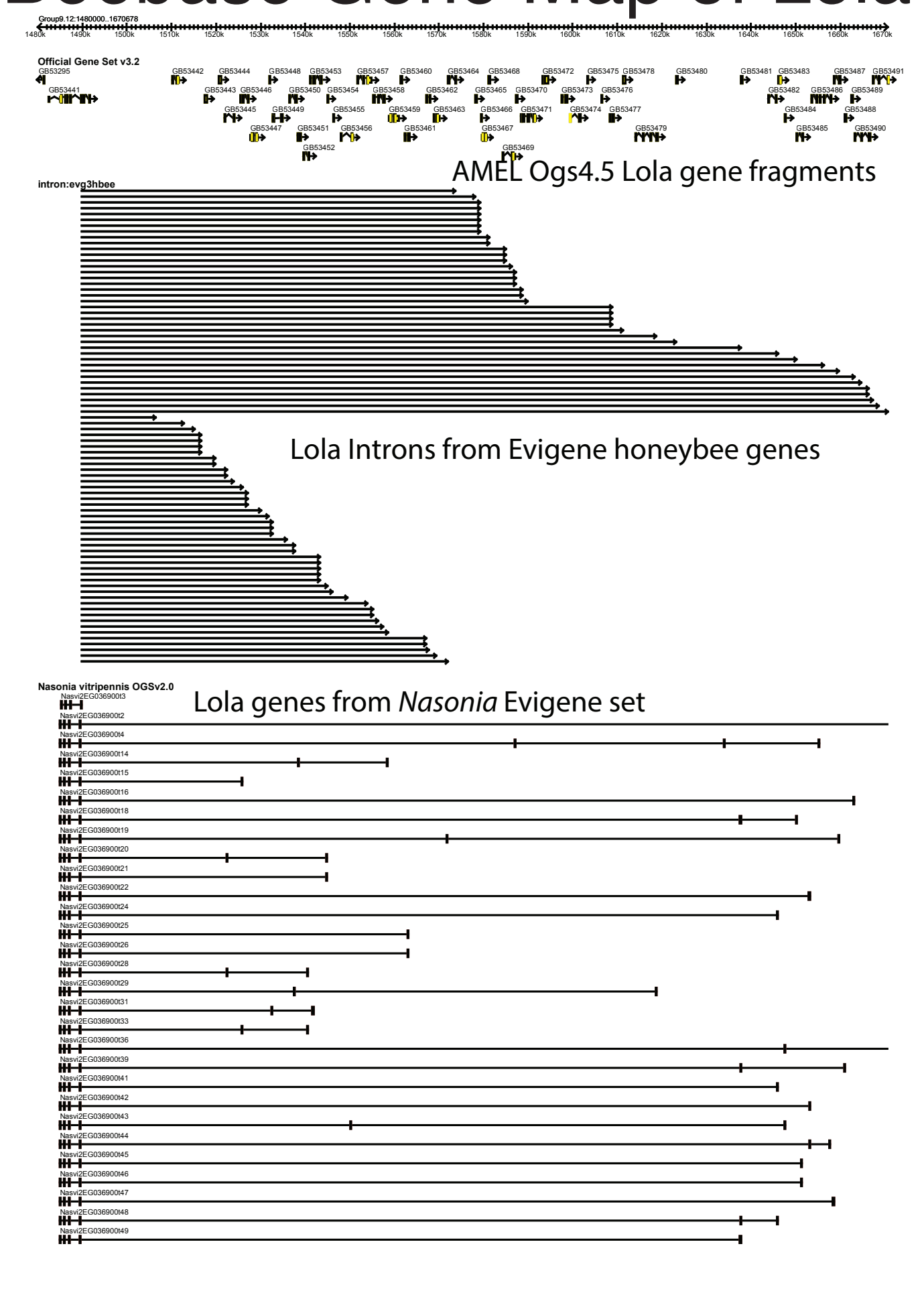
Genesets: Evg3=Evigene evg3hbee 2014, Ncbi4=NCBI Apimel r102 2014, Ogs4=Amell Ogs4.5 2012

Beebase Gene Map of DSCAM



DSCAM has 3 Ogs45 model fragments; 1 Evigene transcript with 1 of 100 alternates shown, and 1 NCBI Refseq model.

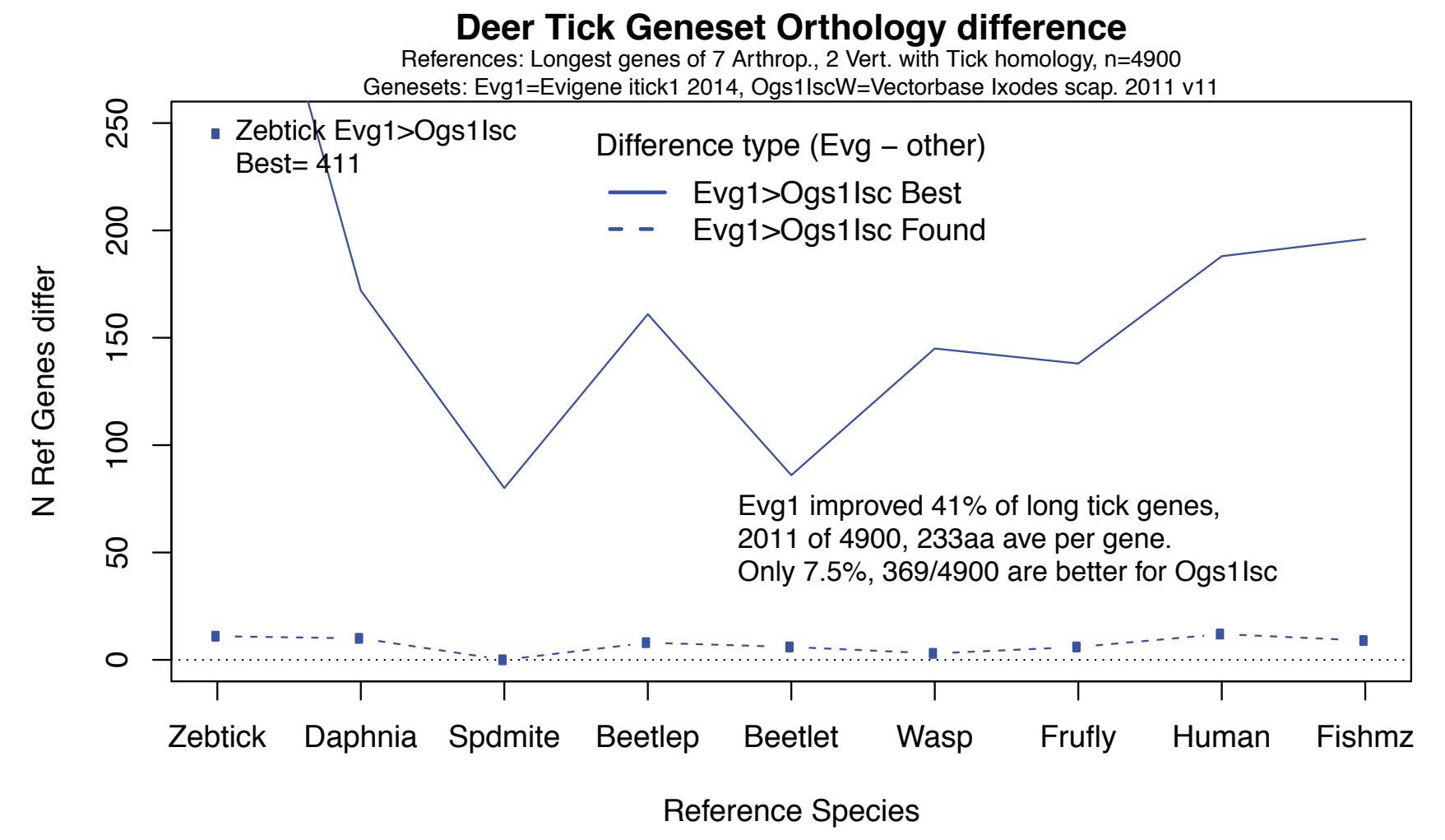
Beebase Gene Map of Lola (200Kb)


Official Gene Set v3.2
AMEL Ogs4.5 Lola gene fragments
intron:evg3hbee
Lola Introns from Evigene honeybee genes
Nasonia vitripennis OGSv2.0
Lola genes from Nasonia Evigene set

Ixodes scapularis Evigene set

http://arthropods.eugenes.org/EvidentialGene/arthropods/deertick/

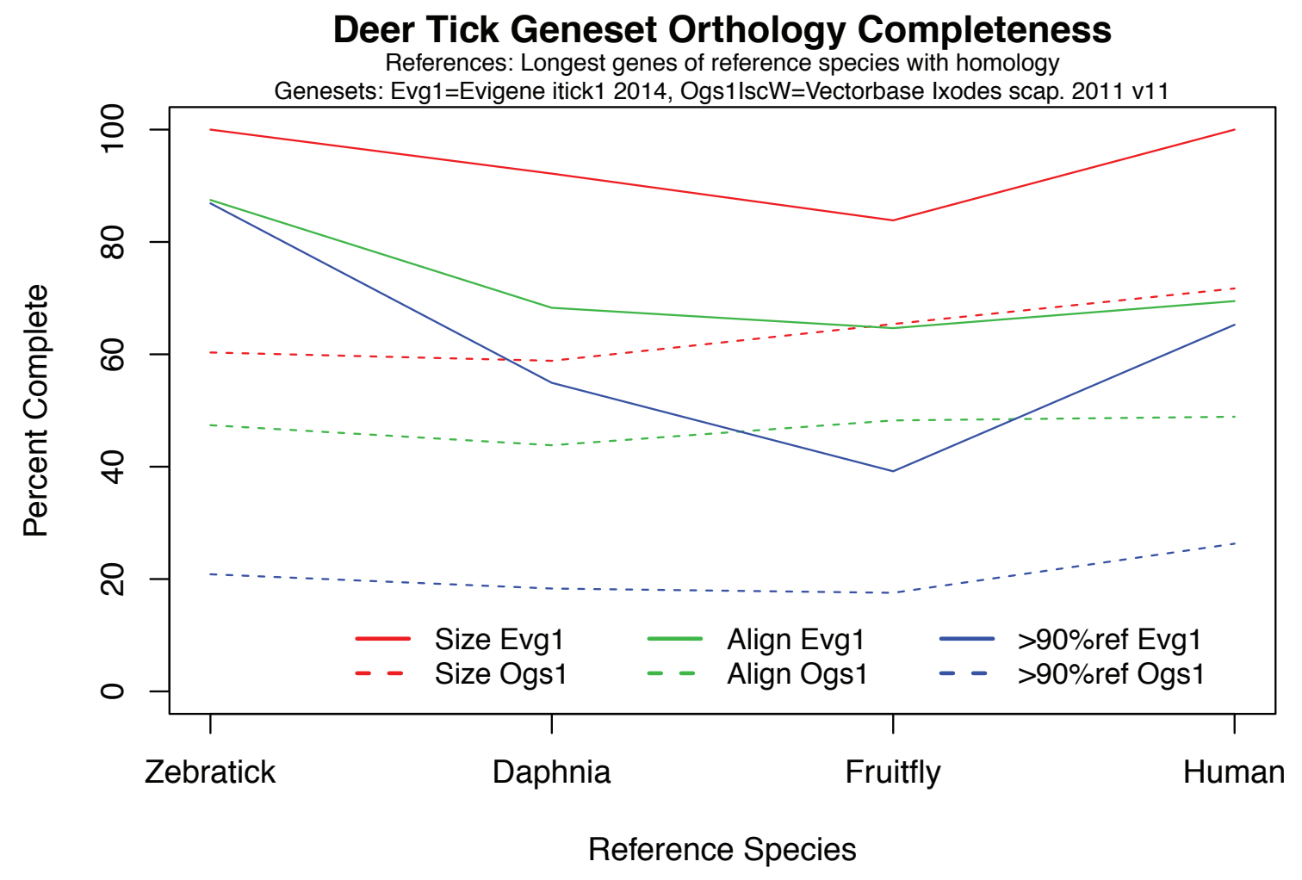
```
#t2ac: EvidentialGene tr2aacds.pl VERSION 2014.05.25
#t2ac: bestorf_cds= evg1itick.cds nrec= 5895165
#t2ac: nonredundant_cds= evg1iticknr.cds nrec= 2792966
#t2ac: nofragments_cds= evg1iticknrcd1.cds nrec= 1725680

Class Table for evg1itick publicset
class       okay%   okay
----------------------------
althi       4.5     78435
altmid      0.6     10610
altmfrag    0.2     3983
main        6.9     119588
noclass     1.7     29975
fragments   0       0
----------------------------
total       14%    242591  or 4% of input.tr
==================================
AA-quality for top 1000 of evg1itick okay set
average=1738; median=1464; min,max=1187,10671
```


Deer Tick Geneset Orthology difference

References: Longest genes of 7 Arthrop., 2 Vert. with Tick homology, n=4900

Genesets: Evg1=Evigene itick1 2014, Ogs1IscW=Vectorbase Ixodes scap. 2011 v11

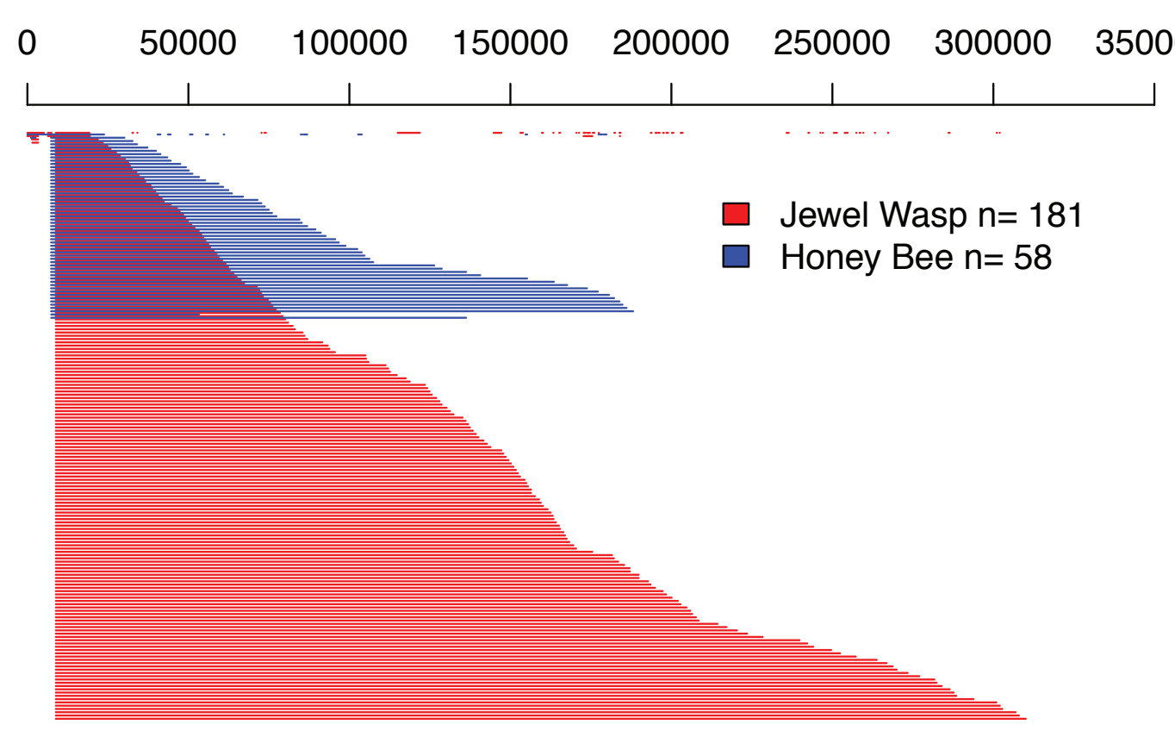
Evg1 improved 41% of long tick genes, 2011 of 4900, 233aa ave per gene. Only 7.5%, 369/4900 are better for Ogs1Isc


Deer Tick Geneset Orthology Completeness

References: Longest genes of reference species with homology

Genesets: Evg1=Evigene itick1 2014, Ogs1IscW=Vectorbase Ixodes scap. 2011 v11

Gene *lola* has most alternate introns in Apis (55+) and Nasonia (180+)



If you study bee/wasp/ant social behavior, lola alternates guide brain axon growth:
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC218142/

Daphnia magna Evigene set

http://arthropods.eugenes.org/EvidentialGene/daphnia/daphnia_magna/

```
#t2ac: EvidentialGene tr2aacds.pl VERSION 2014.05.25
#t2ac: bestorf_cds= dmagset56tx.cds nrec= 9843887
#t2ac: asmdupfilter_cds= dmagset56tx.trclass
# Class Table for dmagset56tx.trclass
class           okay    drop    okay    drop
--------------------------------------------
althi           3.6     11.9    65845   216558
althi1          7.4     24.4    134815  441933
altmfrag        0.6     0.6     12317   11065
altmid          0.5     0.2     9073    5359   # paralogs here
main            1.9     2       34402   37263
noclass         0.3     3.4     7207    63161
parthi          0       26.5    13      479367
parthi1         0       12.3    0       222455
--------------------------------------------
total           14.7    85.2    266330  1541950
============================================
# AA-quality for okay set of dmagset56tx.aa.qual (no okalt):
okay.top  n=1000; average=1935; median=1607; min,max=1229,22277;
```

Daphnia magna gene set is delayed in public release. These water fleas are the problem child of genome informatics, with their genomes densely packed with genes, many tandem gene duplicates, that are hard to resolve accurately with any/all available methods. Watch above URL, due real soon now..

Take home message: Gene-ome construction is solved.

mRNA-assembly of genes, done properly, produces more complete gene sets for species, with no genome assembly, nor gene predicting, nor other species genes to inform and add mistakes. Adding genome-sequence methods, taking errors of both into account, will improve on mRNA-assembly.

Done properly is not yet common practice: Use several de-novo assemblers, with multi-kmers, digital-normalization and re-assembly of unassembled, low expression genes. Use Velvet/Oases (best), SoapDenovoTrans (2nd best and quick), Trans-Abyss (also good), Trinity (4th place) and/or Cufflinks (many more mistakes).

All assemblers together, with EvidentialGene classifying the best mRNA genes, is what you want. Use protein orthology, coding sequence quality metrics to classify best gene sets, as EvidentialGene does. The common practice of using transcript size, N50 and genome metrics is useless for genes, as it rewards mistakes and has little biological relevance. Projects for gene discovery should be using such methods now. Gene assembly, with mRNA-seq and genome sequencing, is a solved engineering problem, and can be done with reliability, accuracy and completeness above the 95% level.

EvidentialGene mRNA-assembly pipeline

EvidentialGene tr2aacds.pl is my new, somewhat easy to use pipeline for processing large piles of transcript assemblies into a biologically useful "best" set of mRNA, classified into primary and alternate transcripts.

Classification is based primarily on CDS-dna local alignment identity. Transcripts at one locus share exon-sized or larger identities. Perfect fragment CDS are dropped, those with some CDS base differences are kept, with longest CDS as primary transcript. UTR identity is ignored (for now) because many of the mis-assemblies are from joined/mangled genes in UTR region.

Algorithm of tr2aacds:

0. collect input transcripts.tr, produce CDS and AA sequences, work mostly on CDS.
1. perfect redundant removal with **fastanrdb**
2. perfect fragment removal with **cd-hit-est**
3. **blastn**, basic local align high-identity subsequences for alternate tr.
4. classify main/alternate cds, okay & drop subsets by CDS-align, protein metrics.
5. output sequence sets from classifier: okay-main, okay-alts, drops.

See http://eugenes.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html

Other Evigene scripts for mRNA assembly
evigene/scripts/rnaseq/trformat.pl : regularize and unique IDs in transcript.fasta, adding prefixes for parameter sets.
evigene/prot/namegenes.pl : add gene function names from UniProt and Conserved Domains (CDD) with delta-blastp.
evigene/scripts/rnaseq/**asmrna_trimvec.pl** : process NCBI vector screen, and trim end gaps in transcripts.
evigene/scripts/evgmrna2tsa.pl : check mRNA, add annotation, create public IDs and sequence files, write Genbank TSA format for public submission.

Genome collaborators and data providers: Daphnia Genome Consortium, Generic Model Organism Database, Indiana Univ. Center for Genomics & Bioinformatics, International Aphid Genomics Consortium, Nasonia jewel wasp Genome project, Cacao chocolate tree Genome project. *Funding*: NSF mostly, NIH, Mars. *Computers*: TeraGrid/XSEDE, NCGAS