The Ecoresponsive Genome of *Daphnia pulex*
John K. Colbourne*, et al.*
Science **331**, 555 (2011);
DOI: 10.1126/science.1197761

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by clicking here.

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines here.

The following resources related to this article are available online at www.sciencemag.org (this infomation is current as of February 3, 2011):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/content/331/6017/555.full.html

Supporting Online Material can be found at:
http://www.sciencemag.org/content/suppl/2011/02/02/331.6017.555.DC1.html

This article **cites 19 articles**, 7 of which can be accessed free:
http://www.sciencemag.org/content/331/6017/555.full.html#ref-list-1

This article has been **cited by** 1 articles hosted by HighWire Press; see:
http://www.sciencemag.org/content/331/6017/555.full.html#related-urls

This article appears in the following **subject collections:**
Genetics
http://www.sciencemag.org/cgi/collection/genetics

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2011 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

The Ecoresponsive Genome of *Daphnia pulex*

John K. Colbourne,[1]* Michael E. Pfrender,[2]† Donald Gilbert,[1,3] W. Kelley Thomas,[4]
Abraham Tucker,[3,4] Todd H. Oakley,[5] Shinichi Tokishita,[6] Andrea Aerts,[7] Georg J. Arnold,[8]
Malay Kumar Basu,[9]‡ Darren J. Bauer,[4] Carla E. Cáceres,[10] Liran Carmel,[9]§ Claudio Casola,[3]
Jeong-Hyeon Choi,[1] John C. Detter,[7] Qunfeng Dong,[1]‖ Serge Dusheyko,[7] Brian D. Eads,[1,3]
Thomas Fröhlich,[8] Kerry A. Geiler-Samerotte,[5]¶ Daniel Gerlach,[11]# Phil Hatcher,[4]
Sanjuro Jogdeo,[4]** Jeroen Krijgsveld,[12]†† Evgenia V. Kriventseva,[11] Dietmar Kültz,[13]
Christian Laforsch,[14] Erika Lindquist,[7] Jacqueline Lopez,[1] J. Robert Manak,[15]‡‡ Jean Muller,[16]§§
Jasmyn Pangilinan,[7] Rupali P. Patwardhan,[1]‖‖ Samuel Pitluck,[7] Ellen J. Pritham,[17]
Andreas Rechtsteiner,[1]¶¶ Mina Rho,[18] Igor B. Rogozin,[9] Onur Sakarya,[5]## Asaf Salamov,[7]
Sarah Schaack,[3,17] Harris Shapiro,[7] Yasuhiro Shiga,[6] Courtney Skalitzky,[15] Zachary Smith,[1]
Alexander Souvorov,[9] Way Sung,[4] Zuojian Tang,[1]*** Dai Tsuchiya,[1] Hank Tu,[7]##
Harmjan Vos,[12]††† Mei Wang,[7] Yuri I. Wolf,[9] Hideo Yamagata,[6] Takuji Yamada,[16] Yuzhen Ye,[18]
Joseph R. Shaw,[1,19] Justen Andrews,[1,3] Teresa J. Crease,[20] Haixu Tang,[1,18] Susan M. Lucas,[7]
Hugh M. Robertson,[21] Peer Bork,[16] Eugene V. Koonin,[9] Evgeny M. Zdobnov,[11,22] Igor V. Grigoriev,[7]
Michael Lynch,[3] Jeffrey L. Boore[7,23,24]

We describe the draft genome of the microcrustacean *Daphnia pulex*, which is only 200 megabases and contains at least 30,907 genes. The high gene count is a consequence of an elevated rate of gene duplication resulting in tandem gene clusters. More than a third of *Daphnia*'s genes have no detectable homologs in any other available proteome, and the most amplified gene families are specific to the *Daphnia* lineage. The coexpansion of gene families interacting within metabolic pathways suggests that the maintenance of duplicated genes is not random, and the analysis of gene expression under different environmental conditions reveals that numerous paralogs acquire divergent expression patterns soon after duplication. *Daphnia*-specific genes, including many additional loci within sequenced regions that are otherwise devoid of annotations, are the most responsive genes to ecological challenges.

Daphnia pulex, or the water flea, is a keystone species of freshwater ecosystems: a principal grazer of algae, a primary forage for fish (*1*), and a sentinel of lentic (still water) inland ecosystems. Their populations are defined by the boundaries of ponds and lakes, are sensitive to modern toxicants in the environment, and thus are used to assess the ecological impact of environmental change (*2*, *3*). *Daphnia* exhibit a range of context-dependent development of specialized phenotypes, such as switching between clonal and sexual reproduction in response to environmental conditions (*4*). They are phenotypically plastic, in that some species alter diurnal migration behavior and develop exaggerated morphological defenses in response to predators (*5*). Physiological responses to abiotic environmental fluctuations can include the rapid rise of hemoglobin levels when ambient oxygen levels fall (*6*). The genus *Daphnia* is speciose, with multiple lineages independently colonizing and adapting to diverse habitats (*7*). Their short generation time, large brood sizes, and ease of laboratory and field manipulation have assured their importance for setting regulatory standards by environmental protection agencies, for testing chemical safety, for monitoring water quality (*2*, *3*), and as a model for ecological and evolutionary research (*8*).

Daphnia pulex is a crustacean arthropod, the group most closely allied with the insects (*9*), and thus allows the cataloging of genes that likely evolved in the pancrustacean ancestor of at least some lineages of insects and Crustacea (fig. S1). Although the branchiopod *D. pulex* represents only a single crustacean lineage—which contains more than 40,000 known species with striking levels of phenotypic diversity—the genus and its order (the Cladocera) date to the Permian (*10*).

Because *Daphnia*'s ecology is superbly understood, access to its genome sequence (fig. S2 and table S1) allows studying environmental influences on gene functions in ways that are difficult in even the best-developed genomic model species. Traits observed in laboratories are likely a small subset of the phenotypic variation that is expressed in natural ecosystems, and a focus on laboratory studies may partly explain why over 50% of many eukaryotic genomes are without experimentally determined functional annotations (*11*).

Genome sequence, assembly, and mapping to chromosomes. The *D. pulex* genome was assembled using JAZZ (*12*) from 1,554,564 quality-filtered nuclear sequence reads (8.7-fold coverage) from a naturally inbred isoclonal daphniid dubbed "The Chosen One" (TCO) [supporting online material (SOM) I.1]. The version 1.1 draft genome assembly comprises 19,008 contigs arranged within 5191 scaffolds that sum to a genome size of ~200 Mb (table S2). Two-hundred-eighty scaffolds link to construct 118 super-scaffolds (tables S3 and S4). Microsatellite markers (*13*) place 73 large scaffolds (73.9 Mb total) on the 12 chromosomes (table S5). We estimate that the draft assembly is high quality and includes ~80% of *Daphnia*'s nuclear genome (SOM I.2, tables S6 and S7, and figs. S3 to S5). We determine that 3598 missing regions (59%) contain duplicated genes, whereas others are heterochromatic regions, including the centromeres and telomeres. We estimate that 25% of the genome may be heterochromatic (table S8 and fig. S6). The ends of *D. pulex* chromosomes appear to consist of long stretches of TTAGG repeats with flanking regions (30 to 40 Kb) internal to these repeats consisting of repetitive sequences, including at least two kinds of satellite sequences (SOM 1).

Gene inventory. A minimum set of 30,907 protein-encoding genes was predicted for *D. pulex*, with 26,867 gene models having the following support (tables S9 to S14 and fig. S7): (i) 145,578 expressed sequence tags (ESTs) from 37 separate conditions validating 10,578 genes; (ii) whole-genome tiling microarrays examining gene expression under six different conditions that detect 186,269 transcriptionally active regions (TARs) validating 57,294 exons from 14,135 genes (additional TARs suggest gene models not yet included within the minimum set); (iii) similarity to proteins from other (non-daphniid) genomes that detect 19,641 *D. pulex* genes (blast $e < 10^{-5}$); (iv) 18,765 genes identified in protein similarity searches against a preliminary draft genome sequence for *D. magna* (SOM 2), which belongs to a separate subgenus (*7*); (v) more than 11,000 *D. pulex* peptide sequences detected by tandem mass spectrometry, of which 93% map to 1273 gene models in the minimum set; (vi) 716 highly conserved single-copy eukaryotic genes, of which *D. pulex* is missing only two (table S15), confirming that expected genes are included in the assembly; and (vii) 13,105 loci identified as paralogs by nucleotide sequence similarity searches for each predicted gene against the complete gene list ($e < 10^{-20}$). Measures of the relative rate of nonsynonymous nucleotide substitutions to the substitution rate at synonymous sites (K_a/K_s) indicate that the paralogs within our gene set generally show evidence of purifying selection (fig. S8).

To ensure that the gene count was not inflated by the erroneous assembly of alleles of the same locus as unique gene copies, we conducted comparative genomic hybridizations of labeled TCO DNA on microarrays. We detected no correlation between the read coverage and the mean fluorescing units of probes representing genes (fig. S9). Counts can also be inflated by inclusion of pseudogenes. However, manual annotations suggest that pseudogenes account for only 4 to 6% of large paralogous family members in *Daphnia* [see companion studies (*14*)].

Many non–protein-encoding genes were also identified in the *D. pulex* genome (SOM 3). Fifty microRNA genes are annotated, and 27 are validated using tiling microarrays (table S16 and fig.

S10). We estimate 468 ribosomal RNA loci and find 3798 transfer RNA genes. As in *Drosophila melanogaster* and *Caenorhabditis elegans*, these loci are clustered (fig. S11). Transposable elements constitute 9.4% of the assembled genome (table S17), consisting of 275 families of retrotransposons (Class I) and DNA transposons (Class II) (table S18). Intra-element pair-wise divergence among termini for intact elements of long terminal repeat retrotransposons ranges from 0 to 25.3% among the three superfamilies, BEL, gypsy, and copia (averaging 2%), indicating many recent transpositions (fig. S12).

Attributes of a compact genome. Comparison with gene-structure statistics for insects, nematode, and mouse, reveals reduced intron size in *Daphnia* (table S19 and fig. S13), resulting in a mean gene span of ~1000 base pairs (bp) shorter than the mean *Drosophila* gene length. However, average protein length is similar in these two species. Aside from introns, most other structures of the *D. pulex* genome are approximately equal in size or in number to those of the nematode, or exceed measurements in other species. The reduced intergenic regions compared with insects may partly be attributed to smaller repeated elements (table S19).

The average length of EST-validated *D. pulex* introns is 170 bp; only 10% of introns are larger than 210 bp. The intron density of *Daphnia pulex* genes is similar to that of *Apis mellifera,* having more than twice as many introns per gene as *Drosophila*. About 50% of introns are shared among respective orthologs in *Daphnia* and *Apis* (tables S19 to S23 and fig. S14). The *Daphnia* lineage shows an estimated intron gain/loss ratio substantially greater than 1 (table S24 and fig. S15). We estimate that 78% of these intron gains are unique to this lineage and that 22% occurred in parallel with gains in other lineages (fig. S16).

Origin and preservation of *Daphnia pulex* genes. *Daphnia*'s gene catalog shows more universal bilaterian genes than other arthropods (8096; black in Fig. 1A) and thus shares the highest number of genes with human (table S25). Only 1383 genes (4.5%) appear pancrustacean (green in Fig. 1A). Remarkably, over 36% of the minimal set of *D. pulex* genes have no detectable homology to those in the other species (Fig. 1A), which can partly be explained by the disproportionate expansion of gene families distinctive to this crustacean lineage ($\chi^2 = 450.55$, $P < 0.0001$) (table S26 and Fig. 1B) and fast divergence for some genes (enlarged beige fraction in Fig. 1A). A phylogenetic accounting of the expansions and contractions of all gene families within pancrustacean and representative deuterostome genomes (tables S27 and S28) suggests a net increase in the number of paralogs within the lineage leading to *Daphnia* (Fig. 1C). By reconstructing gene family histories across a phylogeny (SOM IV.2), we count 17,424 new and 1079 lost genes in the branch leading to *Daphnia*. By contrast, the sum of inferred gains and loss along the longest series of branches in the insect phylogeny—originating from the shared pancrustacean ancestor with *Daphnia*—only reaches 8981 gained loci with 3040 gene losses. Therefore, the overall elevated *Daphnia* gene count appears to result from both gaining and retaining more genes.

To better understand gene duplication in the *Daphnia* genome, we examined the age distribution of gene duplicates by estimating K_s for 66,502 pair-wise combinations of paralogs showing >40% sequence similarity and by comparing this distribution to that of 12,570 nematode and 64,783 human gene pairs (Fig. 1D). The single-pair duplicates within the youngest cohort ($K_s < 0.01$) suggest that *D. pulex* genes duplicate at a rate three times as high as those measured for fly and nematode and 30% greater than human, even when we exclude nearly identical gene copies that may be biased by gene conversion (table S29 and figs. S17 and S18).

In the genomes of many species, new duplicate genes are found in clusters (fig. S19) (*15*). The *D. pulex* genome shows ~20% (table S30) of all genes tightly arranged in clusters of 3 to 80 paralogs and with elevated numbers of tandemly duplicated genes at intervening intervals of 1000 to 2000 bp (fig. S20). The age distribution and positioning of gene duplicates indicate that *Daphnia* has not experienced whole-genome duplication, but the genome is instead characterized by a high and historically steady rate of tandem duplication (Fig. 1D).

Nine gene families have expanded independently in *Daphnia* and other aquatic lineages, including vertebrates (tables S26 and S31). These include photoreactive or photoresponsive gene families (cryptochromes, opsins, and G proteins). The *D. pulex* genome shows 46 opsins (table S32 and figs. S21 and S22), of which 42 derive from two rhabdomeric subfamilies, one ciliary pteropsin subfamily, and a newly discovered lineage that forms a sister group to rhabdomeric opsins that we have named arthropsins (SOM 4). Arthropsins are ancestral to the chordate melanopsin lineage and thus appear to have been retained in *Daphnia*, despite their loss from all other available bilateral animal genomes. The expansion of these gene families suggests that adaptations to a more com-

---

[1]Center for Genomics and Bioinformatics, Indiana University, 915 East Third Street, Bloomington, IN 47405, USA. [2]Department of Biology, Utah State University, 5305 Old Main Hill Road, Logan, UT 84322–5205, USA. [3]Department of Biology, Indiana University, 1001 East Third Street, Bloomington, IN 47405, USA. [4]Hubbard Center for Genome Studies, University of New Hampshire, 35 Colovos Road, Durham, NH 03824, USA. [5]Department of Ecology, Evolution, and Marine Biology, University of California–Santa Barbara, Santa Barbara, CA 93106, USA. [6]Laboratory of Environmental and Molecular Biology, Environmental Sciences Division, School of Life Sciences, Tokyo University of Pharmacy and Life Sciences, 1432-1 Horinouchi, Hachiouji, Tokyo, 192-0392, Japan. [7]Department of Energy Joint Genome Institute (JGI), 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. [8]Laboratory for Functional Genome Analysis (LAFUGA), Gene Center, Ludwig-Maximilians-Universität München, Germany. [9]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA. [10]School of Integrative Biology, University of Illinois, 515 Morril Hall, Urbana, IL 61801, USA. [11]University of Geneva Medical School and Swiss Institute of Bioinformatics, 1 rue Michel-Servet, 1211 Geneva, Switzerland. [12]Biomolecular Mass Spectrometry and Proteomics Group, Bijvoet Centre for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, and Netherlands Proteomics Center, Utrecht University, Sorbonnelaan 16, 3584 CA, Utrecht, Netherlands. [13]Department of Animal Science, University of California, Davis, Meyer Hall, One Shields Avenue, Davis, CA 95616, USA. [14]Department of Biology II and GeoBio Center Munich, Ludwig-Maximilians-University Munich, Großhadernerstrasse 2, 82152 Planegg-Martinsried, Germany. [15]Gene Expression, Roche NimbleGen Inc., 504 South Rosa Rd, Madison WI 53719, USA. [16]Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL) Heidelberg, Meyerhofstrasse 1, D-69117 Heidelberg, Germany. [17]Department of Biology, University of Texas, Arlington, Box 19498, Arlington, TX 76019, USA. [18]School of Informatics and Computing, Indiana University, Informatics Building, 901 East Tenth Street, Bloomington, IN 47408–3912, USA. [19]School of Public and Environmental Affairs, Indiana University, 1315 East Tenth Street, Bloomington, IN 47405, USA. [20]Department of Integrative Biology, University of Guelph, Guelph, Ontario N1G 2W1 Canada. [21]Department of Entomology, University of Illinois at Urbana-Champaign, 505 South Goodwin Avenue, Urbana, IL 61801, USA. [22]Imperial College London, South Kensington Campus, SW7 2AZ London, UK. [23]Genome Project Solutions, 1024 Promenade Street, Hercules, CA 94547, USA. [24]Department of Integrative Biology, University of California, Berkeley, CA 94720, USA.

*To whom correspondence should be addressed. E-mail: jcolbour@indiana.edu

†Present address: Department of Biological Sciences, University of Notre Dame, 109B Galvin Life Sciences, Notre Dame, IN 46556, USA.

‡Present address: J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA.

§Present address: Department of Genetics, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram, Jerusalem 91904, Israel.

‖Present address: Departments of Biological Sciences and Computer Science and Engineering, University of North Texas, Denton, TX 76203, USA.

¶Present address: Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA.

#Present address: Research Institute of Molecular Pathology (IMP), Dr. Bohr-Gasse 7, A-1030 Vienna, Austria.

**Present address: Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR 97331, USA.

††Present address: Genome Biology Unit, EMBL Heidelberg, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

‡‡Present address: Department of Biology and Roy J. Carver Center for Genomics, University of Iowa, Iowa City, IA 52242, USA.

§§Present address: Laboratoire de Diagnostic Génétique, CHU Strasbourg Nouvel Hôpital Civil, 1 place de l'hôpital, 67000 Strasbourg, France, and IGBMC (Institut de Génétique et de Biologie Moléculaire et Cellulaire), CNRS/INSERM/Université de Strasbourg, 67404 Illkirch cedex, France.

‖‖Present address: Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA.

¶¶Present address: Department of Molecular, Cell, and Developmental Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA.

##Present address: Life Technologies Corporation, Foster City, CA 94404, USA.

***Present address: Center for Health Informatics and Bioinformatics, New York University Langone Medical Center, 333 East 38th Street, New York, NY 10016, USA.

†††Present address: Department of Molecular Cancer Research, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, Netherlands.

plex light regime in aquatic environments (*16*, *17*) can be influential in shaping the gene content of these organisms.

Implications of *Daphnia*'s genome structure. Tandemly duplicated gene clusters are predisposed to homogenization by gene conversion and unequal crossing-over (*18*). If common, concerted evolution can maintain sequence and functional similarities among paralogs. We examined copied DNA segments among all paralogs in the genome (SOM V.1) and observed that 47% of the genes show tracts of nonallelic gene conversion compared with 12 to 18% of genes in five *Drosophila* species (tables S33 to S38 and figs. S23 and S24). Thus, concerted evolution is affecting more than 1 Mb (8%) of all protein-coding sequences in *Daphnia*, especially when duplicates are oriented on the same strand, with a similar conversion rate (converted pairs of paralogs/total pairs of paralogs analyzed) and number of events per pair as *Drosophila*. The greater proportion of converted genes in *D. pulex* is mainly attributed to the greater number of targets for gene conversion within the genome, including tandemly duplicated gene clusters with intervening genes. Conversion events in *Daphnia* are less common among the youngest duplicates and within gene families containing only two paralogs.

One example of widespread gene conversion is found in the di-domain hemoglobin genes. Hemoglobin levels in the hemolymph of daphniids can rise by more than one order of magnitude in response to reduced oxygen availability in aquatic habitats, which fluctuates in diurnal and seasonal cycles (Fig. 2A). In *Daphnia*, a tandemly duplicated gene cluster of hemoglobin (Hb) genes contributes to the protein's varying composition (*19*). We sequenced and assembled the full *D. magna* cluster to compare with the arrangement of eight clustered *D. pulex* hemoglobin genes (figs. S25 and S27). (*D. pulex* also has three nonclustered Hb genes.) Notably, the two species show almost identical gene arrangements within an interval of ~23.5 Kb (table S39) except for the obvious absence of Hb6 from the *D. magna* cluster (Fig. 2B). In both species, a noncoding RNA gene interrupts the cluster between Hb4 and Hb5, and hypoxia response elements plus ancillary sequences are preserved within the regulatory regions of each gene. Thus, the duplication and subsequent divergence of hemoglobins must have occurred before the divergence time of *D. pulex* and *D. magna*.

However, a phylogenetic analysis of protein-coding sequences (SOM V.2) suggests that most hemoglobin genes have duplicated independently within each species (Fig. 2C). A separate phylogenetic reconstruction using sequences from intergenic regions recovers a tree that is consistent with duplication before speciation (Fig. 2D). Because the support values at nodes for both trees are equally strong, we conclude that gene conversion tracts are homogenizing the protein coding regions. The hemoglobin gene clusters in both species are homologs because of ancestral gene duplications, yet the duplication history of

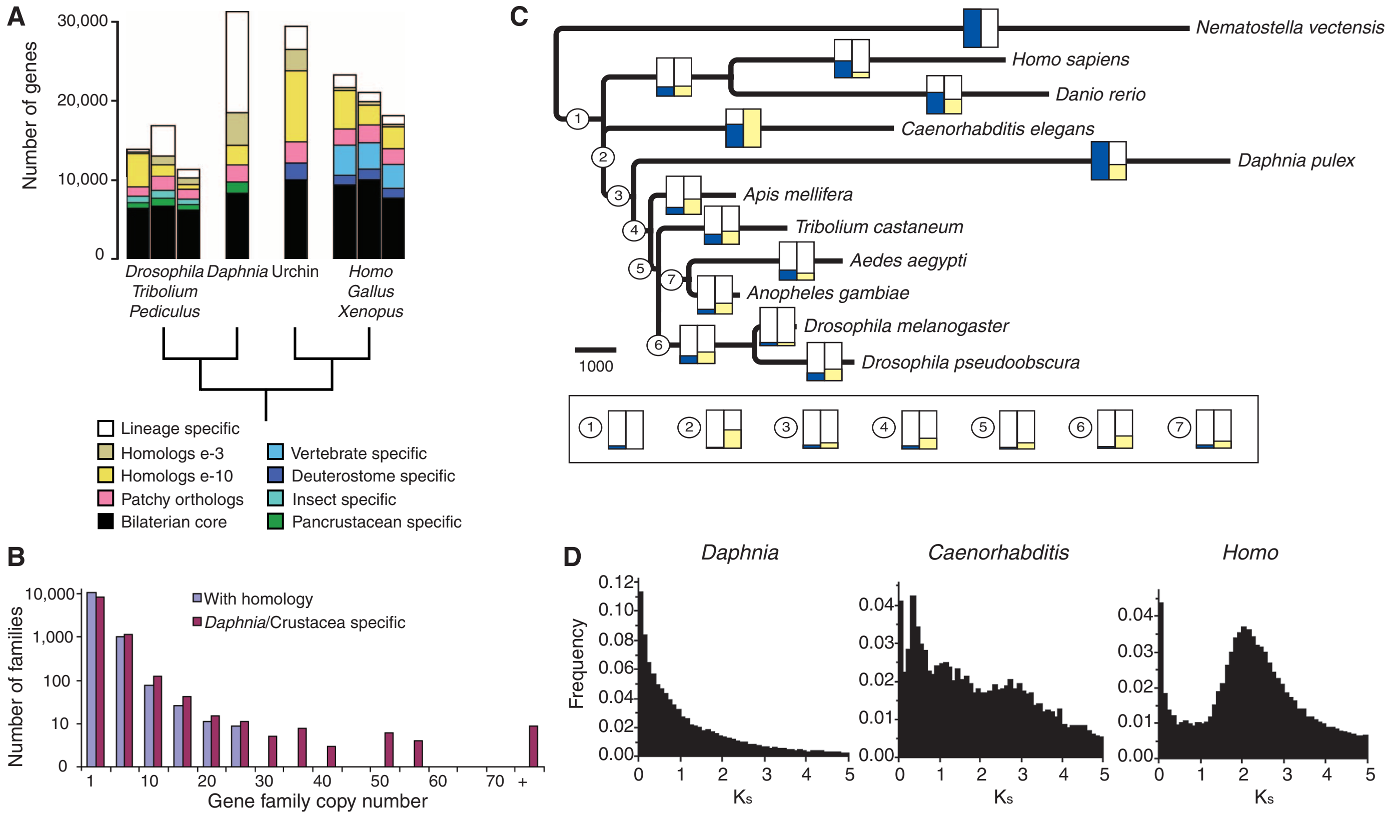


Fig. 1. The *Daphnia pulex* gene repertoire. (**A**) Comparison of genes among *D. pulex*, *Drosophila melanogaster*, *Pediculus humanus*, *Tribolium castaneum*, and *Strongylocentrotus purpuratus* (urchin), and *Gallus gallus*, *Xenopus tropicalis*, and *Homo sapiens*, showing the core bilaterian genes (black), vertebrate (blue), insect (aqua), and pancrustacean (green) specific genes, patchy or ancient orthologs present in at least one arthropod and one deuterostome genome but lost in other lineages (pink), multiple copy homologs (yellow and beige), and species-specific genes (white). (**B**) Distribution of *D. pulex* gene family sizes comparing genes with and without detectable homology to other genomes. (**C**) History of gene family expansions and losses among pancrustacean plus representative deuterostome genomes with the outgroup *Nematostella vectensis*. Tree topology is fixed from the assumed species phylogeny and used to map gene family histories by a combination of gene similarity and character-state optimization with Dollo parsimony (SOM IV.2). Branch lengths scaled to differences between inferred gene gains and losses. Scale bar corresponds to 1000 genes gained. Gene gains along each branch of the tree are scaled by the maximum value along the branch leading to *D. pulex* (blue); gene losses along each branch are scaled by the maximum loss along the branch leading to *Caenorhabditis elegans* (yellow). (**D**) Frequency of pair-wise genetic divergence at silent sites (K_s) among all gene duplicates in the *D. pulex*, *C. elegans*, and *H. sapiens* genomes, for genes with >100 aligned amino acids and percent identity >40% (66,502, 12,570, and 64,783 pair-wise comparisons for the three genomes, respectively). The vertical axis differs for *D. pulex*.

genes is obfuscated by independent gene conversions facilitated by their ordered arrangement in the genomes.

Evolutionary diversification of duplicated genes. Gene duplication is an important source of evolutionary novelty. After duplication, one copy is commonly disabled by mutation and becomes a pseudogene. This fate may be avoided if selection maintains both copies through gene dosage, novel function, or by subdividing the gene's original function into multiple components (*20*). We conducted microarray experiments to determine the magnitude of functional divergence among paralogs, then traced (*21*) and tested (*22*) whether their patterns of gene transcription differ in 1 to 12 ecologically relevant conditions as a function of K_s (table S40 and SOM VI.1). As expected, many recent duplicates (K_s < 0.05) have indistinguishable gene expression patterns for the tested conditions (47%) (Fig. 3A). Within many gene families, divergence in expression patterns correlates with age (figs. S28 and S29). We found that long-wavelength opsins most similar in sequence have the same expression patterns (correlation > 0.9) but then diverge in their response to shared conditions as they age, at an estimated rate of 0.6% per 10% synonymous nucleotide substitutions. A similar pattern is observed for the di-domain hemoglobins, albeit with more rapid divergence in expression.

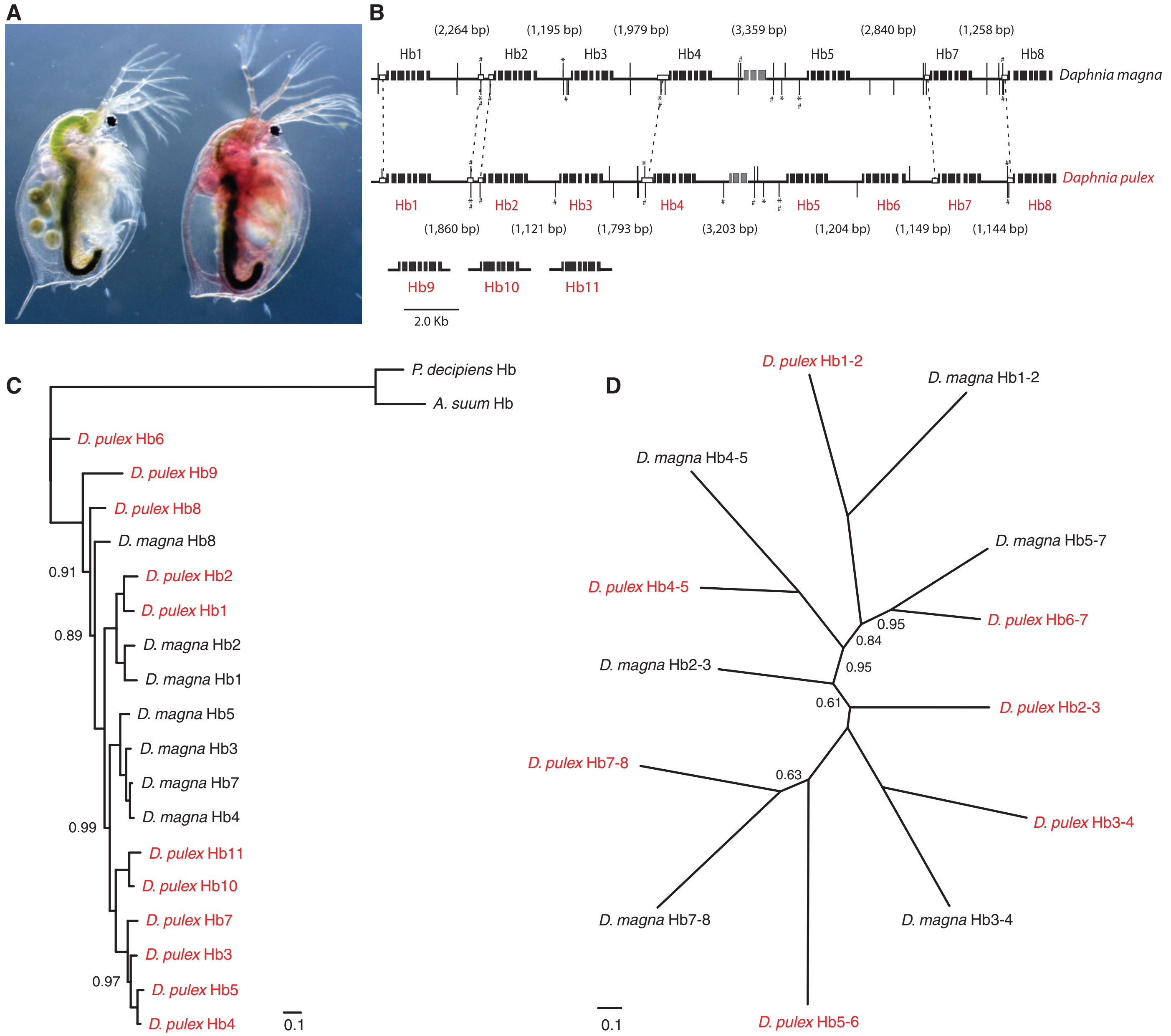


Fig. 2. Evolution of *Daphnia* di-domain hemoglobin (Hb) genes. (**A**) When deprived of oxygen, many species (here *D. magna*) increase hemoglobin concentration in the hemolymph by a factor of 15 to 20 within a single molting, coloring the body red. (**B**) Organization of the Hb gene cluster in the *D. magna* and the *D. pulex* genomes. Black boxes are exons. Gray boxes represent an RNA gene. Vertical bars are hypoxia response elements (HREs), and asterisks show ancillary elements. Conserved HREs are linked by hatches. Open boxes represent highly similar sequences. The lengths of intergenic regions are shown in parentheses. *Daphnia pulex* genes Hb9 to Hb11 are located on separate sequence scaffolds. (**C**) Phylogenetic tree (SOM V.2) from nucleotide sequences of Hb genes in *D. pulex* (red) and in *D. magna* (black). Outgroup Hb cDNA sequences are from *Ascaris suum* and *Pseudoterranova decipiens*. Scale bar shows mean number of differences (0.1) per nucleotide along each branch. Posterior probability node support <100% are shown. (**D**) Phylogenetic tree based on nucleotide sequences of intergenic regions between the stop codons and the downstream TATA of the neighboring gene. Posterior probabilities <100% are shown.

In contrast to the steady expression divergence of many duplicates, we observed an equally large fraction of recently arisen paralogs—with nearly identical sequences—that differ in their expression in at least one condition (Fig. 3A). Although we could confidently detect locus-specific expression for only a fraction of the youngest duplicates represented on the microarray (table S40), a plot of the maximum difference in the expression response of paralogs to an identical condition suggests that, on average, newly duplicated genes may differ in expression by as much as a factor of 1.9 (Fig. 3B). These may be cases in which new regulatory programs were created by the gene duplication itself through a failure to copy regulatory elements or when a duplicate is integrated within a new genomic location (*23*).

Gene conversion, homogenizing nonregulatory nucleotide sequences, can contribute to this class of highly differentially expressed (DE) paralogs at low sequence divergence (K_s). We tested whether gene conversion accounts for the differences in the evolutionary rates of expression divergence by comparing duplicates ($K_s < 2$) on the basis of their structural arrangements in the genome (SOM VI.2). Neighboring paralogs within tandem gene clusters were just as likely to diverge in expression as dispersed duplicates outside of clusters ($\chi^2 = 0.027$, $P = 0.87$). Globally, gene conversion reduces the expression-level divergence of paralogs ($\chi^2 = 11.9$, $P = 0.0005$) (table S41), yet we detected no significant impact on the observed fractions of divergently expressed paralogs when we removed duplicated genes with signatures of gene conversion (table S42). Although adjacent genes are often coexpressed (*24*), the local placement of genes within tandem gene clusters has no clear effect on gene expression divergence in *D. pulex*. We thus conclude that paralogs, even in tandem, frequently acquire divergent expression patterns at, or soon after, the time of duplication.

Functional importance of expanded gene families. To investigate the functional role of paralogs and their preservation, we examined interacting genes with known function. A total of 1908 genes representing 563 enzymes were charted onto the global metabolic pathway for *D. pulex* by referencing the metabolic enzyme networks of three insects and four vertebrates (Fig. 4) (SOM VII.1). Of these, 38 gene families were amplified in Pancrustacea, and 32 are expanded in the lineage leading to *Daphnia* (figs. S30 and S31 and tables S43 and S44). Half (19 of 38) of the amplified genes are nonrandomly clustered within seven distinct pathways ($P < 0.03$ by exact binominal test and $P < 0.03$ by network permutation analysis) (Fig. 4, A to G, and fig. S32). These data, showing coexpansion of genes within pathways, suggest that duplicated genes can be interdependent.

A study of the expression patterns of duplicated genes from this metabolic network (SOM VII.2) reveals greater average similarity between genes from coexpanding and interacting families (same Kyoto Encyclopedia of Genes and Genomes map ID in table S43) than between genes from nonassociating families ($t = 3.30$, $P = 0.0025$). This pattern suggests nonindependent functional divergence of expanding genes within pathways (e.g., tables S45 to S48 and figs. S33 and S34). One example involves nine phylogroups of fucosyltransferase paralogs that share 95% amino acid similarity (colored lines in fig. S35) and have independently diversified to express seven transcriptional profiles shared with interacting glycosyltransferase paralogs. Such a pattern of codivergence suggests a decoupling of duplication history and functional association. To test this prediction, we estimated the ratio of among-group variance to total variance in differential expression (D_{st}) for phylogroups of fucosyltransferase paralogs and for expression profile clusters (SOM VII.2). We detect no significant subdivision of expression patterns for fucosyltransferase paralogs based on phylogeny (blue nodes in fig. S35; $D_{st} = 0.0042$, $P = 0.89$). By contrast, clusters based on transcriptional profiles, and including distantly related paralogs and interacting glycosyl transferase paralogs, show significant subdivision ($D_{st} = 0.0836$, $P = 0.002$).

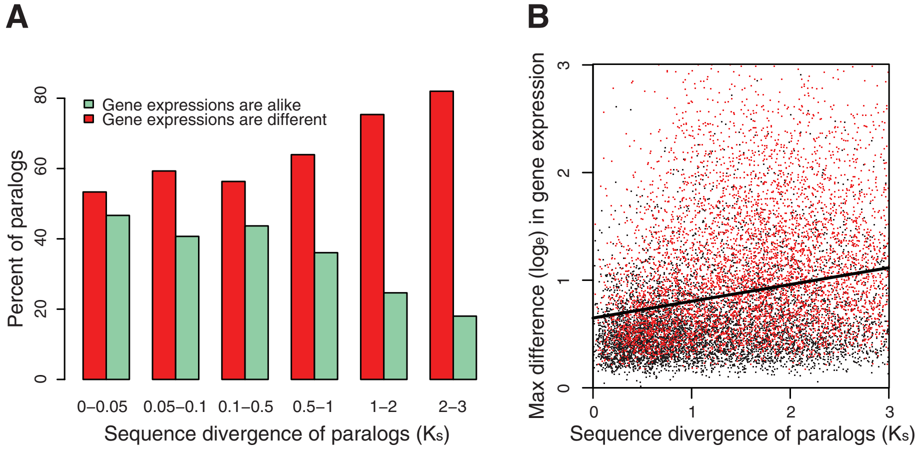


Fig. 3. Functional diversification of duplicated genes from 12 microarray experiments. (**A**) The fraction of duplicated genes with similar versus divergent DE patterns as a function of their pair-wise divergence at silent sites (K_s). (**B**) Regression ($r = 0.29$) of the maximum observed difference (treatment versus control) between duplicated genes among the 12 conditions as a function of the age of duplicated genes inferred from K_s. Red points are significant values ($P < 0.05$, analysis of variance). The regression line *y*-axis intercept (ln 0.642 ± 0.009) suggests that, on average, newly duplicated genes may differ in expression by as much as a factor of 1.9 at particular conditions, which is significantly different from zero ($t = 68.7$, $P < 2 \times e^{-16}$) and validated by tiling microarray data ($r = 0.16$; $t = 75.3$, $P < 2 \times e^{-16}$).

Ecoresponsive genes. The *D. pulex* genome contains many duplicated genes with unknown homology. Although this may diminish with the availability of more crustacean genomes, these unknown genes appear to play important roles in the animal's ecology. ESTs from 37 cDNA libraries representing transcriptomes of daphniids exposed to biotic ecological challenges, abiotic ecological stressors, and different life-history stages in laboratory environments (table S10) show that genes unique to the *Daphnia* lineage, and genes that reside within tandemly duplicated gene clusters, are significantly over-represented within transcriptomes under ecological conditions (Fig. 5A and table S49; $\chi^2 = 265.1$, $P = 2.66 \times e^{-58}$ and $\chi^2 = 41.0$, $P = 1.23 \times e^{-09}$, respectively). Whole-genome tiling-expression microarray experiments show differential expression to be twice as frequent in genomic regions devoid of gene models (intergenic) when *D. pulex* are exposed to environmental challenges compared with conditions of life history (table S50 and fig. S36).

We count 34,844 transcriptionally active regions (TARs) within unannotated regions of the genome, showing predictable exon-intron intervals supporting additional gene models not yet included within the minimum set (TAR-genes) (table S12) and that are condition-dependent in their regulation. By partitioning the differentially expressed genome by experimental conditions, between 72% and 85% of the transcriptome uniquely responded to one of the three conditions (Fig. 5B). In all, 73% of differential regulation under biotic or abiotic stressors requires additional gene models or extensions.

Evolutionary perspectives. *Daphnia pulex* paralogs follow different evolutionary trajectories that are determined, in part, by their initial transcriptional expression patterns. At least half appear to acquire divergent expression patterns at or near the time of origin. Interacting and coexpanding genes can also appear to be codiverging in their responses to environmental conditions. These observations suggest that the persistence of this distinctive class of functionally divergent gene duplicates is due to preservation by entrainment (PBE). Entrainment is defined as the process of increasing the initial probability of preserving a duplicated gene through its functional interaction with existing or newly interacting genes sharing regulatory programs. Because biological processes can be governed by interdependent regulation of interacting genes, there are three likely evolutionary

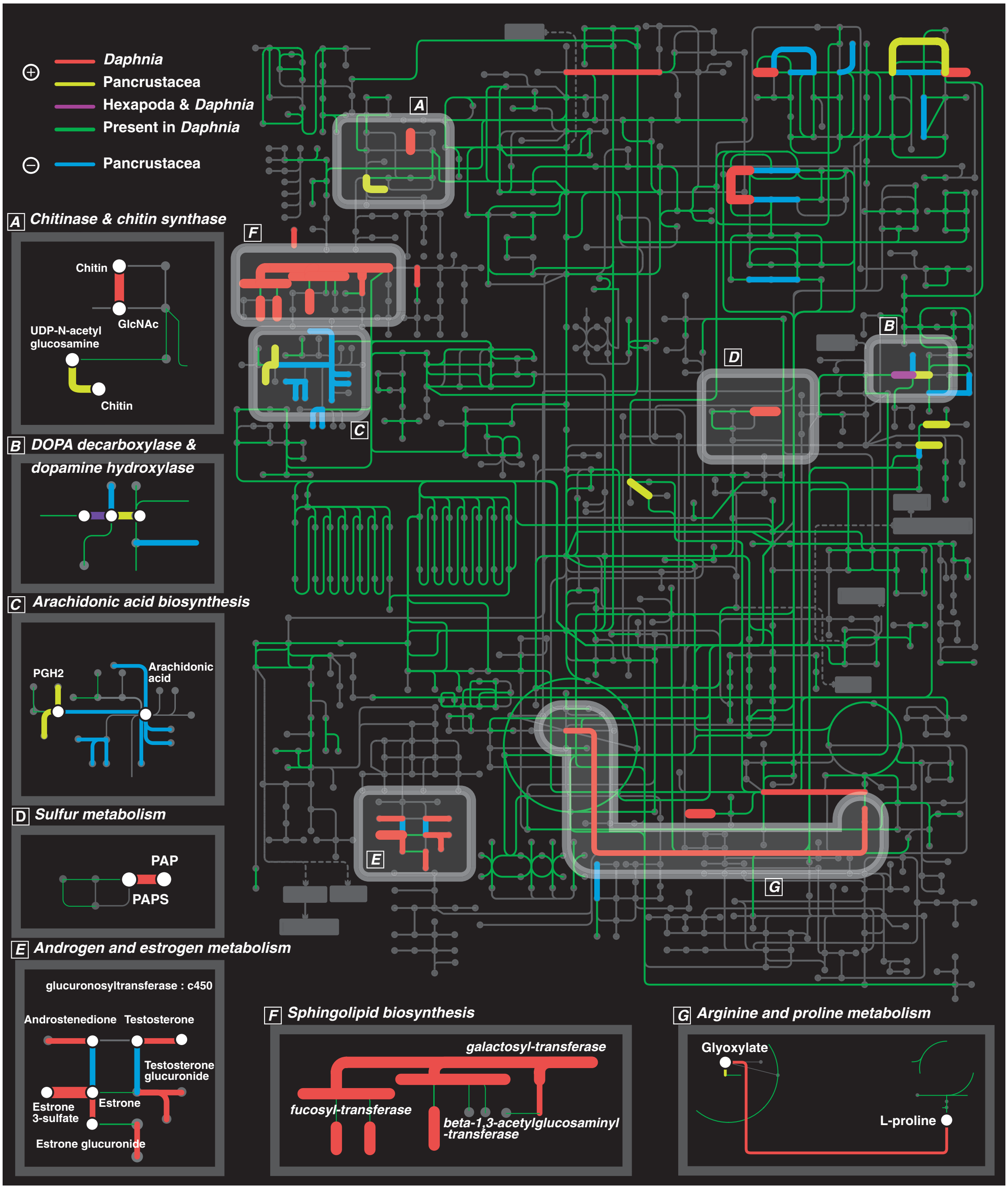

⊕
Daphnia
Pancrustacea
Hexapoda & Daphnia
Present in Daphnia
⊖
Pancrustacea
A
Chitinase & chitin synthase
Chitin
GlcNAc
UDP-N-acetyl glucosamine
Chitin
B
DOPA decarboxylase & dopamine hydroxylase
C
Arachidonic acid biosynthesis
PGH2
Arachidonic acid
D
Sulfur metabolism
PAP
PAPS
E
Androgen and estrogen metabolism
glucuronosyltransferase : c450
Androstenedione
Testosterone
Testosterone glucuronide
Estrone 3-sulfate
Estrone
Estrone glucuronide
F
Sphingolipid biosynthesis
galactosyl-transferase
fucosyl-transferase
beta-1,3-acetylglucosaminyl -transferase
G
Arginine and proline metabolism
Glyoxylate
L-proline

Fig. 4. Map of global KEGG metabolic pathway in *D. pulex* showing significantly expanded or contracted gene families in metabolic pathways. Nodes and edges represent compounds and enzymes, respectively. Expanded gene families in *D. pulex* (red); expanded gene families in Pancrustacea (yellow); independently expanded gene families in *D. pulex* and in insects (purple); contracted gene families in Pancrustacea (blue); and genes present in *D. pulex* (green). Amplification of gene families encoding each highlighted enzyme is supported by the Fisher exact test (thick edges are supported by Bonferroni correction), on the basis of the distribution of the number of genes encoding corresponding enzymes among *Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Tetraodon nigroviridis*, *Drosophila melanogaster*, *Apis mellifera*, and *Anopheles gambiae*. Emphasized pathways (A to G) include at least two cases of expanded interacting enzymes. The nonrandom coexpansion of interacting enzymes is supported by exact binomial test ($P < 0.03$) and by the node permutation test on 1000 randomized metabolic networks ($P < 0.03$).

outcomes for these interacting duplicated genes (Fig. 6). Genes with expression patterns unchanged at the time of duplication may continue to share the condition-specific regulation of existing interacting genes (Fig. 6A). In this scenario, selection for gene dosage may increase the probability that gene duplicates are preserved (*25*). Alternatively, duplicates may initially have divergent expression patterns but have inappropriate transcriptional responses to environmental conditions or lack appropriately coregulated interacting genes (Fig. 6B). Duplicates within this category are most likely lost. In contrast, genes with divergent expression patterns at the time of duplication, yet with regulation sufficiently similar to the expression patterns of a different interacting gene, may have combined products that are beneficial under a distinct environmental condition (Fig. 6C). In this scenario, the likelihood for preservation of these new gene duplicates is increased. Thus, when genes are advantageous at the time of duplication, their coding regions are subject to purifying selection from the start and are entrained to a distinct regulatory pattern dictated by condition-specific gene-gene interactions. Although the likelihood of converging on a beneficial gene expression profile near the time of duplication is very small, in the case of *Daphnia*, PBE is facilitated by the high rate of gene duplication, resulting in co-regulated interacting genes that can potentially define environment-specific transcriptomes, which may increase with the complexity of interactions between organisms and their environments.

In conclusion, by examining genome structure and the functional responses of genes to environmental conditions within species with tractable ecologies, we further our understanding of gene-environment interactions in an evolutionary context. Many responsive genes to ecological conditions have unknown function, and information from laboratory model species may be insuf-

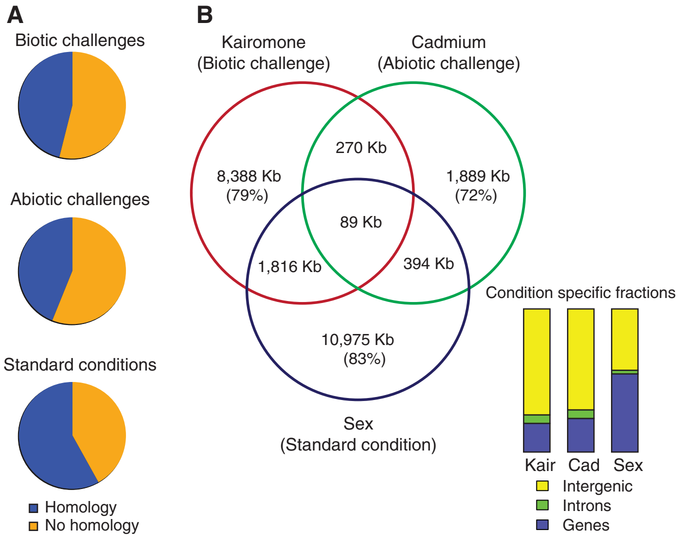
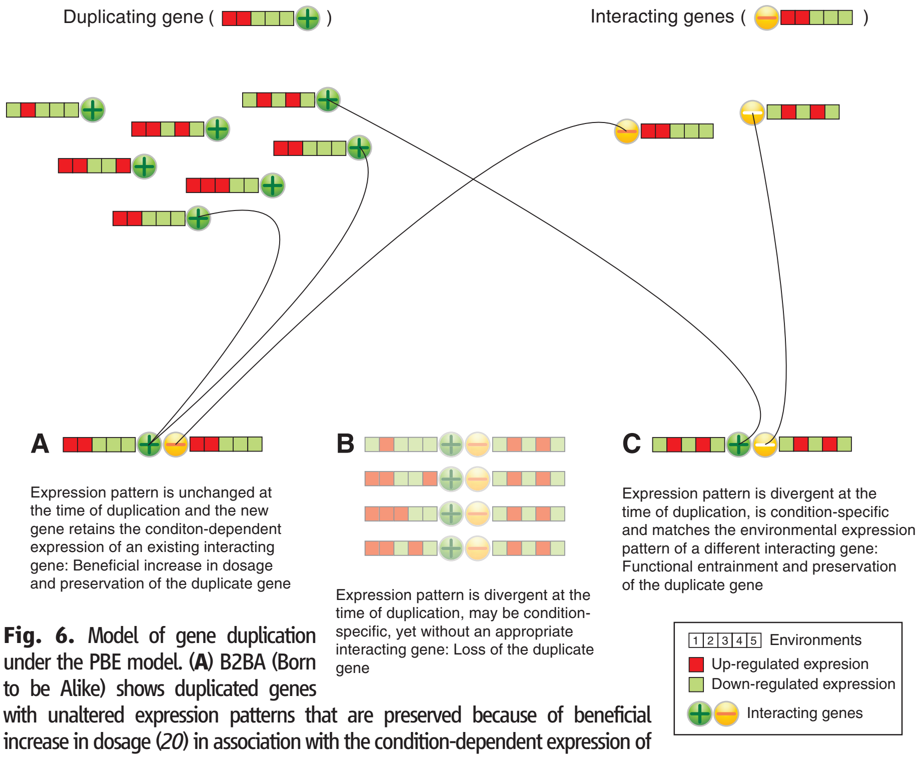


Fig. 5. Function of genes unique to the *D. pulex* lineage. (**A**) Pie charts show the distribution of expressed genes both with and without detectable homology to other sequenced genomes, sampled under exposure to bacterial infection, predators, hormones, varying diets (biotic challenges), environmental toxicants, elevated ultraviolet radiation, hypoxia, acid, salinity, and calcium starvation (abiotic challenges), in addition to various stages of life history within a controlled laboratory environment (standard conditions). (**B**) Differential expression of the genome upon exposure to *Chaoborus* kairomone (Kair), cadmium (Cad), and by sex, measured as nucleotides in kilobases (Kb) on genome tiling microarrays. Comparing three experimental conditions, 79%, 72%, and 83% of transcriptomes are condition-specific (Venn diagram) and twice as pronounced in genomic regions that are currently void of gene models (yellow) when *D. pulex* are exposed to ecological conditions.



Expression pattern is unchanged at the time of duplication and the new gene retains the conditon-dependent expression of an existing interacting gene: Beneficial increase in dosage and preservation of the duplicate gene

Expression pattern is divergent at the time of duplication, may be condition-specific, yet without an appropriate interacting gene: Loss of the duplicate gene

Expression pattern is divergent at the time of duplication, is condition-specific and matches the environmental expression pattern of a different interacting gene: Functional entrainment and preservation of the duplicate gene

Fig. 6. Model of gene duplication under the PBE model. (**A**) B2BA (Born to be Alike) shows duplicated genes with unaltered expression patterns that are preserved because of beneficial increase in dosage (*20*) in association with the condition-dependent expression of an interacting gene. (**B**) B2BU (Born to be Useless) genes with initially divergent expression patterns and with inappropriate condition-dependent responses or interacting genes are most likely lost. (**C**) B2BD (Born to be Different). When the derived expression pattern of a paralog at the time of duplication is shared with a different interacting gene (white negative sign), and when the effect of their combined products is beneficial under a distinct environmental condition, the likelihood for preservation is increased. Color-coding represents condition-dependent expression patterns across multiple environments. Lines represent the process of functional entrainment.

ficient because of a lack of homology or experimentally demonstrated functions in response to the environment. Thus, ecological genomics requires empirical annotations of new genome sequences from a broader diversity of species, tested under a variety of natural conditions.

References and Notes

1. S. R. Carpenter *et al.*, *Ecology* **68**, 1863 (1987).
2. J. R. Shaw *et al.*, in *Advances in Experimental Biology on Toxicogenomics,* C. Hogstrand, P. Kille, Eds. (Elsevier Press, 2008), pp. 165–219.
3. J. Martins, L. Oliva Teles, V. Vasconcelos, *Environ. Int.* **33**, 414 (2007).
4. P. D. N. Hebert, in *Daphnia,* R. H. Peters, R. de Bernardi, Eds. (Memorie dell'Istituto Italiano di Idrobiologia, 1987), vol. 45, pp. 439–460.
5. R. Tollrian, S. I. Dodson, in *The Ecology and Evolution of Inducible Defenses,* R. Tollrian, C. D. Harvell, Eds. (Princeton Univ. Press, Princeton, NJ, 1999), pp. 177–202.
6. B. Zeis, S. Schwerin, R. Pirow, T. Lamkemeyer, R. J. Paul, *Comp. Biochem. Physiol. A* **151**, S38 (2008).
7. J. K. Colbourne, P. D. N. Hebert, D. J. Taylor, in *Molecular Evolution and Adaptive Radiation,* T. J. Givnish, K. J. Sytsma, Eds. (Cambridge University Press, Cambridge, 1997), pp. 163–188.
8. M. Lynch, K. Sptize, in *Ecological Genetics,* L. Real, Ed. (Princeton Univ. Press, Princeton, NJ, 1994), pp. 109–128.
9. J. L. Boore, D. V. Lavrov, W. M. Brown, *Nature* **392**, 667 (1998).
10. D. J. Taylor, T. J. Crease, W. M. Brown, *Proc. R. Soc. London B Biol.* **266**, 791 (1999).
11. L. Peña-Castillo, T. R. Hughes, *Genetics* **176**, 7 (2007).
12. S. Aparicio *et al.*, *Science* **297**, 1301 (2002).
13. M. E. A. Cristescu, J. K. Colbourne, J. Radivojac, M. Lynch, *Genomics* **88**, 415 (2006).
14. www.biomedcentral.com/series/Daphnia.
15. Q. Zhou *et al.*, *Genome Res.* **18**, 1446 (2008).
16. U. C. Storz, R. J. Paul, *J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol.* **183**, 709 (1998).
17. F. Y. Wang, H. Y. Yan, J. S. C. Chen, T. Y. Wang, D. Y. Wang, *Vision Res.* **49**, 1860 (2009).
18. F. G. Hoffmann, J. C. Opazo, J. F. Storz, *Mol. Biol. Evol.* **25**, 591 (2008).
19. S. Kimura, S. Tokishita, T. Ohta, M. Kobayashi, H. Yamagata, *J. Biol. Chem.* **274**, 10649 (1999).
20. H. Innan, F. Kondrashov, *Nat. Rev. Genet.* **11**, 97 (2010).
21. T. Casneuf, S. De Bodt, J. Raes, S. Maere, Y. Van de Peer, *Genome Biol.* **7**, R13 (2006).
22. Z. L. Gu, S. A. Rifkin, K. P. White, W. H. Li, *Nat. Genet.* **36**, 577 (2004).
23. V. Katju, M. Lynch, *Mol. Biol. Evol.* **23**, 1056 (2006).
24. M. J. Lercher, T. Blumenthal, L. D. Hurst, *Genome Res.* **13**, 238 (2003).
25. F. A. Kondrashov, I. B. Rogozin, Y. I. Wolf, E. V. Koonin, *Genome Biol.* **3**, research0008 (2002).
26. We thank M. Frazer (JGI), P. Cherbas (CGB), R. Green, and T. Takova (Roche NimbleGen, Inc.). The work conducted by the U.S. Department of Energy Joint Genome Institute (JGI) was supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-05CH11231 and in collaboration with the *Daphnia* Genomics Consortium (DGC). This project was also supported by NSF grants 0221837 and 0328516 and NIH grant R24GM07827401. Coordination infrastructure for the DGC is provided by the Center for Genomics and Bioinformatics (CGB) at Indiana University, which is supported in part by the METACyt Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc. Additional contributions and acknowledgments are provided in the SOM. Our work benefits from and contributes to the *Daphnia* Genomics Consortium. *Daphnia pulex* genome assembly version 1.1 and annotations are deposited at DNA Data Bank of Japan, European Molecular Biology Laboratory, and GenBank databases under accession ACJG00000000. ESTs (FE274839 to FE425949) are in GenBank. Microarray platforms GPL11200 to GPL11201 and data GSE25823 are deposited at National Center for Biotechnology Information Gene Expression Omnibus database.

Supporting Online Material

www.sciencemag.org/cgi/content/full/331/6017/555/DC1
Materials and Methods
SOM Text
Figs. S1 to S36
Tables S1 to S50
References

14 September 2010; accepted 16 December 2010
10.1126/science.1197761