Aphid and Water flea have a High Rate of Gene Duplications Compared to Other Arthropods 2009

Don Gilbert and Mike Pfrender, in collaboration with Daphnia and Aphid genome consortia

Abstract

Many new genomes are being deciphered with the advent of rapid, low cost next generation sequencing. New genome discoveries are facilitated with ready access for biologists to informatics tools on shared cyberinfrastructure. Genome informatics tools have been integrated and are provided publicly for genome scientists on the US **TeraGrid** (**gmod.org/Genome_grid**). Arthropod genomes provide examples of rapid genome discovery using this, with interesting results from comparing aphid and water flea genomes.

Comparison of gene orthology and paralogy groups among thirteen arthropod genomes (**insects.eugenes.org/arthropods/**) finds the pea aphid *Acyrthosiphon pisum* and the water flea *Daphnia pulex*, both cyclical asexuals, have four times the gene duplications of other arthropods. These duplicates are not from whole genome duplication, and artifacts do not account for this.

This observation suggests a new phenomenon of interest to ecological, evolutionary and biomedical genomics. A catalog of overabundant, annotated genes indicates that both species have independently acquired extra genes for mitosis and chromosomal maintenance. Speculation on reasons for this expansion include mitosis-related genes have evolved with asexuality. High phenotypic plasticity and rapid environmental responses are exhibited by both species, and likely are facilitated by the many gene duplications. There are apparent effects of gene conversion, differentiation and gene dosage in the preservation of these duplicate genes.

(DG) Biology Dept., Indiana U., Bloomington, IN 47405, gilbertd@indiana.edu;
(MP) Biology Dept., Utah State U., Logan, UT 8432, pfrender@biology.usu.edu

Table 1. Arthropod duplicate and singleton gene numbers

	Gene Count		Relative to Dipterans	
	Single	**Double**	**Single**	**Double**
Aphid	17600	14500	1.4	4
Daphnia	17100	14400	1.4	4
Nasonia	13900	5200	1.1	1.5
Tribolium	12700	3300	1	1
Dipterans	12500	3600	1	1
Apis	13200	2300	1.1	0.7
Ixodes	14700	2200	1.2	0.6
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.

Introduction

Rapid and low cost sequencing technology yields a wealth of new genomes, but the informatics component is now a bottleneck to genome discovery. Discoveries in newly sequenced arthropod genomes are reported here, from application of standard genome informatics. These results expand on genome publications of the International Aphid Genomics Consortium [1] and the Daphnia Genomics Consortium [2], and are in debt to their encouragement of collaborations.

During annotation of the aphid genome an unusual number of gene duplications were found. Another arthropod, the crustacean *Daphnia pulex*, has many more gene duplications than insect relatives. Duplication is an important aspect genome biology, as evidence rises for their existence and role in metabolic pathways, adaptation to environment, and evolution. To assess significance of these duplications, and provide a comparative gene reference for experts, a public web database the gene catalogs of thirteen arthropods has been developed.

Gene prediction Grid

A complete automated prediction and annotation gene set of the Pea Aphid *Acyrthosiphon pisum* genome was produced and provided to the International Aphid Genomics Consortium. Genome informatics uses several open-source genome analysis tools, run on **TeraGrid** cyberinfrastructure, with genome data partitioning and parallelization. These are based on those developed for several eukaryotes including insects. These genome analysis and annotation tools are deployed for public use on TeraGrid cyberinfrastructure in a community folder, with documentation at gmod.org/Genome_grid. This is available to any scientist with TeraGrid access; scientists are encouraged to contact the author for advice.

A gene prediction set has been built with **AUGUSTUS** predictor using a full range of evidence. The result is 33,700 gene models, plus 7,000 alternate transcripts. Of these, 73% of models are supported by EST or protein homology. Annotations include evidence IDs, description and protein functions drawn from **UniProt** and **RefProt** for best homology, and other data. *Nasonia* and *Tribolium* are the best gene homology sources. The largest group with at least 2,700 genes is transposon-related. Associated data include 25,000 cDNA/EST assemblies produced with **PASA** software from 160,000 *Acyrthosiphon pisum* ESTs. Homologous arthropod genes from *Tribolium*, *Nasonia* and *Daphnia were* mapped to Aphid with **Exonerate**. These Aphid gene annotations are searchable by function, key words and homology gene names, at the DroSpeGe insect genome database (insects.eugenes.org). BLAST searches and **GBrowse** genome map views are also available. These genes can be searched via Internet search engines, for instance, **Google** for *"pea aphid gene" topoisomerase* will return *"Pea aphid gene ... DNA topoisomerase 2 .. insects.eugenes.org/genepage/aphid/DGIL_AUG5s5510g8t1"*.

Arthropod gene groups

A need among expert annotators of new genomes is comparative genome results, for interpreting homology, functional assignments and potential gene gains and losses. This has been provided with a web-accessible orthology database of arthropod genomes **insects.eugenes.org/-arthropods/** . Protein gene models of 13 arthropod genomes were collected from genome providers, and gene groups were computed with **OrthoMCL** to identify orthologs and recently arisen paralogs, or in-paralogs. Different criteria for retaining predictions are used by genome data providers. Some exclude predictions that lack homology or ESTs, others include all predictions. Gene sets were chosen to increase consistency of methods as far as possible. Arthropod annotated gene groups are available for searching, browsing and download, including sequence searches by BLAST and **Psi-BLAST**. These can be searched via **Google** e.g. *"arthropod gene group" topoisomerase.*

Gene models for these 13 arthropod genomes are summarized in Fig 2 in categories of orthologs or species unique, duplicate paralogs or singleton genes. This indicates the large difference in gene counts from 16,000 in dipterans to over 30,000 in Aphid and *Daphnia*. The main effects for high versus low gene count appear to be paralogs. Aphid has 2,475 genes in 625 groups compared to 621 average insect genes, a rate of 3.8 to 1. *Daphnia* has 2,621 duplicated genes in 589 groups, compared to 610 average insect genes, a rate of 4.4 to 1. These two species are also abundant in genes in many groups that lack homology (Fig. 2B). Transposon genes (TE) are abundant in *Nasonia* and Aphid. Mosquitoes and fruitflies have more orthologs due to inclusion of 3 related species from each clade.

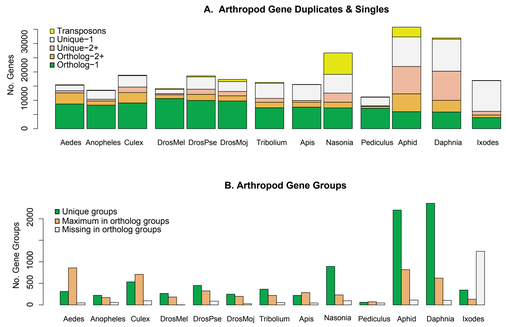

Fig 1. Arthropod genes classed by orthology and duplication (A) Arthropod gene duplicates and single gene counts, separated in those with orthologs (Ortholog-1, Ortholog-2+), those without orthologs (Unique-1, Unique-2+), and transposon-related genes. Aphid has expanded duplicated genes (ortholog-2+ and unique-2+), as has *Daphnia*. **(B)** Arthropod gene group counts, classed by unique, maximum and minimum per species. Aphid and *Daphnia* have abundant unique gene families.

Discussion

This variation in duplicate numbers is much larger than can be accounted for by transposons, allelic heterozygosity, or different prediction methods. Refinements in gene prediction and validation, and genome assemblies can be expected to change these gene catalogs, as it has for other genomes.

It is interesting to speculate whether similarities in Aphid and *Daphnia* that have led to this. Both species are asexual parthenogenic during much of their population history, but alternate with sexual reproduction. 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?

Alternate speculations include a reduced gene set in diperans, or genome size variation. Phylogeny is a weak explanation: the parasitic insect *Pediculus* is taxonomically closest to aphids yet is at the other extreme with few duplicate genes; the two hymenoptera also diverge. Genome size does not correlate well with gene numbers, and none of these have evidence of whole genome duplication. As genomes of parasitic organisms can be reduced and specialized, Aphid and *Daphnia* may share another relationship for expanded gene numbers. The largest class of gene duplications is clade-specific for Aphid and *Daphnia*. This agrees with species-specific duplications in nematode and plants, and support conclusions that duplicates are involved in rapid adaptation to changing environments.

Duplicated genes can be adaptive by developing new functions (**neofunctionalization**) or by remaining identical for increased **gene dosage**. Preliminary evidence from genome tile expression for *Daphnia* supports both: some near identical duplicates show different expression to environmental factors, others share the same expression. These are compelling results that *Daphnia* and Aphid have over-abundant gene duplications. Speculation suggests chromosomal maintenance and mitosis genes contribute to this 4-fold higher duplication rate, and may be related to their asexuality.

Acknowledgements

These analyses were performed with support from the NSF (DBI-0640462) and the NIH, including TeraGrid award (TG-MCB060059). Members of the International Aphid Genomics Consortium and the Daphnia Genomics Consortium have shared significant insights and data contributing to this work. *Daphnia pulex* sequencing and portions of the analyses were performed at the DOE Joint Genome Institute in collaboration with the Daphnia Genomics Consortium (DGC, daphnia.cgb.indiana.edu). Pea aphid genome sequencing, assembly and portions of the analyses were performed at Baylor College of Medicine, and by the International Aphid Genomics Consortium (www.hgsc.bcm.tmc.edu/projects/aphid/).

Aphid & Daphnia: why more duplicates?

Gene duplicates are more abundant in Aphid and *Daphnia*. Table 2 summarizes this major distinction in genes, irrespective of orthology. Aphid and *Daphnia* both have 400% as many duplicate genes as dipterans, even excluding poor quality gene models. In *Daphnia*, these duplicates are often found in tandem arrays of gene copies [2]. The arthropods have similar numbers of singleton genes, though Aphid and *Daphnia* appear to have 40% more of these.

Investigating gene groups that *Daphnia* and Aphid share as over-abundant to others turns up interesting **mitosis**-related and **chromosome maintenance** genes: Structural maintenance of chromosomes 6 (*Smc6*), DNA topoisomerase 2, and several mitotic kinases. There are also duplicated genes in these two species in **DNA/RNA processing** functions. These functions include of chromatin assembly, microtubule genes, transcription initiation and regulation, and RNA polymerases and binding genes, where *Daphnia* and Aphid have two to three times as many genes (Figure 2).

Fig 2. DNA replication gene expansions in Aphid and *Daphnia*, compared to Insects. Aphid and *Daphnia* have more genes than the insect average of 1 gene in these groups.

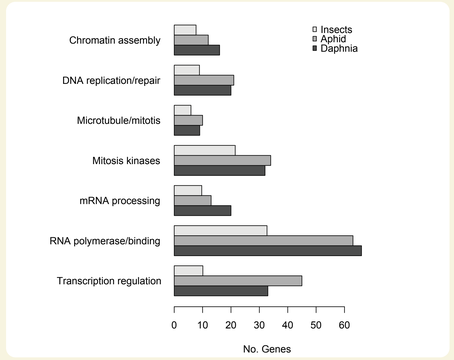

Fig. 3. Eukaryote chromosome maintenance genes trees. Multiple alignments of eukaryote forms for genes *Smc6*, *Top2* and *Topbp1* , are displayed as phylogenetic trees, and show duplicates in Aphid, *Daphnia* and *C. elegans.*

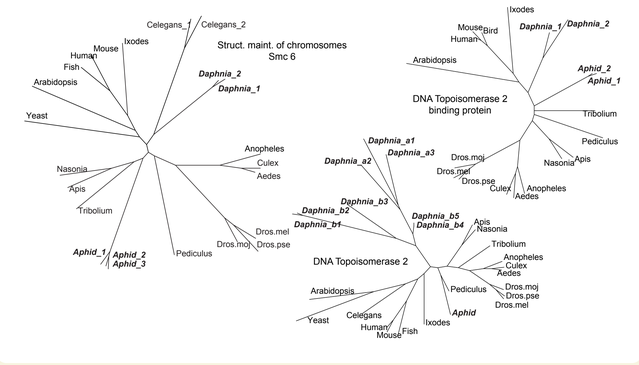

Table 2. Arthropod gene structure statistics

	Daphnia[1]	**Aphid**[1]	**Wasp**[1]	**Beetle**	**Mosquito**	**Fruitfly**[1]	**Worm**
Genome size	200	460	290	180	580	180	100
No. of genes	32,000	32,800	27,300	16,400	18,900	13,700	20,100
Gene density	0.175	0.063	0.120	0.100	0.055	0.168	0.250
Gene length	2,200	4,500	2,900	3,700	3,700	3,200	3,000
CDS size	1,320	1,200	1,510	1,370	1,280	1,650	1,300
Exons/gene	6.5	6.3	6.0	4.4	3.5	4.3	6.0
Exon size [2]	210	200	260	310	360	410	200
Intron size [3]	72	71/640	79/310	57/1200	65/1100	69/400	65/400
Intr > Exon	10%	41%	24%	31%	37%	27%	33%
Alternate Tr.	10%	24%	23%	--	--	36%	--

insects.eugenes.org/arthropods/data/summaries/ Updated 2009/06/08.

References

1. International Aphid Genome Consortium (2009) Genome sequence of the pea aphid *Acyrthosiphon pisum* : an insect dependent on host plants and symbiotic bacteria. PLoS Biology, Submitted.

2. Daphnia Genome Consortium (2009) The genome sequence of *Daphnia pulex.* in preparation.

3. Gilbert, D.G. 2009. Aphid and Waterflea have a High Rate of Gene Duplications Compared to Other Arthropods. PLoS ONE, provisionally accepted, May 2009.

4. Gilbert, D.G. 2008. Pea aphid genome annotation and analysis . http://insects.eugenes.org/apl

5. Gilbert, D.G. 2007. Daphnia gene duplicates. http://wfleabase.org/genome-summaries/