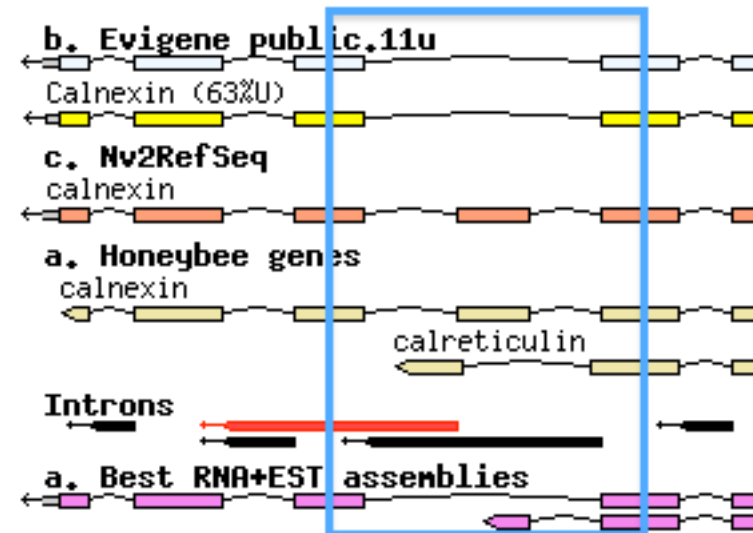


Perfect~ Arthropod Genes Constructed from Gigabases of RNA



Don Gilbert

May/June 2012

Biology Dept., Indiana University

gilbertd@indiana.edu



Perfect Arthropod Genes

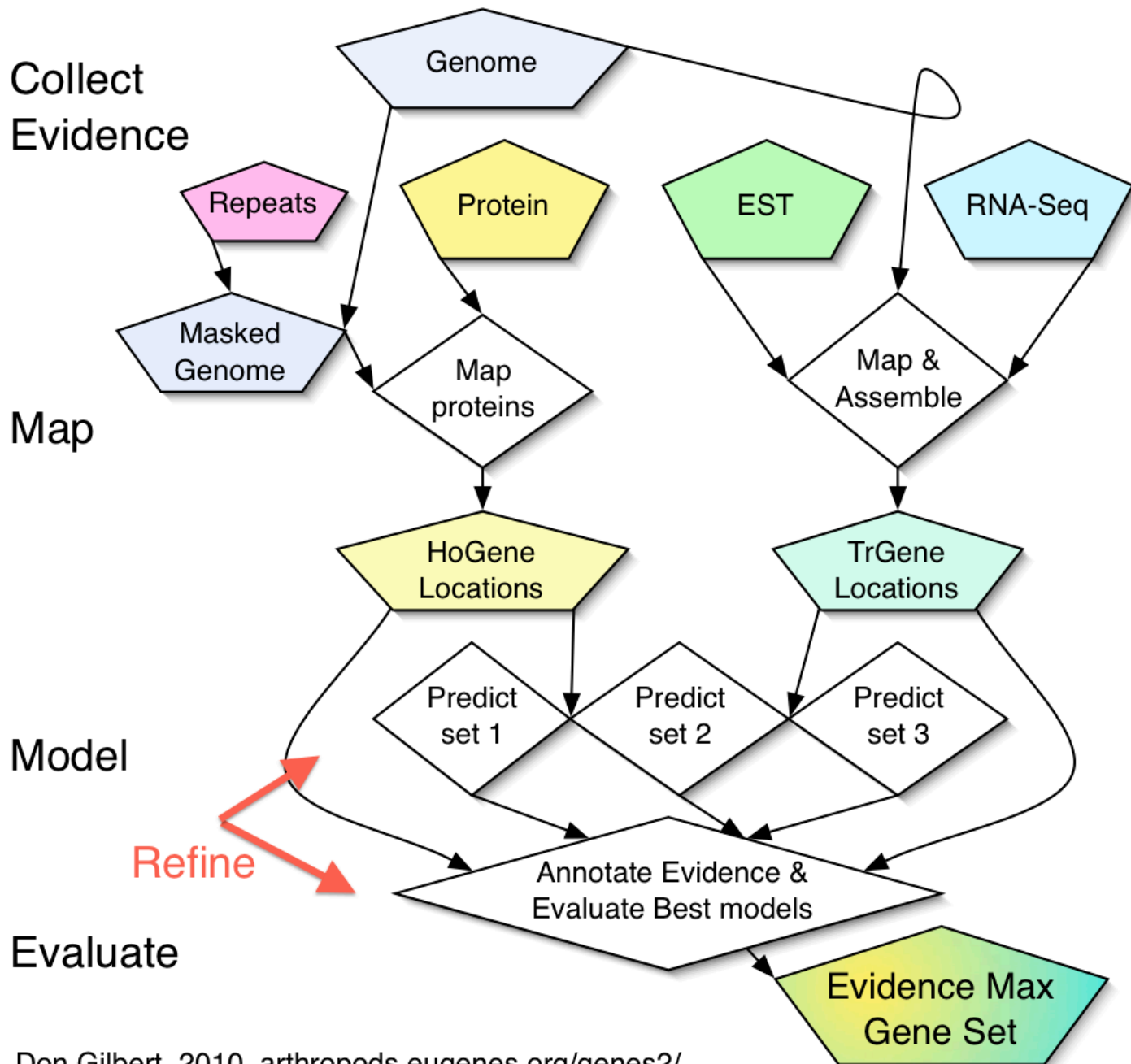
- Gen 2 genome informatics
 - Gene ~~prediction~~ **construction** recipe
 - Wrestling with RNA-Seq
 - Software lags behind data
- Perfect genes for Aphid, Daphnia, Wasp, ..
 - Augustus gene models + RNA assembly
 - + Protein orthology + Details
 - = much improved gene sets
- Daphnia magna genes and expression



Gene construction, not prediction

- The decade of gene prediction is over; gene construction with transcript sequence surpasses predictions for biological validity.
- To paraphrase others: “.. *over half the gene predictions were imperfect, with missing exons, false exons, wrong intron ends, fused and fragmented genes*”.
- Gene assembly from RNA has similar problems.
- Perfecting this means using all (best) data and tools, plus quality tests, to build accurate genes.

Evidence Directed Gene predictions for Eukaryotes





Evidential Genes 2.0

Pea aphid v2, 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 v2, 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



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.



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.

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

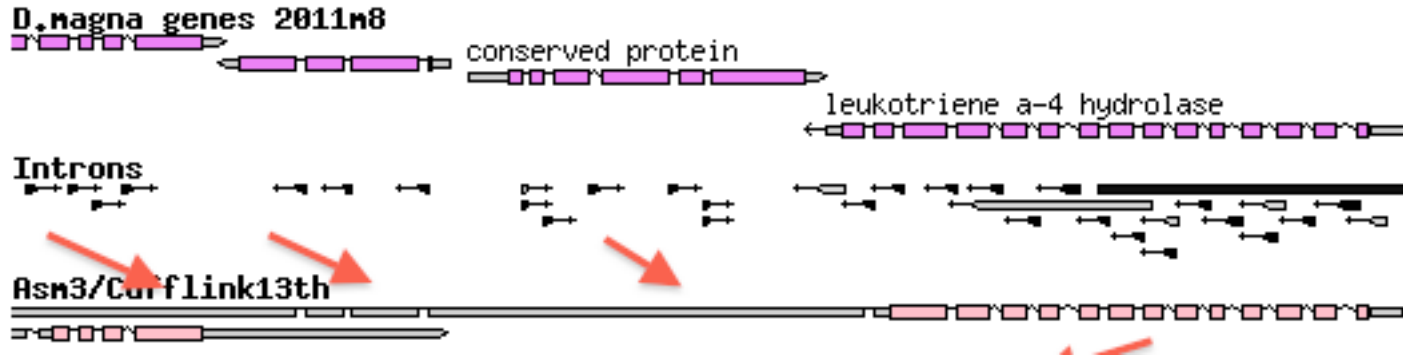
Continually evaluate/replace software with best of breed.



Too much data or not enough?

- Transcript assemblies can be more accurate than predictions, but effortful to resolve conflicts.
- RNA data quality sets limits, 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, not 50 Million, may be enough
- ✓ Mate paired with staggered inserts (200 – 600 bp); strand specific helpful.
- ✓ Long (454) + Short (Illumina) better, both insert paired

RNA assembly good, bad



Too much data &/or tool problems

Coding quality (subset of genes)

Method	Ngene	%CDS	wCDS	wUTR	cds<33%
VelvetO	5589	73	1683	605	2.8%
Trinity	7709	71	1599	653	8.4%
Cufflk13	5475	45	1641	1995	25.7%

EST coverage

Method	Ngene	Accur	Compl	UTRoff
Genes2011	28561	94	57	20
VelvetO	32298	92	72	11
Trinity	37340	92	71	12
Cufflnk13t	9830	95	65	41

Best homology (subset)

Method	N.Grp	%Grps	Bits
same2,3	5404	63%	750
VelvetO	1414	17%	704
Trinity	1364	16%	706
Cufflk13	321	3%	822

Daphnia magna
RNA assemblies

Introns valid

Method	Nintron	Valid%
Genes2011	115267	64%
Trinity	74575	58%
VelvetO	72153	56%
Cufflnk13t	61209	47%

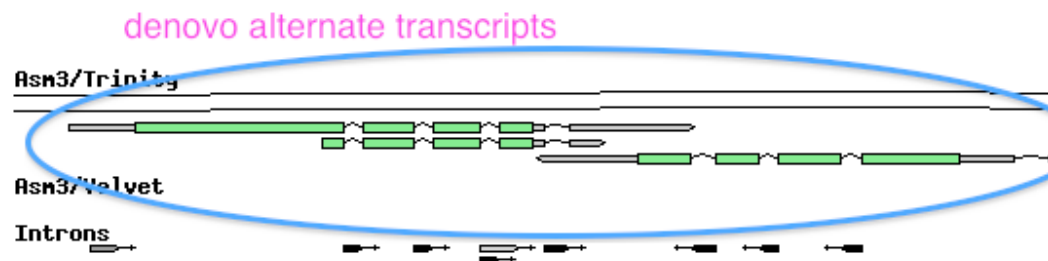
Genes without genomes?

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

Gene set	Bits	Δ Size
Daphnia	502	3
<i>Locust.vel</i>	482	-20
Beetle	475	16
Wasp	470	28
<i>Locust.trin</i>	452	-87
Fruitfly	447	89

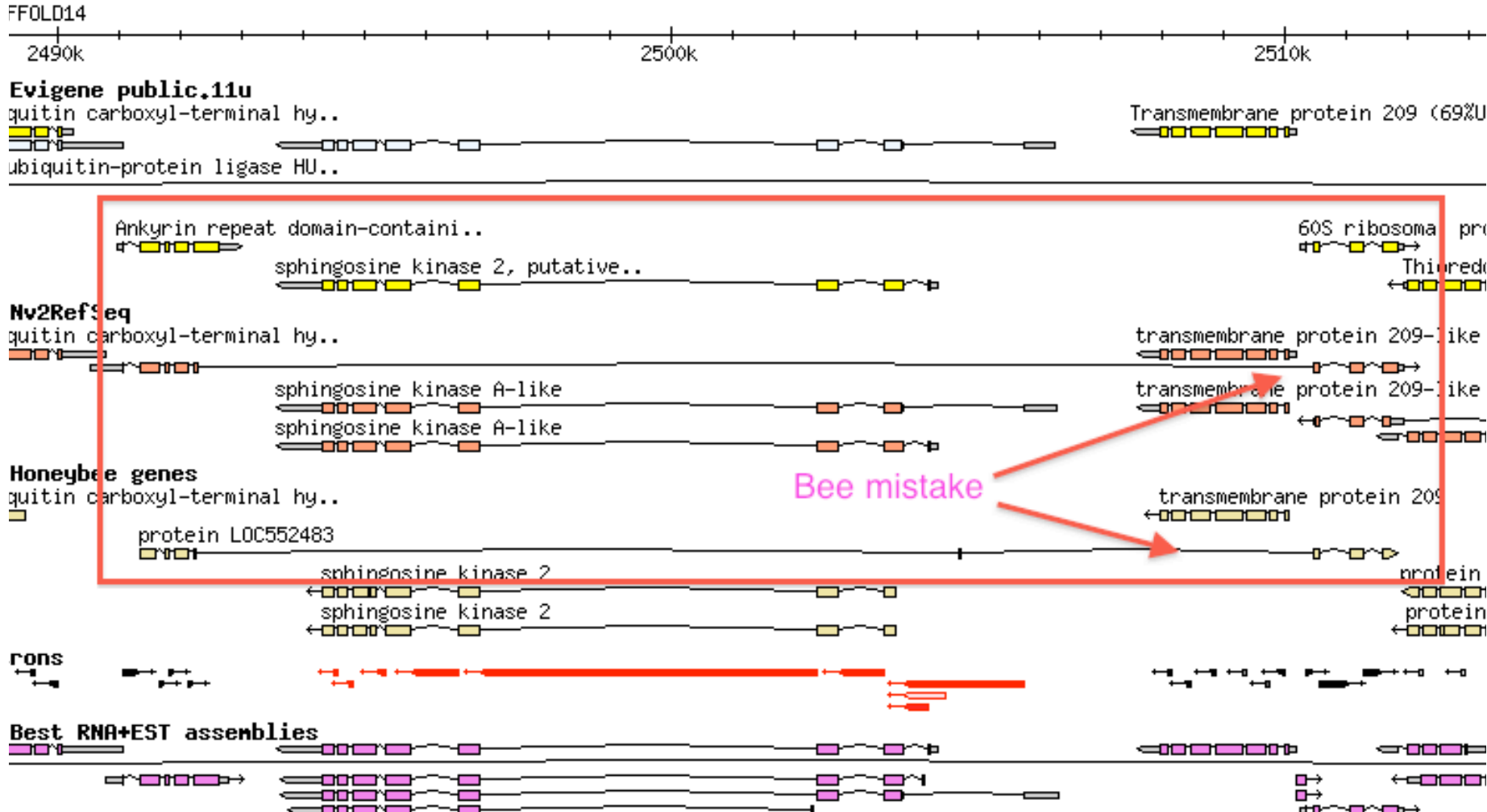
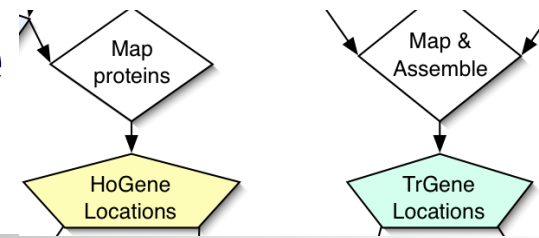
But..

- Alternates, paralogs, bad guesses are resolved with a genome.
- Contaminants don't map to genome. E.g. *mouse* genes in 2 sets of arthropod reads.
- Best gene assembly uses gene structure signals from genome.



But, both ways is better.

Is that a honey bee gene in your wasp genome?



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





Daphnia magna workshop

Don Gilbert

May/June 2012

Biology Dept., Indiana University
gilbertd@indiana.edu

Daphnia magna genes

Genes

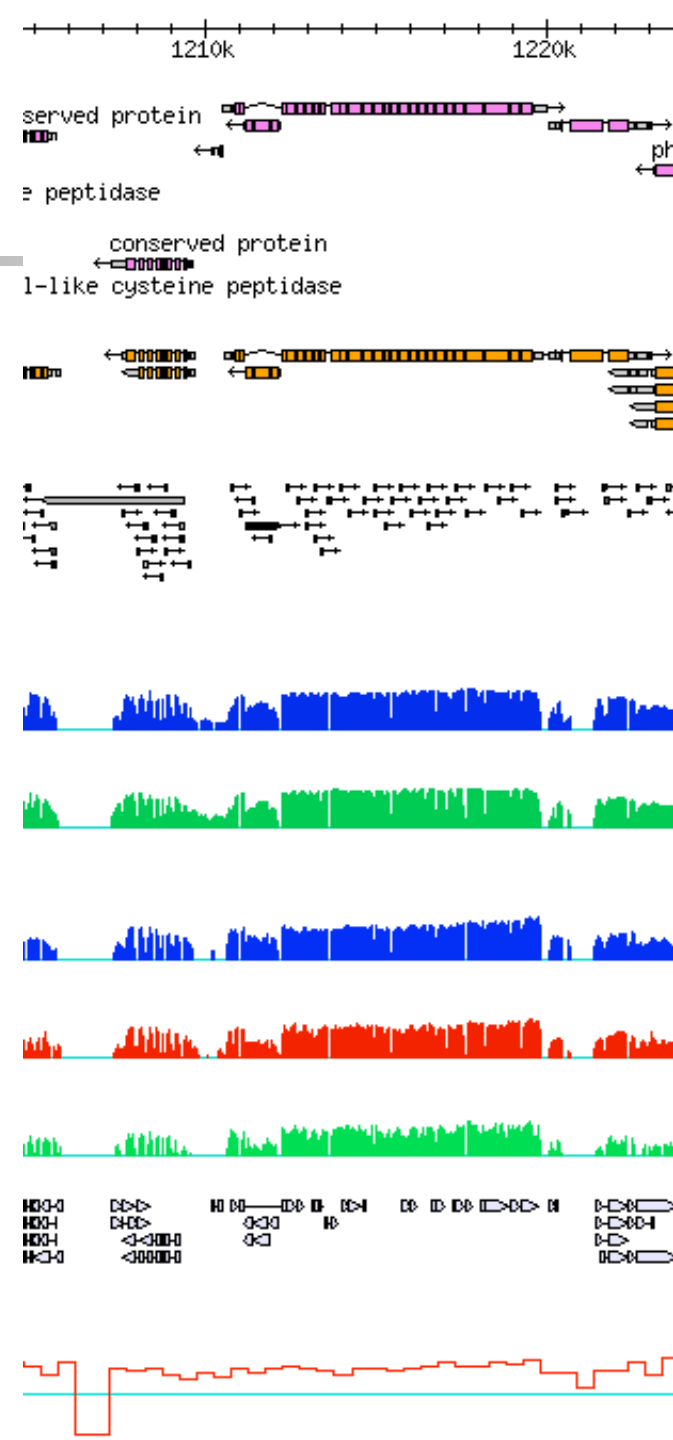
wfleabase.org/genome/Daphnia_magna/prerelease/

Genome maps

server7.wfleabase.org:8091/gbrowse/cgi-bin/gbrowse/daphnia_magna2/

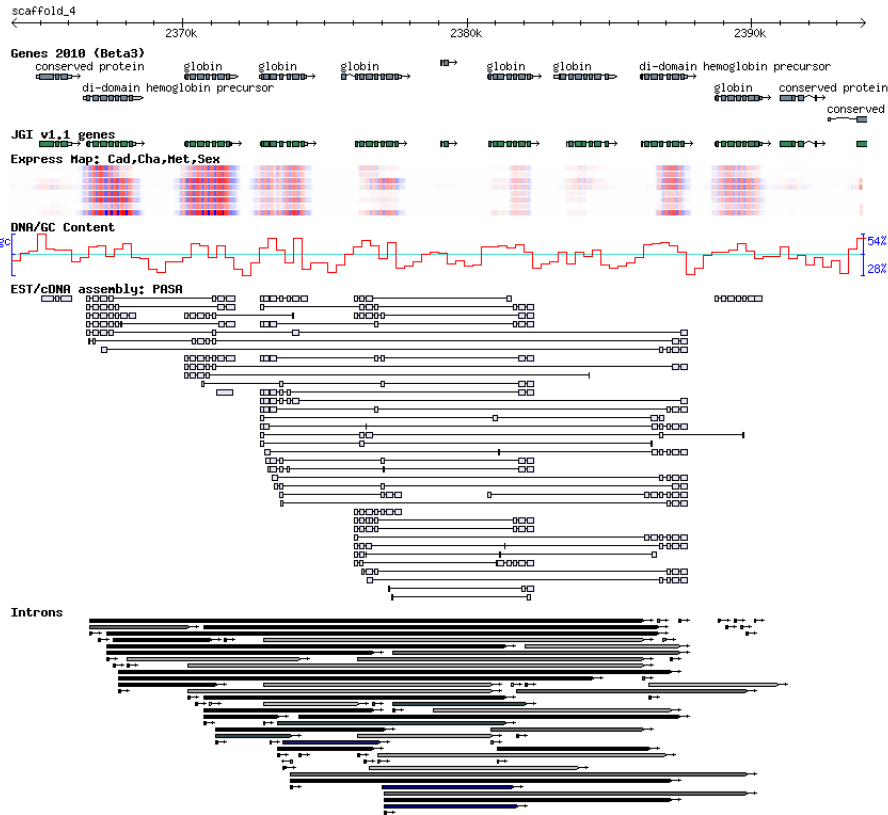
2012 draft gene models (newer rna but not newest rna)
gene-predictions/daphmagna_201205/

Differential expression for StressFlea RNA on 2012 draft genes
gene-predictions/daphmagna_201205/de_mag3mtv3/
counts of read/transcript, includes unmapped genes
edgeR rough-draft DE stats from these counts

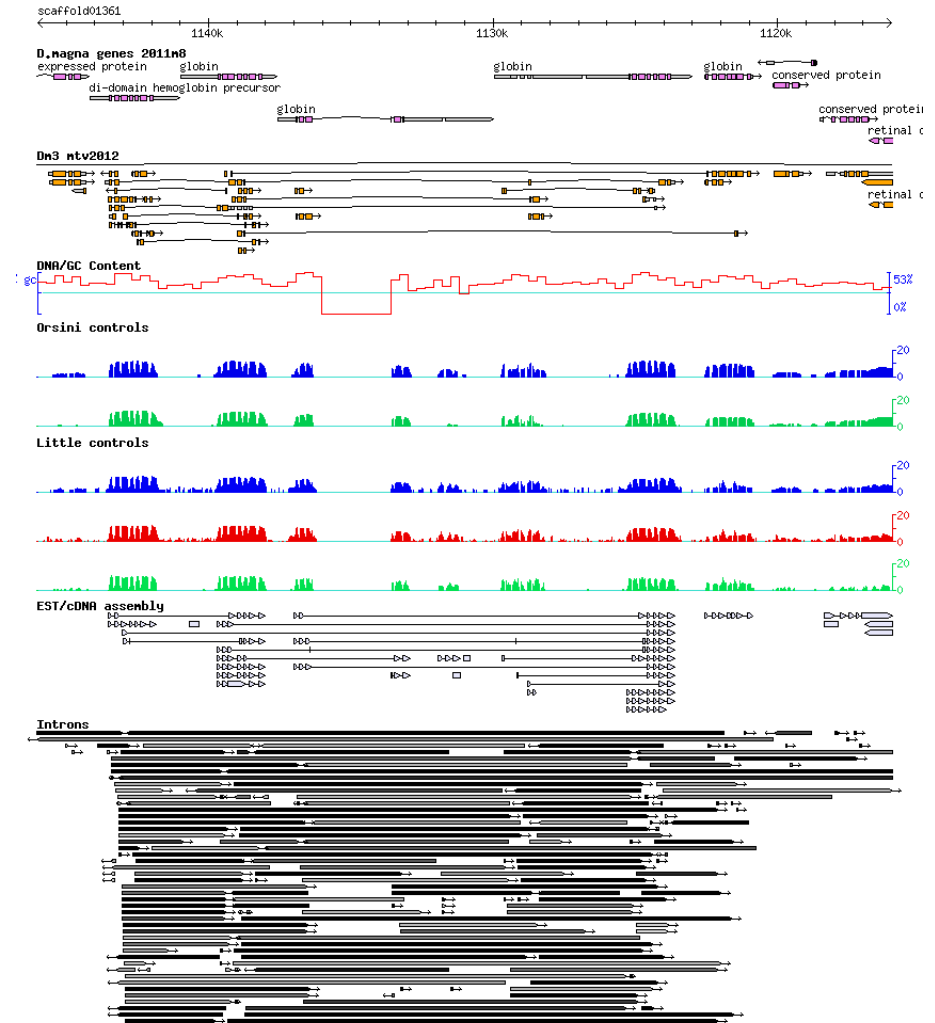


Genes are unfinished

Dap pulex Hemoglobin-8



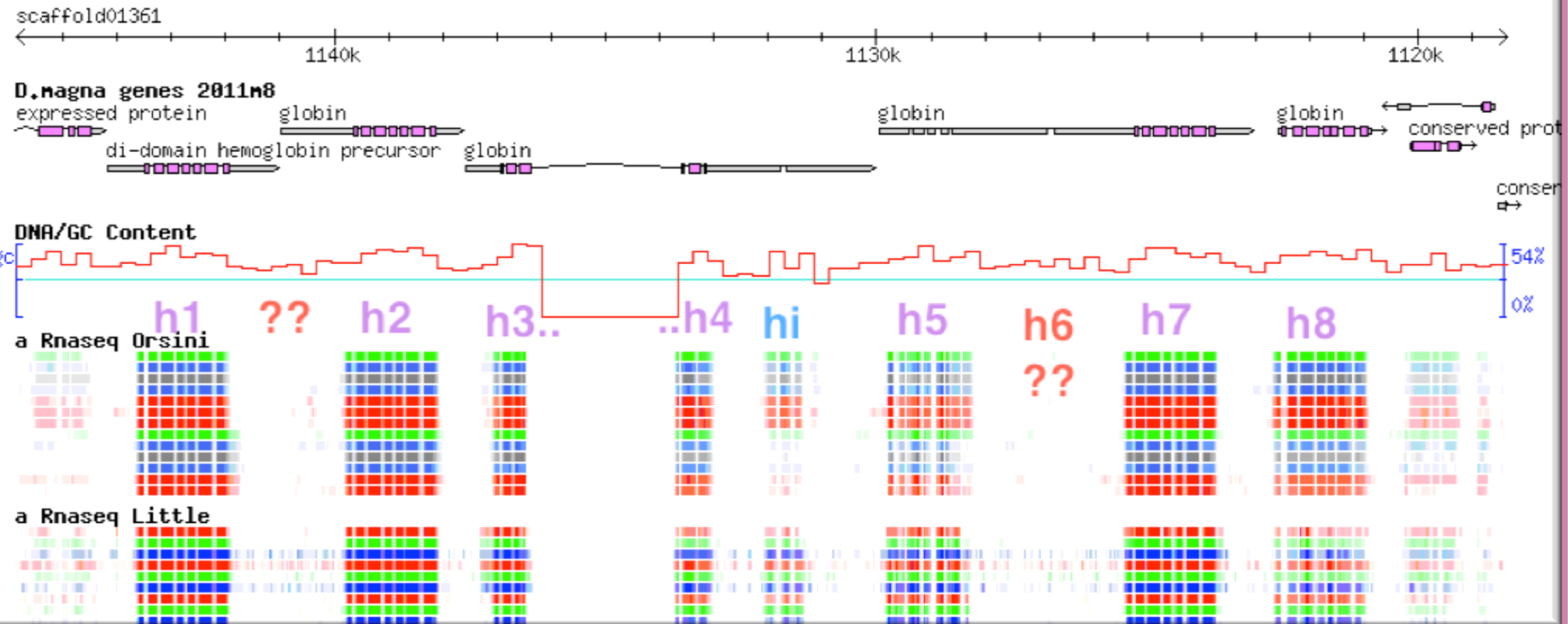
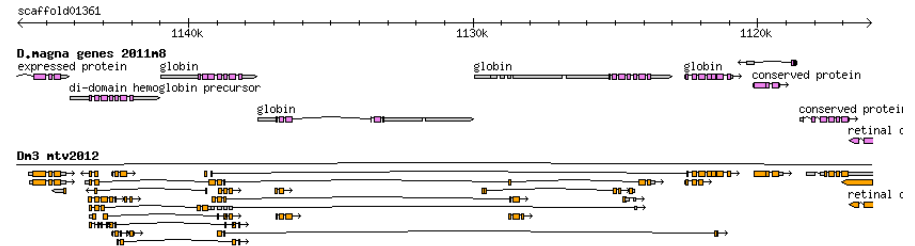
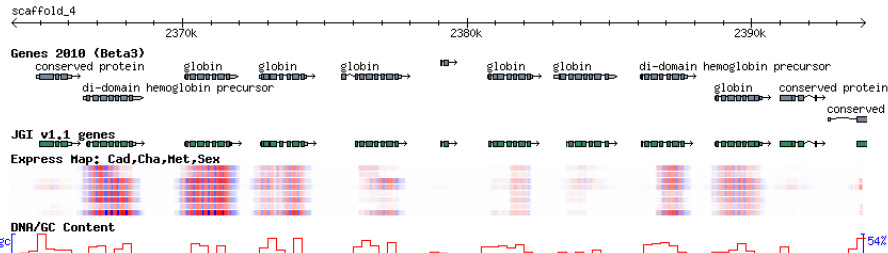
Dap magna Hemoglobin-7 (or 8?)



Hemoglobin Mystery Spans

Dap pulex Hemoglobin-8

Dap magna Hemoglobin-7 (or 8?)



You can annotate genes!

The image displays a genome browser interface with a genomic track from 740k to 780k. Several gene models are visible, including 'A User Choice', 'D.magna genes 2011m8', and 'Dm3 mtv2012'. A pop-up window provides detailed information for the gene **daphmag3mtv3l171t1**, including its locus, quality metrics, cross-references, and update options.

Gene daphmag3mtv3l171t1
Locus [scaffold01361:738525-773910](#)
Quality intron=24/24, aa=2118, 73% mapped, homl=24%, estcov=41%
Crossref UniProt:E9FWM0_DAPPU, oid:daphmag3tri7trimsu2loc1957c0t5
[View Gene Details ..](#)

Update Choices [View Changes table](#)

No change . . Alternate transcript . .

Best model . . None are good at locus

Drop model . . Skip locus, no gene here

Note:

Saves time to record notes, choice, when viewing

Daphnia magna DE genes

CA Carbaryl -

CO Control ;

daphmag3mtv3.edger3x.
CO.CA.txt

A = logConc;

M = logFoldChange

Gene ID	A	M	Pr	FDR	Annotation
daphmag3mtv3l12646t1	-19.3	2.8	1.90E-12	5.30E-09	Aromatic-L-amino-acid decarboxylase
daphmag3mtv3l31768t1	-20.7	2.4	3.10E-09	5.30E-06	CA+BX, chitinase /ARP2_G1856
daphmag3mtv3l22251t1	-19.2	2.3	3.20E-09	5.30E-06	CA+BX, chitinase /ARP2_G1856
daphmag3mtv3l12793t1	-17.5	2.1	1.10E-07	0.0001	Unknown,DAPPU_314986
daphmag3mtv3l15011t1	-16.7	1.9	9.70E-07	0.0008	Sulfotransferase family/ARP2_G20
daphmag3mtv3l12757t2	-17.9	1.8	2.60E-06	0.0021	glucosyl/glucuronosyl transferases
daphmag3mtv3l7093t1	-20.4	1.7	8.10E-06	0.006	salivary gland-expressed bhlh
daphmag3mtv3l7446t1	-16.3	1.7	8.80E-06	0.0064	ABC membrane transporter
daphmag3mtv3l29356t1	-21.4	1.7	1.50E-05	0.0101	chitinase /ARP2_G1856

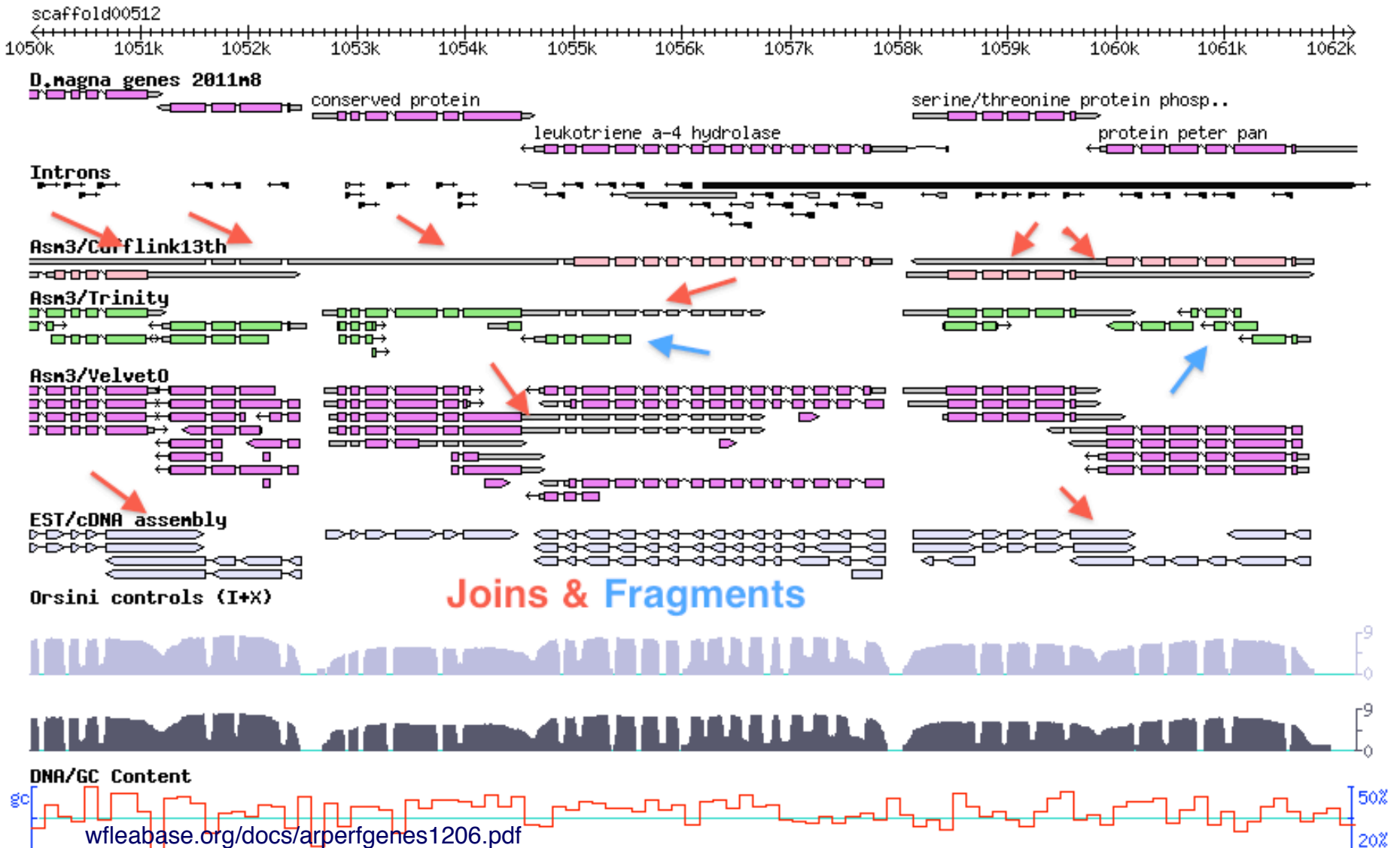
Gene ID	A	M	Pr	FDR	Annotation
daphmag3mtv3l43412t1	-26	6.2	3.80E-18	7.50E-14	unmapped , Unknown
daphmag3mtv3l18392t1	-19.8	2.4	1.00E-09	5.40E-06	CD9 antigen/ARP9_G1857
daphmag3mtv3l22251t1	-19.2	2.4	1.50E-09	5.90E-06	CA+BX, chitinase /ARP2_G1856
daphmag3mtv3l31768t1	-20.7	2.4	1.70E-09	6.30E-06	CA+BX, chitinase /ARP2_G1856
daphmag3mtv3l29567t1	-20.3	2.4	4.90E-09	1.60E-05	unmapped , Unknown
daphmag3mtv3l24735t1	-19.9	1.8	2.50E-06	0.0037	Unknown
daphmag3mtv3l25123t1	-21.2	1.7	1.70E-05	0.0184	unmapped , Unknown
daphmag3mtv3l17144t1	-20.4	1.5	7.50E-05	0.057	Secreted protein/ARP9_G473

BX Bacteria toxic

- CO Control ;

daphmag3mtv3.edger3x.C
O.BX.txt

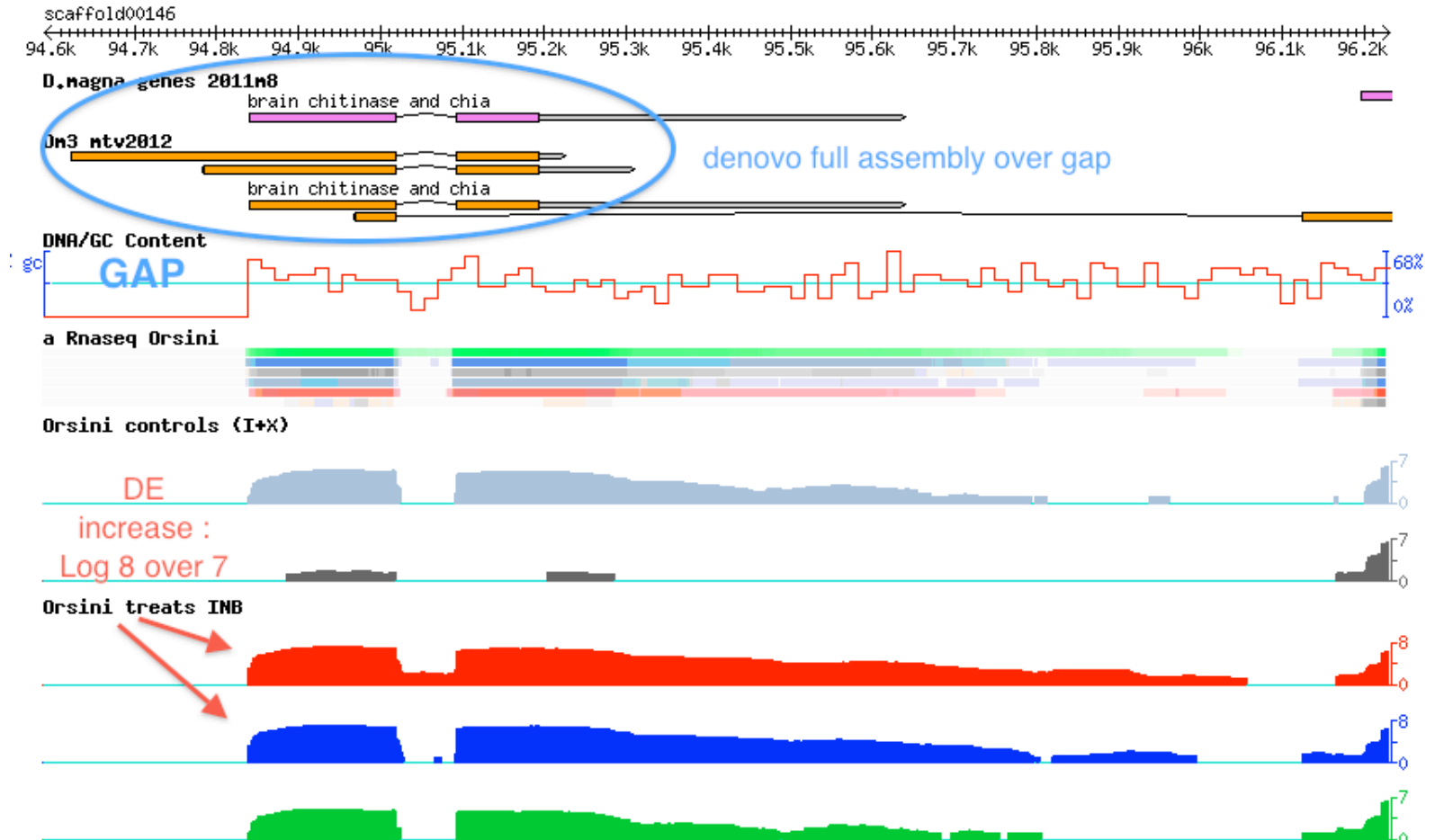
Too little data and too much? Asm-RNA joins & fragments



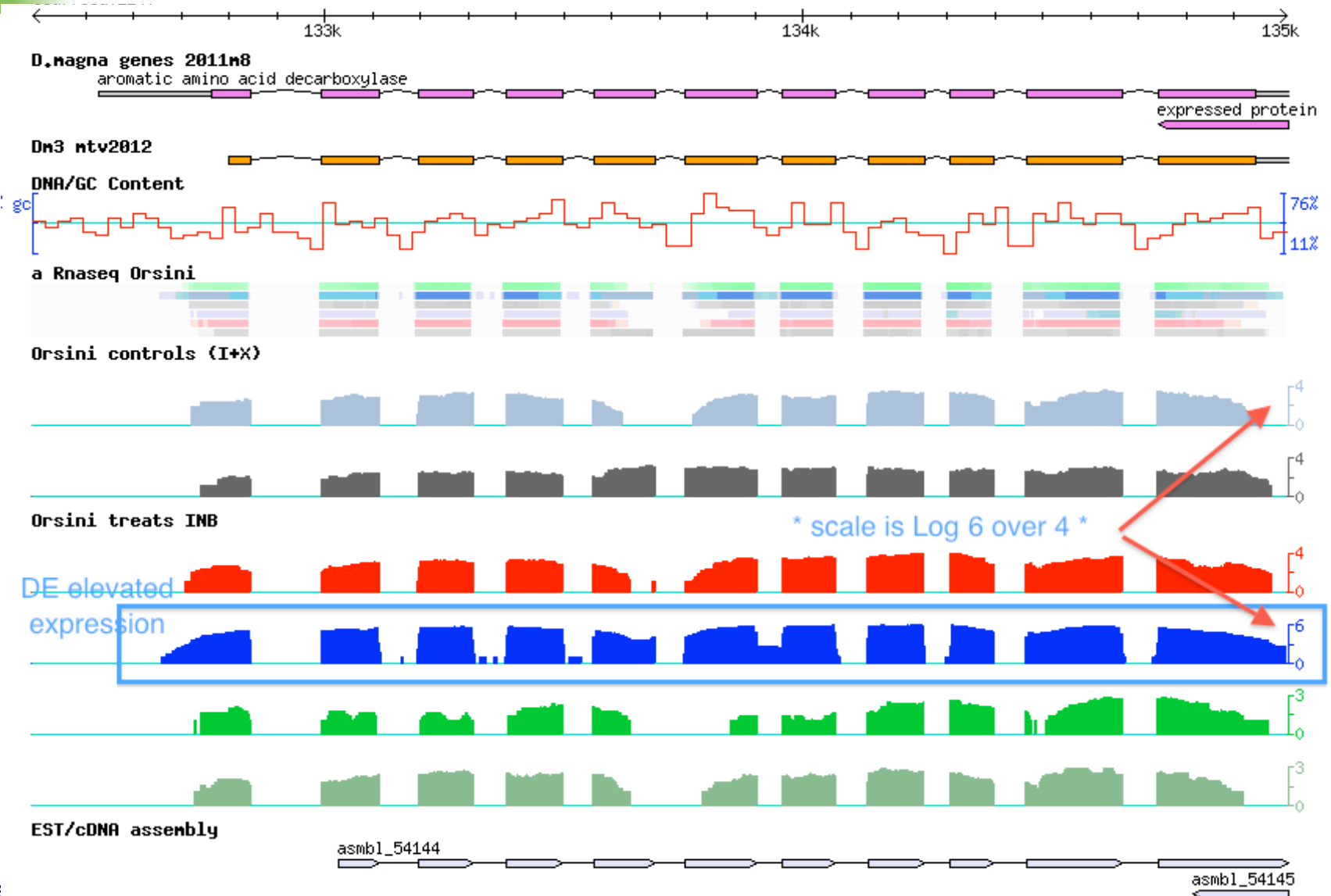
When you need a genome: De novo: alternate or paralog?



When to avoid the genome: de novo RNA spans genome gaps



DE elevated expression



DE extra exons

scaffold01663

110k

111k

112k

113k

D.nagna genes 2011n8

conserved protein

Dn3 mtv2012

conserved protein

DNA/GC Content

gc

64%

10%

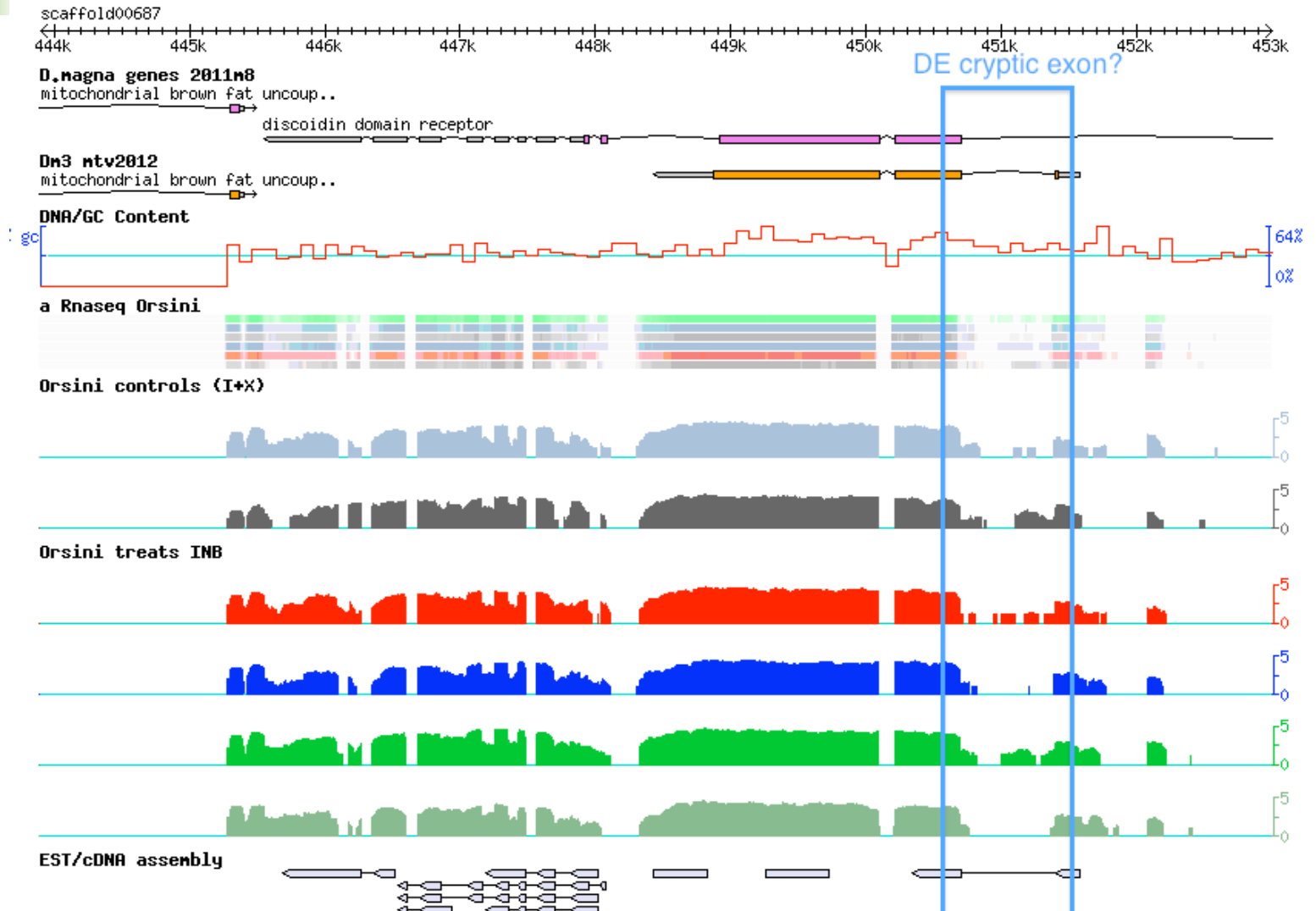
a Rnaseq Orsini

Orsini controls (I+X)

Orsini treats INB

DE extra exons

DE cryptic exons?





Daphnia notes

gilbertd@indiana.edu

<http://wfleabase.org/genome/> for magna, pulex,

→ server7.wfleabase.org/genome/Daphnia_magna/prerelease/

Genome maps

server7.wfleabase.org:8091/gbrowse/cgi-bin/gbrowse/daphnia_magna2/

server7.wfleabase.org:8091/gbrowse/cgi-bin/gbrowse/daphnia_pulex/

2011 gene models

gene-predictions/daphmagna_2011/

2012 draft gene models (from newer rna asm but not this newest rna)

gene-predictions/daphmagna_201205/

Differential Expression for StressFlea RNA on 2012 draft genes

gene-predictions/daphmagna_201205/de_mag3mtv3/

counts of read/transcript, includes unmapped genes

edgeR rough-draft DE stats from these counts



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

... and others

Links to this work

arthropods.eugenes.org/

arthropods.eugenes.org/EvidentialGene/

wfleabase.org

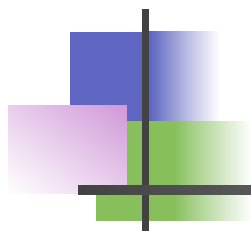
www.bio.net

14+ Bug genomes

perfecting Bug genes

Daphnia genomics

Arthropod news/discussion list



Arthropods/euGenes database

- ✓ New in progress..
- crustacea/tick/insect balance
- ✓ OLD 2010 OrthoMCL orthology for 263,000 current genes of 14 species
- ✓ Web-searchable gene pages, ..
- ✓ Summaries of gene structure ..

gene ortholog search at

Find where *Pediculus* is missing gene groups that all other arthropods have, with the search at Arthropods.euGenes.org. Or find groups with 1-1 orthologs in all species. You can search for groups by gene name, function or ID.

Results from Arthropod Gene Clusters v2

Query: arthropod2xml+all:arthropoda +Aedes:[1 TO 999] +Anopheles:[1 TO 999] +Culex:[1 TO 999] +DrosMel:[1 TO 999] +DrosMoj:[1 TO 999] +DrosPse:[1 TO 999] +Apis:[1 TO 999] +Nasonia:[1 TO 999] +Tribolium:[1 TO 999] +Aphid:[1 TO 999] +Pediculus:0 +Daphnia:[1 TO 999] +Ixodes:[1 TO 999] +Bombyx:[1 TO 999]

No. matches = 58 of 28773 documents, in 0.021 sec.

#	GeneID	ntaxa	ngene	group-identity	occurrence	description
1	ARP2_G99	13	66	27	Aed Ano Aph Api Bom Cul Dap Dro Dro Ixo Nas Tri	alpha-amylase
2	ARP2_G159	13	50	31	Aed Ano Aph Api Bom Cul Dap Dro Dro Ixo Nas Tri	CRAL/TRIO domain-containing protein
3	ARP2_G324	13	33	31	Aed Ano Aph Api Bom Cul Dap Dro Dro Ixo Nas Tri	lactosylceramide 4-alpha-galactosyltransferase ...
4	ARP2_G419	13	29	26	Aed Ano Aph Api Bom Cul Dap Dro Dro Ixo Nas Tri	conserved hypothetical protein
5	ARP2_G442	13	28	35	Aed Ano Aph Api Bom Cul Dap Dro Dro Ixo Nas Tri	conserved hypothetical protein
6	ARP2_G540	13	25	60	Aed Ano Aph Api Bom Cul Dap Dro Dro Ixo Nas Tri	calcium/calmodulin-dependent serine protein k ...
7	ARP2_G1068	13	19	30	Aed Ano Aph Api Bom Cul Dap Dro Dro Ixo Nas Tri	abhydrolase domain-containing protein
8	ARP2_G1144	13	19	57	Aed Ano Aph Api Bom Cul Dap Dro Dro Ixo Nas Tri	60S acidic ribosomal protein P2
9	ARP2_G1273	13	18	35	Aed Ano Aph Api Bom Cul Dap Dro Dro Ixo Nas Tri	conserved hypothetical protein
10	ARP2_G1421	13	18	37	Aed Ano Aph Api Bom Cul Dap Dro Dro Ixo Nas Tri	ribonuclease H1

Arthropod gene group ARP2_G1421

basic information						occurrence	
GeneID	species	ntaxa	ngene	group-eval	group-identity	date	Get cluster proteins
ARP2_G1421	Arthropoda	13	18	9.4e-11	37	20091228	

Descr.: ribonuclease H1

similar_genes			
spp	acc	eval	iden def
<i>Aedes aegypti</i>	aedes_AAEL017101-PA	4e-19	37
<i>Anopheles gambiae</i>	anopheles_AGAP008088-PA	2e-18	34
<i>Acyrtosiphon pisum</i>	aphid_ncbi_hmm235253	1e-16	32 similar to ribonuclease H1

dbx: Aae.19968, XP_317369, Entrez: 1277863, Aga.43603, ARP1_G1274, XP_001948714, XM_001948679.1

arthropods.eugenes.org



ARP3x Arthropods Summary

- *Daphnia* maintains the most homology to human and other eukaryotes, followed by *Ixodes*. Among insects, *Tribolium* has most non-insect homology &/or best gene models.
- Gene duplication rate is more variable than singleton rate.
- xxx
- // As yet unfound ortholog genes exist in most of these genomes.



Best practices for perfect genes

- Gene construction software and methods continue to improve, but are imperfect.
- Current best strategy uses several methods, extract the best of their many results.
- Rough edges need smoothing: predictor models and transcript assemblies each have qualities the other lacks, for coding sequences and sequence signals, gene holes and mash-ups.
- Multiple lines of gene evidence scores the quality of competing gene constructions to select a best, if not yet perfect, gene set.

You can annotate genes!

The screenshot displays a genome browser interface with a scale from 0M to 3M. The main view shows a genomic region from 1050k to 1062k. Annotations include:

- D.magna genes 2011 m8**: conserved protein, leukotriene a-4 hydrolase, serine/threonine protein phosph., protein peter pan.
- Introns**: Diagram showing intron-exon structure.
- Asm3/Cufflink13th**: Gene model with exons in red.
- Asm3/Trinity**: Gene model with exons in green.
- Asm3/Velvet0**: Gene model with exons in purple.
- EST/cDNA assembly**: Diagram showing transcript alignments.
- Orsini controls (I+X)**: Diagram showing control elements.
- DNA/GC Content**: Graph showing GC content (red line) and GC percentage (blue line).

A dialog box is open for the gene **Gene dmag3cuf13th_Gsc00512g804t1**, located at **Locus scaffold00512:1049059-1057929** with a **Quality aa=569**. The dialog provides options for **Update Choices**:

- No change ..
- Best model ..
- Drop model ..
- Alternate transcript ..
- None are good at locus
- Skip locus, no gene here

The **Note** field contains the text: "This is a horrible join of 5 genes". An **Ok** button is visible.

Saves time to record notes, choice, when viewing