Daphnia Genomes at the Crossroads

2010 March

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

Daphnia genomes bring new insights to many areas of biology (basic, health and biotech)

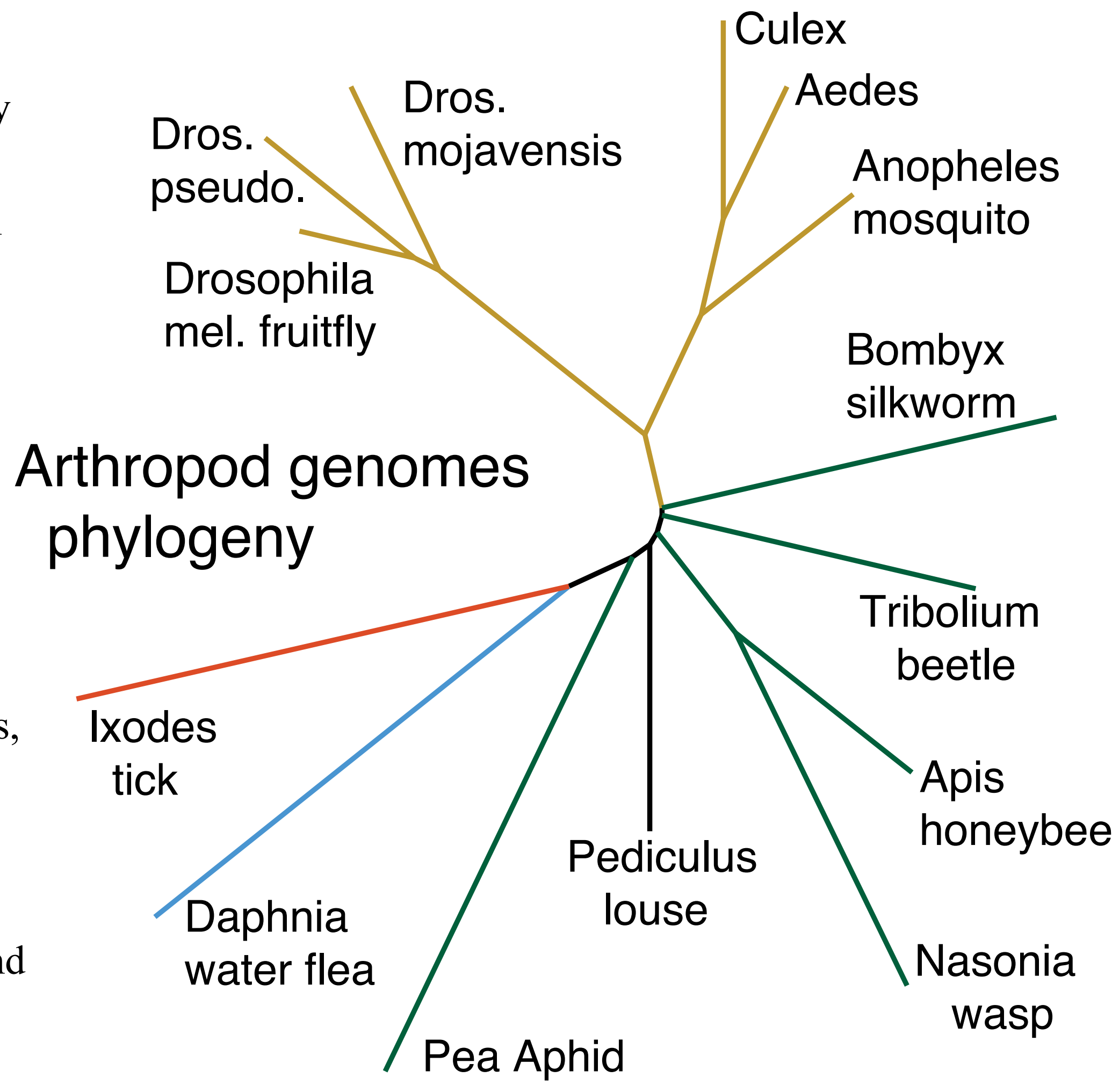
- Crossroad for environmental and ecological genomics

- Crossroad for study of gene expression divergence and evolution of new genes.

- Crossroad for understanding basic DNA replication and gene creation mechanisms.

- Crossroad for modeling human genetic diseases

With the closest homology to human genes among insects and known arthropod genomes, *Daphnia* offers a new model for health genomics. With the highest rate of gene duplications known among arthropods, it provides an important model for biomedical study of gene copy-number variation, as found in cancers.



Crossroads of Arthropods and Human models

Daphnia has best matches and longest alignments to human and other model eukaryote gene sets (*Tribolium* has the best of the insects). *Daphnia* has significantly more matches to model genes than *Tribolium* (p < e-15) [1]. Using phylogenetic orthology methods (protein alignment and tree construction), Phylomedb [4] and PHiGs [5] both find similar results, Daphnia > Tribolium > other insects.

Best match to model genes (SwissProt reference)

Arthropod	Human	Mouse	Worm	Plant
Daphnia	3286	2849	573	1004
Ixodes	2465	2180	279	447
Tribolium	1969	1707	283	524
Apis	1724	1486	235	482
Pediculus	1593	1352	204	410
Drosophila	563	463	134	330

Percent model genes found (SwissProt reference)

Arthropod	Human	Mouse	Worm	Plant
Daphnia	90.4	91.5	94.9	88.1
Tribolium	89.7	90.7	93.3	85.7
Apis	88.9	90.1	92.7	83.5
Drosophila	87.6	88.7	93.1	86.9
Pediculus	89.0	89.8	92.3	81.7
Ixodes	87.7	88.8	90.5	80.1

Many gene duplications: the stuff of evolution

Gene models for 14 arthropod genomes are summarized in Table C in categories of duplicated or singleton genes. This indicates the large difference in gene counts from 16,000 in Dipterans to over 30,000 in Aphid and *Daphnia*. *Daphnia* and aphid have a high rate of four times more duplicated genes than dipteran insects [3].

Table C. Arthropod duplicate and single gene counts

	Gene Count		**Relative to Dipterans**	
	Single	**Double**	**Single**	**Double**
Daphnia	17100	14400	1.4	4
Aphid	17600	14500	1.4	4
Ixodes	15500	4800	1.3	1.2
Nasonia	13900	5200	1.1	1.5
Tribolium	12700	3300	1	1
Dipterans	12500	3600	1	1
Apis	13200	2300	1	0.7
Pediculus	10200	800	0.8	0.2

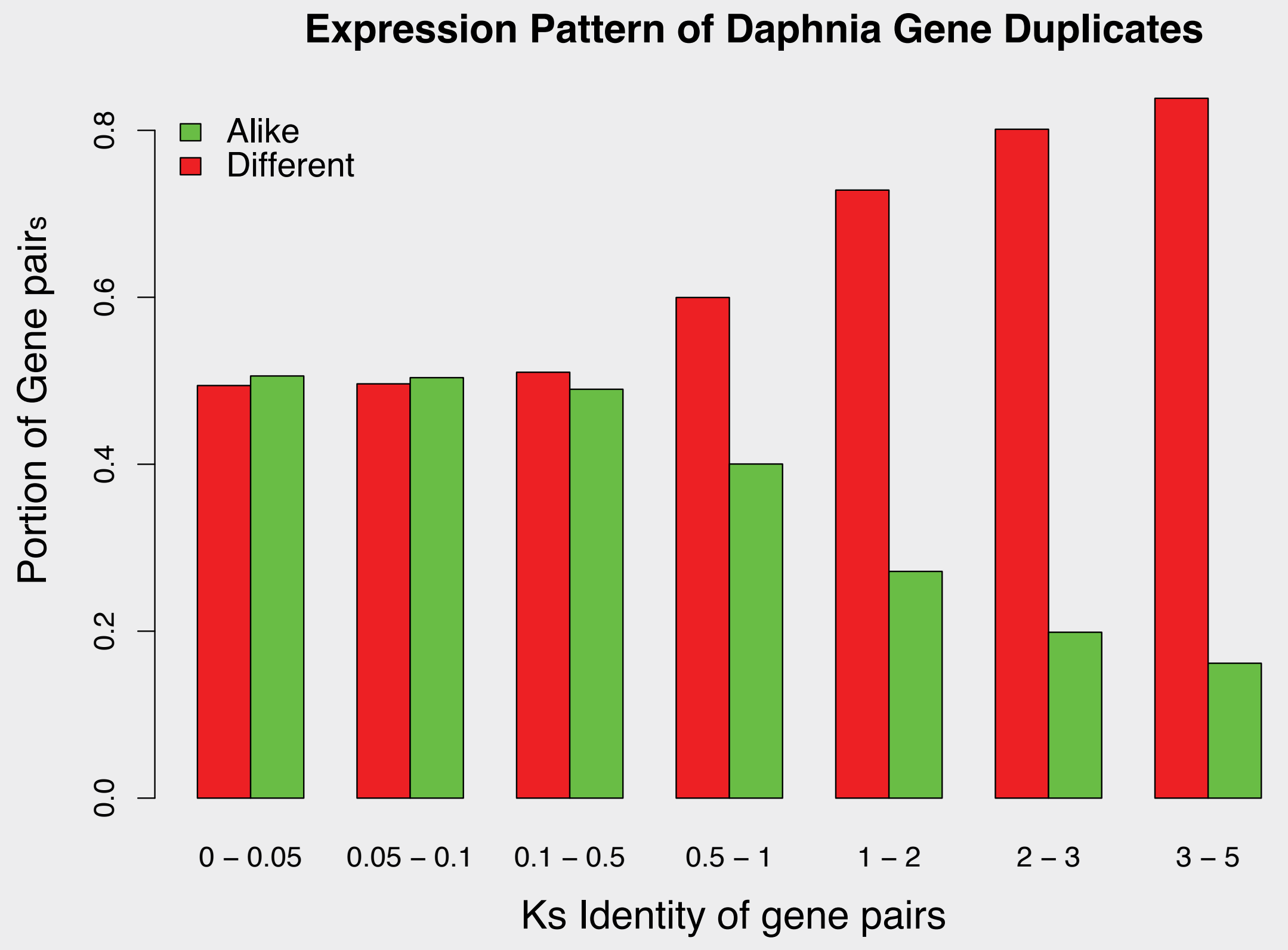
Single = single copy gene, Double = 2+ paralogous genes, after removing poor gene models. Poor models are transposons and short/partial genes, for Aphid (5,400), *Nasonia* (7,000) and *Daphnia* (5,000), less than 500 for other species. Dipterans are the average of 6 fly genomes. Genes with and without orthologs are combined.

Are there similarities in Aphid and *Daphnia* have led to this? Both species are asexual parthenogenic during much of their population history. There is evidence that asexuality includes mitotic recombination, where these species may have diverged from sexual species. Might this include a greater propensity for gene duplications?

Phylogeny is not an explanation: the parasite *Pediculus* is taxonomically closest to aphids yet is at the other extreme with few duplicate genes. The largest class of gene duplications is clade-specific for both. This agrees with nematodes and plants, and says that duplicates are involved in rapid adaptation. Evidence from expression for *Daphnia* supports both uses: near identical duplicates with different expression, others sharing the same expression.

Born Different: Expression Diverges in New Genes

Daphnia's many duplicate genes appear to be born different; 50% of the most identical genes have different expression patterns [2]. This may help explain the abundance of duplications, as they appear to function differently in complex metabolic and developmental pathways. Expression divergence increases as the gene pair sequences diverge.



Daphnia magna expands this crossroads model

In the large EST set produced by the *Daphnia magna* genome collaboration, are 25,000 uniquely located EST assemblies (on the draft magna genome), with an additional 10,000 alternate transcript forms [6]. This is a large number compared to other arthropod EST assemblies. However, these are not all complete genes, as many will map to one gene when a complete genome is available. Alternate *D. magna* transcripts have an unusually high portion of retained introns, compared to other arthropod ESTs. Protein homology for *Daphnia magna* ESTs is similar to *D. pulex*, with many ribosomal and cuticular proteins, hemoglobins, opsins and others.

Summary of EST assemblies

	Daphnia magna	*Daphnia pulex*	*Drosophila melanog.*
Total EST	1274539	166289	567759
Any alignment	1020785	145578	561200
Valid align	879441	114128	533435
Assemblies	58021	18211	42618
Subclusters	44353	15827	33329

Table N. Gene Function inferred for *Daphnia magna* and *pulex* EST assemblies from homology

N *Daphnia magna*	N *Daphnia pulex*
54 Ribosomal protein	96 Cuticle protein5a
40 Cytochrome Pnn	93 Ankyrin repeat protein
40 Mitochondrial ribosomal protein	73 Cytochrome Pnn
36 Eukaryotic translation initiation factor	68 Zinc finger protein
35 NnS ribosomal protein	62 Ribosomal protein
35 Zinc finger protein	57 Cuticular protein
33 Cuticle protein5a	57 Integrase
32 Solute carrier family	56 Mitochondrial ribosomal protein
24 Transmembrane protein	48 Focal adhesion kinase
17 Male sterility domain-containing protein	41 Sptzle 2 protein
15 1-acylglycerol-3-phosphate acyltransferase	39 Trypsin serine protease
15 Serine protease	37 Heat shock protein
15 Zinc metalloproteinase	34 Eukaryotic translation initiation factor
13 Hemoglobin	33 Tubulin folding cofactor D
13 Opsin	32 Oviductin
13 Threonine-protein kinase	32 Transmembrane protein
12 Cuticle protein	31 Opsin
12 Glutathione S-transferase	31 Secreted protein, putative
12 Peroxinectin	29 Inositol 1,4,5,-tris-phosphate receptor
11 Cuticular protein	28 Germinal histone H4 gene
11 DEAD box ATP-dependent RNA helicase	27 ORF2-encoded protein
11 Dehydrodolichyl diphosphate synthase	27 Pao retrotransposon peptidase family protein
11 Sptzle 2 protein	27 Solute carrier family

Informatics of genomes: an essential for biology

Recipe for genome annotation

Take a genome assembly, and a good set of gene evidence from EST sequences, proteins of related species, and next generation data of tiling and RNA-seq expression, then one can model genes rather accurately according to gene evidence. Building a new gene set with current software is in progress for Daphnia, completed for Aphid and other insects. **PASA** is used for EST assembly and gene validation. **BLAST** is used to locate related proteins (tblastn), and annotate predicted genes (blastp). **Exonerate** refines protein gene mappings. **Augustus** models genes using all evidence of ESTs, mapped proteins, tiling and RNA-Seq expression. Other predictors, such as fgenesh, GeneID, SNAP, Gnomon, are valuable additions. Methods for combining predictions to one best set are still problematic; one such is **EvidenceModeler** that uses evidence weightings.

Gene models are D. pulex JGI V11 (official release 1 from 2007), D pulex NCBI Gnomon (2007), and new predictions (Aug25, 2010). Statistics are the proportion of bases matching evidence, and overlaps are the number of features overlapping evidence. Evidence includes (1,2) EST assemblies for D. pulex and D. magna, (3) Proteins-Arp2 the complete protein sets from 6 closest arthropod genomes (aphid, apis, crab, ixodes, pediculus, tribolium), (4) Tile genes, the genome tile expression gene-like regions unpredicted by JGI genes. **Best1** is selected from all 3 predictors (Augustus, JGI, Gnomon) to maximize evidence scores. This improves overall quality. The best sources are 25030 AUG25, 10688 JGI, 9760 NCBI_GNO. I discard 5,000 of the no-evidence, single-predictor models, giving a total as above of 34,000 as protein coding, 6,000 classed as non-coding RNA for trivial CDS spans in expression-supported long exon spans. Evidence statistics are for scaffolds 1-9 subset.

Table X. New gene models (Augustus) for *Daphnia pulex* compared to 2007 genes

Evidence	N	Statistic	JGI_v11	Gnomon	Augustus	Best1
				Exon Sensitivity		
EST-D.pulex		bases	0.726	0.725	0.906	0.889
	16510	overlaps	14632	14884	16225	16074
EST-D.magna		bases	0.782	0.823	0.931	0.906
	23829	overlaps	19545	20222	22911	22373
Proteins-Arp2		bases	0.753	0.870	0.904	0.923
	35660	overlaps	28726	31351	31759	32901
Tile genes		bases	0.000	0.108	0.780	0.663
	10223	overlaps	0	1365	7537	6287
				Exon Specificity		
All evidence		bases	0.777	0.798	0.516	0.577
	284561	overlaps	25302	25902	28828	31793
				Gene model Accuracy		
Proteins-Arp2	2314	found gene	2235	2263	2265	2309
		CDS bases	0.632	0.680	0.681	0.704
		split genes	0.113	0.115	0.171	0.114
		join genes	0.008	0.011	0.046	0.012
				Gene Totals		
Coding bases		bases	30Mb	36Mb	41Mb	45Mb
Genes count		count	31K	37K	36K	40K*

* 34K mRNA; 6K noncoding-RNA

References

1. Gilbert DG, 2009. OrthoMCL clustering among 14 arthropod proteomes (ARP2). http://arthropods.eugenes.org/
2. Daphnia Genome Consortium (2010) *Daphnia pulex* genome paper, in preparation.
3. Gilbert, D.G. 2009. Aphid and Waterflea have a High Rate of Gene Duplications Compared to Other Arthropods. manuscript, May 2009.
4. Huerta-Cepas J, Bueno A, Dopazo J, Gabaldón T. PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Res. 2008 36:D491-6. http://phylomedb.org/
5. Dehal PS, Boore JL. 2006. A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. BMC Bioinformatics 2006, 7:201 doi:10.1186/1471-2105-7-201
6. Daphnia magna genome prerelease, http://wfleabase.org/genome/Daphnia_magna/