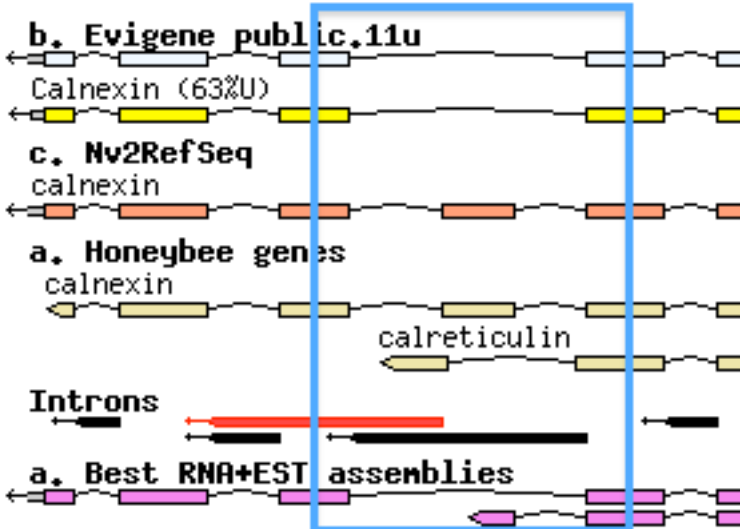
Perfect Arthropod Genes Constructed from Gigabases of RNA



Don Gilbert May/June 2012

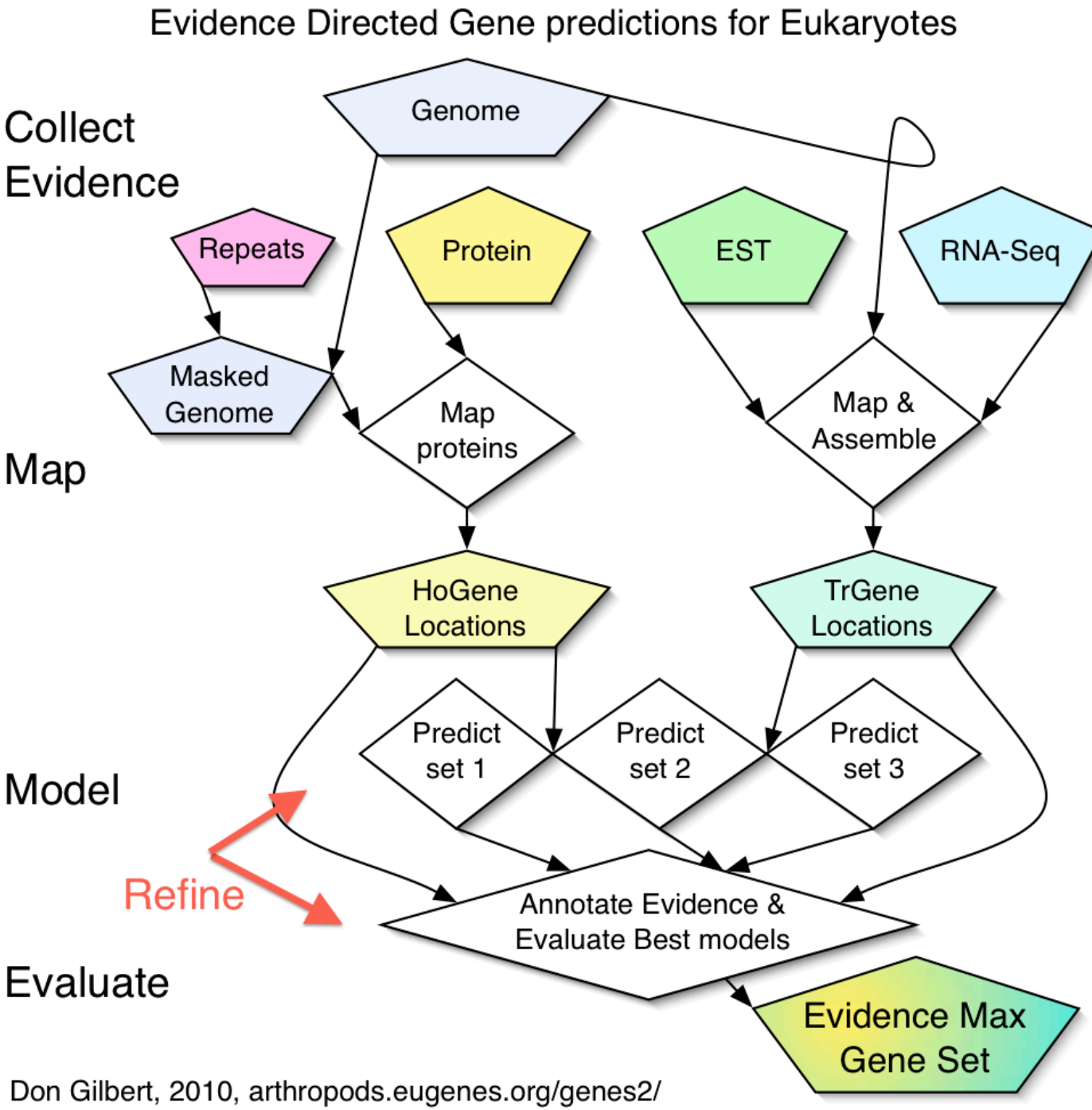
Biology Dept., Indiana University
gilbertd@indiana.edu

Gene construction, not prediction

- The decade of gene prediction is over; genes constructed with transcript sequence *can* surpass predictions for biological validity.
- *".. over half the gene predictions were imperfect, with missing exons, false exons, wrong intron ends, fused and fragmented genes"* w/r/t © 2006 gene set.
- *but..* Gene assembly from RNA has similar problems.
- Perfecting this means using all of best data and tools, plus quality tests, to build accurate genes.

EvidentialGene

- No gene set is best at all loci, alternate sets are useful
- Tries to match expert choices
- Deterministic evd. scoring, not majority vote
- Same result for 1 locus or 50,000
- Can update 100 w/ new evidence

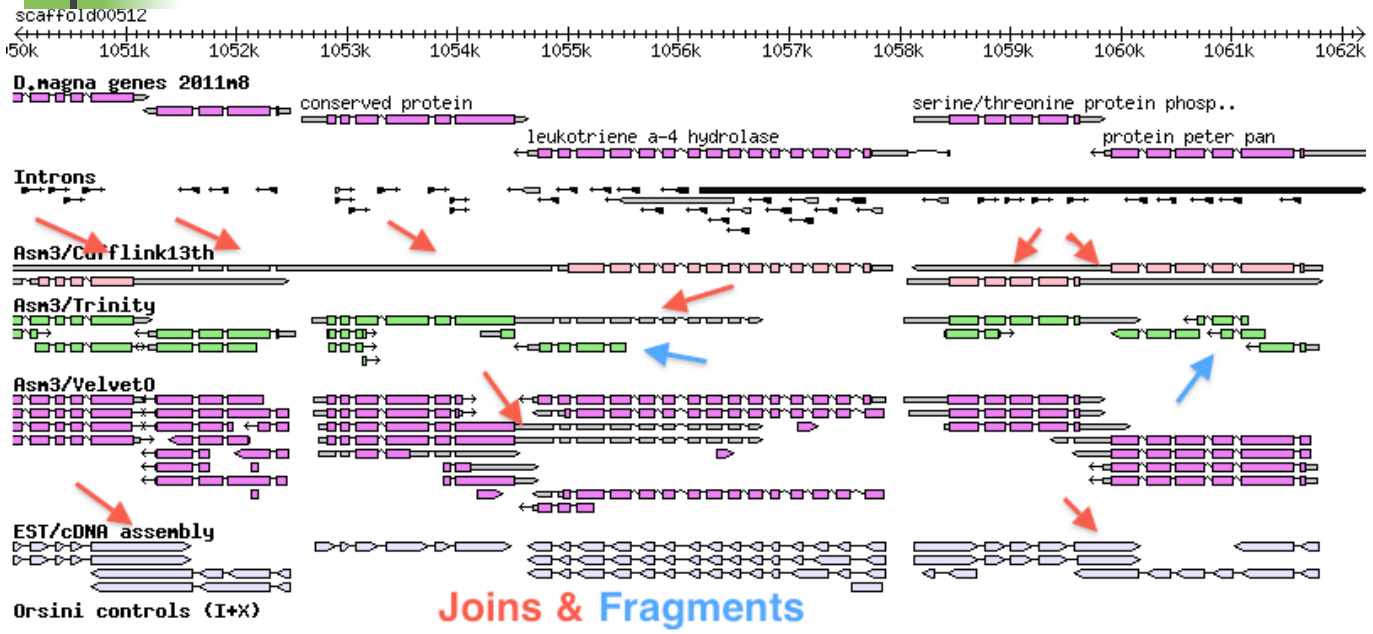


Too much data or not enough?

- Transcript assemblies can be more accurate than predictions, but not at 90+% of loci. Effort is needed to perfect them.
- RNA data quality sets limits, imperfect software struggles at both ends of the data river.
- Data reduction a major task: 10^9 RNA reads assemble to 10^6 competing models, selecting $10^{4.5}$ biological genes.

- ✓ 1 Billion short reads, from many tissues/time/environs, not 50 Million, may be enough
- ✓ Mate paired with staggered inserts (200 – 600 bp); strand specific helps.
- ✓ Long (454) + Short (Illumina) better, both insert paired

RNA assembly good, bad



Too much data &/or tool problems

Daphnia magna RNA assemblies

Evidence evaluation (in part)

- CDS/exon ratio, UTR exons *Nomap*
- Protein homology (bitscore, identity; *Nm*)
- RNA read cover, uncover spots (fusions? *Nm*)
- Read intron match (*Map*)
- RNA assembly / reference equivalence

Evidence scores for gene sets

Method	CDS	Homlg	EST.cov	UTR.ok	Intron
Genes2011	62%	565	57%	80%	64%
Velvet/O	73%	577	72%	89%	56%
Trinity	71%	565	71%	88%	58%
Cufflinks13	45%	498	65%	59%	47%

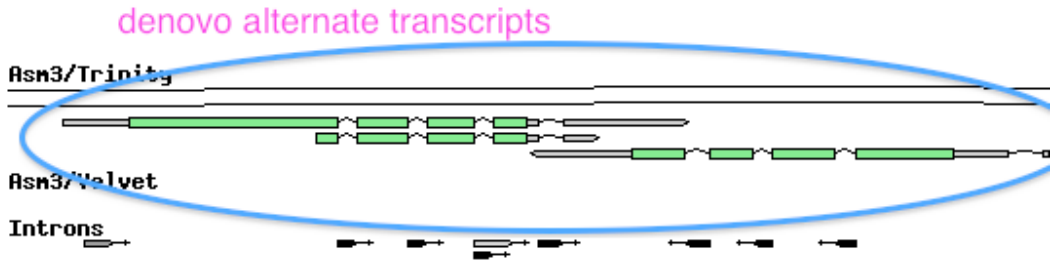
Genes without genomes?

Yes. Locust gene set is assembled without a genome. Orthology gene family score is higher for locust than for insects with genome-map genes (for Velvet assembly, lower for Trinity).

Gene set	Bits	Δ Size
Daphnia	502	3
Locust.Vel	482	-20
Beetle	475	16
Wasp	470	28
Locust.Trin	452	-87
Fruitfly	447	89

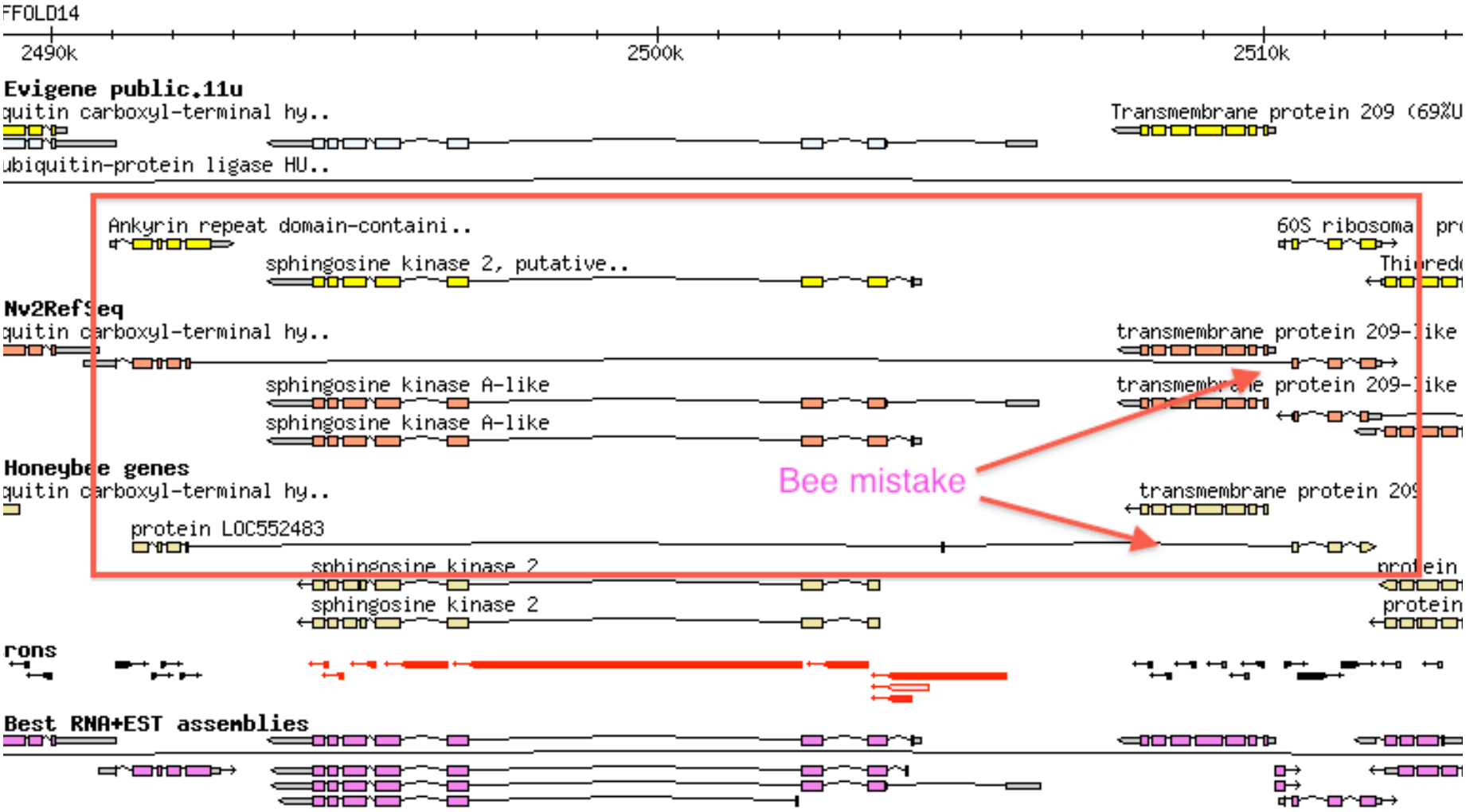
But..

- Paralogs, alternates, bad guesses are resolved with a genome.
- Contaminants don't map to genome. E.g. *mouse* genes in arthropod reads.
- Best gene models match gene structure signals on genome.

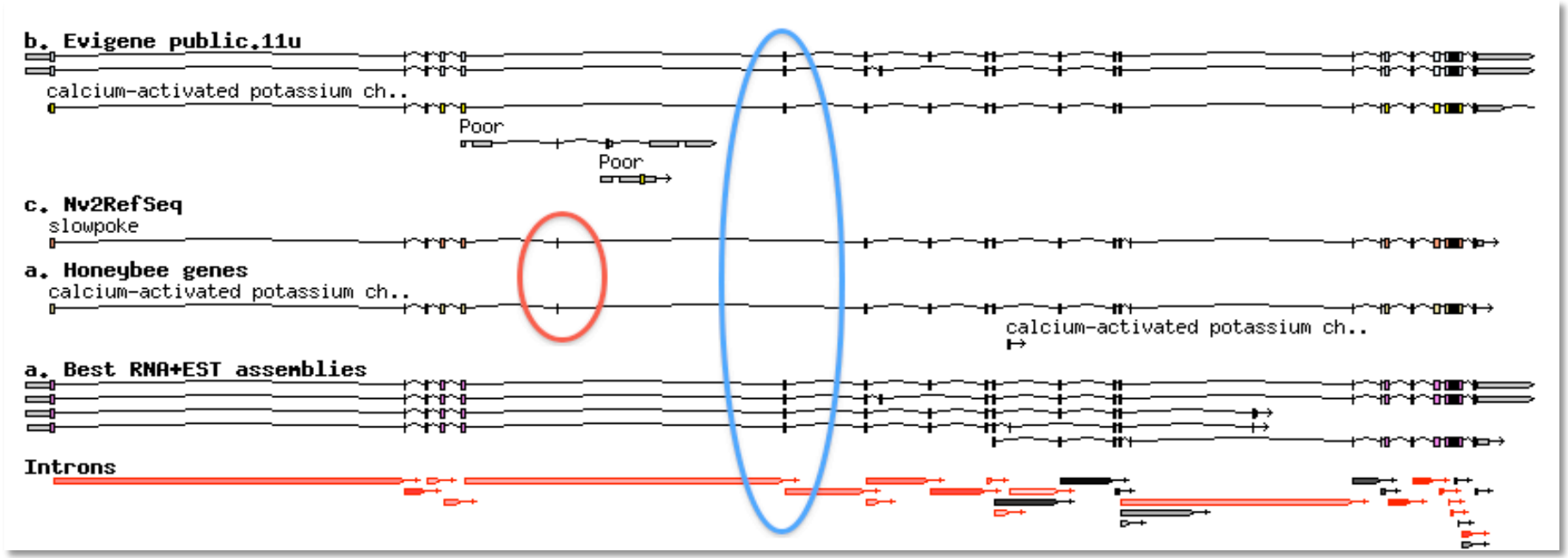


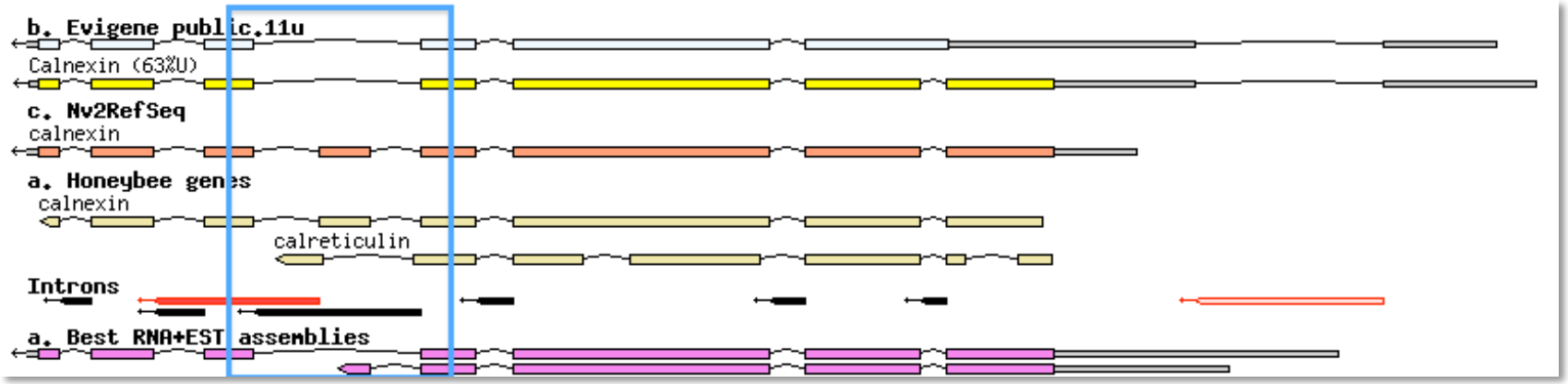
Both ways is better (genomes have holes).

Is that a *honey bee* gene in your *wasp* genome? Mistakes can be transferred



Is that a *honey bee* gene in your *wasp* genome? Exon changes are common





Y B Purrfekt ?

How many gene studies have artifacts of quality?

"Genome annotation emerged as the largest single influencer, affecting up to 30%" of discrepancies in orthology assessments [1] . Gene function studies, differential expression, *etc*. want perfect genes.

Assess current tools with species and data sets.

RNA/gene software is changing rapidly, sometimes not for better.

Several tools combine to give better answers.

EvidentialGene results are not perfect, yet.

But this approach appears to be working. A major remaining need is that tuning out problem cases is not automated.

Expert inspection combined with evidence rescoring reduces such errors, but the last 10% require effort similar to the first 90%.

1. Trachana .. and Bork. 2011. Orthology prediction methods: A quality assessment using curated protein families. Bioessays 33: 769–780.

EvidentialGene Results

Acyrthosiphon pea aphid, 2011 June

Evidence	Evigene	RefSeq2	ACYPI v1
Introns	70%	68%	52%
EST coverage	79%	69%	49%
RNA assembly	49%	43%	27%
Protein score	76%	46%	47%

Nasonia jewel wasp, 2012 Jan

Evidence	Evigene	RefSeq2	OGS v1.2
Introns	97%	90%	85%
EST coverage	72%	67%	51%
RNA assembly	63%	36%	29%
Homology bits	679	635	--

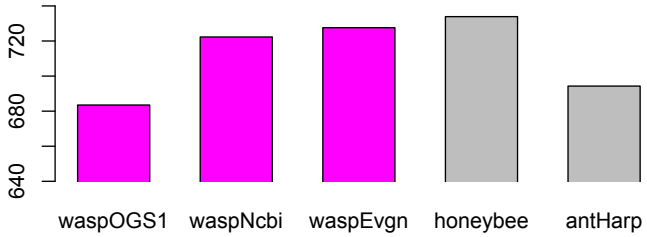
Introns: match to EST/RNA spliced introns
EST coverage: overlap with EST exons
RNA assembly: equivalence to RNA assemblies

Gene sets for pea aphid and jewel wasp are superior on several evidence scores to those of NCBI RefSeq, built with the same available evidence. Evigene results with *Daphnia* and *The. cacao* also improve their genes.

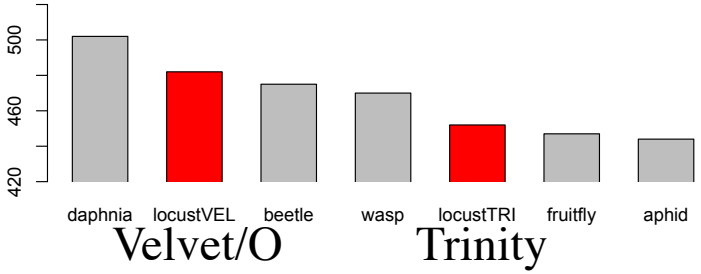
arthropods.eugenes.org/EvidentialGene/

Gene set quality *vs* Orthology rank

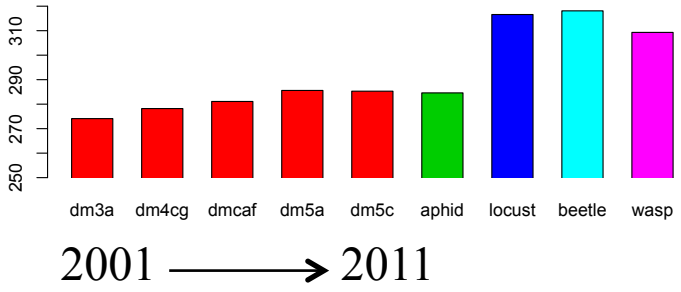
Wasp moves up to Bee



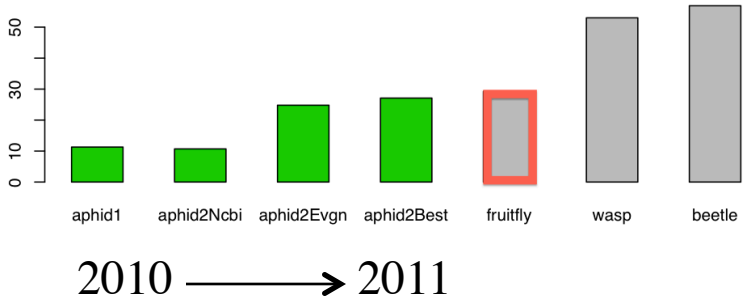
RNA-assembler changes
Locust species rank



Fruitfly genes improve over 10 years



Aphid matches Fruitfly in 1 year



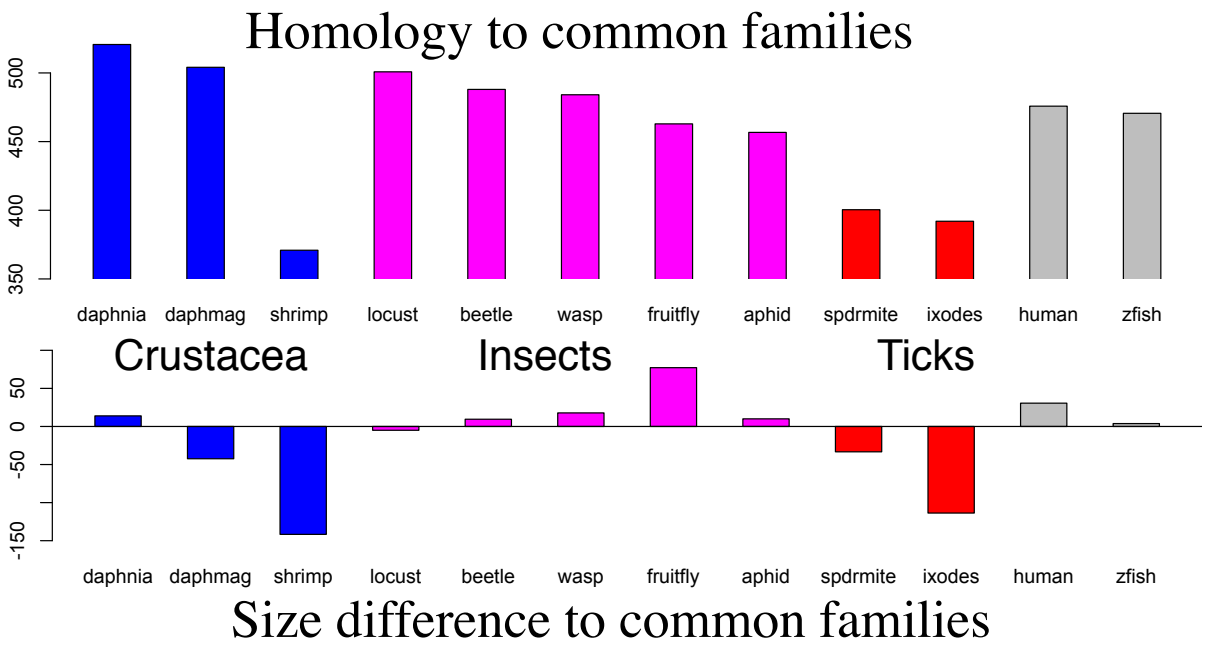
Arthropod Genes Summary

Clade presence of gene families

Clade	Only	Miss	OutOnly	OutMiss
Crustacea	101	580	144	213
Ticks	64	1171	69	471
Insects	519	1683	157	340

OutOnly= 2+ species in clade have outgroup family, not in other clades.
OutMiss= none in clade have outgroup, both other clades have.
Only = all species in clade have family, none of other clades have
Miss = no species in clade has family, both other clades have

But, gene set qualities confound gene family presence



End note

gilbertd@indiana.edu

Genome collaborators and data providers

Daphnia Genome Consortium
Generic Model Organism Database
International Aphid Genomics Consortium
Nasonia Genome project
Cacao Genome project
Indiana U Ctr. Genomics & Bioinformatics

Funding: NSF mostly, NIH, Mars
Computers: TeraGrid/XSEDE, NCGAS

Links to this work

arthropods.eugenes.org/	14+ Bug genomes
arthropods.eugenes.org/EvidentialGene/	Perfecting bug genes
wfleabase.org	Daphnia genomics
www.bio.net	Arthropod news/discussion list

Tools for Gene Building

Augustus to model genes, with mapped EST/RNA and proteins; make many prediction sets from data slices. Other predictors as desired (fgenesh, Gnomon, ...)

Exonerate for protein gene mapping.

GMAP-GSNAP for read mapping RNA/EST.

Velvet/Oases for RNA/EST assembly (de novo) .

Trinity for RNA/EST assembly (de novo) .

Cufflinks for RNA/EST assembly (genome mapped) .

NCBI BLAST locate proteins, annotate genes.

OrthoMCL group gene families and homologs.

Evigene combiner and support scripts, best gene models for evidence @ arthropods.eugenes.org/EvidentialGene/

Continually evaluate/replace software with best of breed.

Perfect Arthropod Genes

- Gen 2 genome informatics
  - Gene ~~prediction~~ **construction** recipe
  - Wrestling with RNA-Seq
  - Software lags behind data
- Perfect genes for Aphid, Daphnia, Wasp, ..

  Augustus gene models + RNA assembly

  + Protein orthology + Details

  = much improved gene sets
- Daphnia magna genes and expression

EvidentialGene Recipe

Evidence annotation and maximization.

- Deterministic evidence scoring (same for 1 locus or 50,000).
- *Not* majority vote, single best scoring model wins
- Attempts to match expert curator choices

Basic steps

1. produce several predictions and transcript assembly sets with quality models.
   No single method/set is best at all loci, variants often have best among them.
2. Annotate models with all evidence, esp. gene model qualities
   (transcript introns, exons, homology, transposons, ...)
3. Score models from weighted sum of evidence.
4. Remove models below minimum evidence score
5. Select from overlapped models/locus the highest score, include fusion metrics
   (**longest is not always best**)
7. Evaluate results, genome-wide averages and with inspection (map views of errors)
8. Iterate 3..7 with alternate scoring to refine final best set.

Is that a *honey bee* gene in your *wasp* genome?

Wasp models