Lawrence Berkeley National Laboratory

Lawrence Berkeley National Laboratory

Title

The ecoresponsive genome of Daphnia pulex

Permalink

https://escholarship.org/uc/item/72r3k7s9

Author

Colbourne, John K.

Publication Date

2011-02-04

eScholarship.org Powered by the California Digital Library
University of California

The Ecoresponsive Genome of *Daphnia pulex*

John K. Colbourne[1‡], Michael E. Pfrender[2,25], 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[9,26], Darren J. Bauer[4], Carla E. Cáceres[10], Liran Carmel[9,27], Claudio Casola[3], Jeong-Hyeon Choi[1], John C. Detter[7], Qunfeng Dong[1,28], Serge Dusheyko[7], Brian D. Eads[1,3], Thomas Fröhlich[8], Kerry A. Geiler-Samerotte[5,29], Daniel Gerlach[11,30], Phil Hatcher[4], Sanjuro Jogdeo[4,31], Jeroen Krijgsveld[12,32], Evgenia V Kriventseva[11], Dietmar Kültz[13], Christian Laforsch[14], Erika Lindquist[7], Jacqueline Lopez[1], J. Robert Manak[15,33], Jean Muller[16,34], Jasmyn Pangilinan[7], Rupali P. Patwardhan[1,35], Samuel Pitluck[7], Ellen J. Pritham[17], Andreas Rechtsteiner[1,36], Mina Rho[18], Igor B. Rogozin[9], Onur Sakarya[5,37], 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[1,38], Dai Tsuchiya[1], Hank Tu[7,37], Harmjan Vos[12,39], 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] and Jeffrey L. Boore[7,23,24]

[1] The Center for Genomics and Bioinformatics, Indiana University, 915 East Third Street, Bloomington, Indiana 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, Indiana 47405 USA

[4] Hubbard Center for Genome Studies, University of New Hampshire, 35 Colovos Road, Durham, New Hampshire 03824 USA

[5] 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] DOE Joint Genome Institute, 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 Maryland 20894 USA

[10] School of Integrative Biology, University of Illinois, 515 Morril Hall, Urbana, Illinois 61801 USA

[11] University of Geneva Medical School, 1 rue Michel-Servet, 1211 Geneva, Switzerland; 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, The Netherlands

[13] Department of Animal Science, University of California, Davis, Meyer Hall, One Shields Avenue, Davis, California 95616 USA

[14] Department of Biology II and GeoBio Center Munich, Ludwig-Maximilians-University Munich, Großhadernerstr.2, 82152 Planegg-Martinsried, Germany

[15] Gene Expression, Roche NimbleGen Inc., 504 S. Rosa Rd, Madison WI 53719 USA

[16] Structural and Computational Biology Unit, EMBL Heidelberg, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

[17] Department of Biology, University of Texas, Arlington, Box 19498, Arlington, Texas 76019 USA

[18] School of Informatics and Computing, Indiana University, Informatics Building, 901 E. 10th Street, Bloomington IN 47408-3912 USA

The Ecoresponsive Genome of *Daphnia pulex*

[19] The School of Public and Environmental Affairs, Indiana University, 1315 East Tenth Street, Bloomington, Indiana 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 S. Goodwin Ave, 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

[25] Present address: Department of Biological Sciences, University of Notre Dame, 109B Galvin Life Sciences, Notre Dame, IN 46556 USA

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

[27] 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

[28] Present address: Departments of Biological Sciences, Computer Science and Engineering, University of North Texas, Denton, TX 76203 USA

[29] Present address: Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138 USA

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

[31] Present address: Center for Genome Research and Biocomputing, Oregon State University, Corvallis, Oregon 97331 USA

[32] Present address: Genome Biology Unit, EMBL Heidelberg, Meyerhofstrasse 1, 69117 Heidelberg, Germany

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

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

[35] Present address: Department of Genome Sciences, University of Washington, Seattle, WA 98195 USA

[36] Present address: Department of MCD Biology, University of California, Santa Cruz, Santa Cruz, CA 95064 USA

[37] Present address: Life Technologies Corporation, Foster City, CA USA

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

[39] Present address: Department Molecular Cancer Research, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands

[‡]Corresponding email: jcolbour@indiana.edu

Acknowledgements:

We thank Marvin Frazer (JGI), Peter Cherbas (CGB), Roland Green and Tsetska 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 No. DE-AC02-05CH11231, and in collaboration with the *Daphnia* Genomics Consortium (DGC). This project was also supported by major NSF grants 0221837 and 0328516, and NIH grant IR24GM07827401A1. 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 acknowledgements are provided in the SOM. Our work benefits from, and contributes to the *Daphnia* Genomics Consortium

Disclaimer:

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

Supporting Online Material (SOM) for:

The Ecoresponsive Genome of *Daphnia pulex*

This file includes:

Contributions and Acknowledgements
Material and Methods I-VIII
Supporting Text 1-4
Supplementary Figures S1-36
Supplementary Tables S1-50
Supplemental References S1-183

CONTRIBUTIONS AND ACKNOWLEDGEMENTS

Steering Committee

Justen Andrews, Jeffrey Boore, Peer Bork, John Colbourne, Teresa Crease, Qunfeng Dong, Donald Gilbert, Igor Grigoriev, Joshua Hamilton, Eugene Koonin, Michael Lynch, Michael Pfrender, Hugh Robertson, Harris Shapiro, Joseph Shaw, Kelley Thomas, Evgeny Zdobnov

DNA Library Construction and Quality Control

Darren Bauer, John Colbourne, Chris Detter, James Haney, Michael Lynch, Michael Pfrender, Sarah Schaack, Kelley Thomas

Genome Sequencing

Jeffrey Boore, Chris Detter, Susan Lucas

cDNA Library Construction, Sequencing and Analysis

Darren Bauer, Jeffrey Boore, John Colbourne, Qunfeng Dong, Brian Eads, Donald Gilbert, James Haney, Erika Lindquist, Rupali Patwardhan, Michael Pfrender, Joseph Shaw, Zachary Smith, Zuojian Tang, Kelley Thomas, Mei Wang

Chromosomal Organization

John Colbourne, Dai Tsuchiya

Sequence Assembly and Validation

Jeong-Hyeon Choi, John Colbourne, Serge Dusheyko, Donald Gilbert, Sanjuro Jogdeo, Jasmyn Pangilinan, Samuel Pitluck, Hugh Robertson, Harris Shapiro, Haixu Tang, Kelley Thomas, Abraham Tucker, Hank Tu

Protein Coding Gene Prediction

Andrea Aerts, Donald Gilbert, Igor Grigoriev, Yuri Kapustin, Boris Kiryutin, Paul Kitts, Terence Murphy, Asaf Salamov, Victor Sapojnikov, Alexander Souvorov

Non-Coding Gene Prediction

Donald Gilbert, Daniel Gerlach, Michael Lynch, Ellen Pritham, Mina Rho, Sarah Schaack, Way Sung, Haixu Tang, Evgeny Zdobnov

Microarray Experiments and Functional Genomics

Jeong-Hyeon Choi, John Colbourne, Karel De Schamphelaere, Brian Eads, Donald Gilbert, Stephen Glaholt, Roland Green, Noah Greenberg, Christian Laforsch, Zhoa Lai, Leigh Latta, Florian Leese, Jacqueline Lopez, John Manak, Michael Pfrender, Ralph Pirow, Andreas Rechtsteiner, Joseph Shaw, Courtney Skalitzky, Ralph Tollrian

Proteomics

Georg Arnold, Thomas Frohlich, Jeroen Krijgsveld, Dietmar Kültz, Christian Laforsch, Harmjan Vos

Comparative Analysis of Genome Organization and of Gene Inventory

Jeffrey Boore, Paramvir Dehal, Donald Gilbert, Igor Grigoriev, Evgenia Kriventseva, Todd Oakley, Asaf Salamov, Onur Sakarya, Yuzhen Ye, Evgeny Zdobnov

Intron Evolution

Malay Kumar Basu, Liran Carmel, Eugene V. Koonin, Igor B. Rogozin, Yuri I. Wolf

Gene Duplication History

Claudio Casola, Donald Gilbert, Matthew Hahn, Phil Hatcher, Kelley Thomas, Abraham Tucker

Hemoglobin Gene Family

Shin-ichi Tokishita, Hideo Yamagata

Opsin Gene Family

Carla Cáceres, Kerry Geiler-Samerotte, Todd Oakley, Carolina Peñalva-Arana, Hugh Robertson, Catherine Seul, Kelley Thomas, Kim Walden

Metabolic Pathways

Peer Bork, Jean Muller, Takuji Yamada

Gene Duplication Model

John Colbourne, Donald Gilbert, Mike Pfrender, Kelley Thomas

Manual Annotation Project

Abderrahmane Tagmount, Abe Tucker, Abhigna Polavarapu, Adrian Maximillian Fischl, Ajna Rivera, Alexandra Jauhiainen, Amanda Callaghan, Andrew Schurko, Angela Omilian, Anke Freeman, Anna Syme, Armin Sturm, Austin Elliott, Bastiaan Jansen, Birgit Pils, Brent Hallahan, Brian Eads, Carla Caceres, Carlos Villacorta Martin, Chris Hill, Chris Vulpe, Christian Laforsch, Christoph Mayer, Claire Conlon, Cornelis Grimmelikhuijzen, D. Carolina Penalva-Arana, Darin Hullinger, Darren J. Bauer, David Innes, David Kehoe, David Van Dyken, Dietmar Kültz, Ditlecadet Delphine, Don Gilbert, Elisabeth Stafflinger, Ellen Decaestecker, Erik Kristiansson, Feseha Abebe-Akele, Florian Leese, Florian Raible, Flynn Picardal, France Dufresne, Francis Poulin, Frank Hauser, Frank Nunes, Gary Stuart, Giuseppe Cazzamali, Haleh Ashki, Harald Parzer, Heinrich Dircksen, Helen Poynton, Henri Wintz, Hugh Robertson, Jacqueline Ann Lopez, Jade Carter, James Costello, Jasleen Kaur, Jeffry Dudycha, Jeong-Hyeon Choi, Joachim Mergeay, John Colbourne, John Logsdon, Jonathon Stillman, Joseph Shaw, Justen Andrews, Karen Wilson, Kim Rewitz, Krystalynne Morris, Lars Heckmann, Lawrence J. Weider, Lee Bjerregaard, Leigh Latta, Le-Shin Wu, Lev Yampolsky, Loren Probst, Margaret Beaton, Mark Blaxter, Martina Schneider, Mehmet Dalkilic, Melania E. Cristescu, Michael Pfrender, Michael Williamson, Mieke Jansen, Mike Wang, Molly Craxton, Philippe Boucher, Piers Napper, Preeti Misra, Puni Jeyasingh, Qunfeng Dong, Rajesh Gollapudi, Ralph Pirow, Ralph Tollrian, Ramya Sabbineni, Rebecca Klaper, Rick Zuzow, Robert Sterner, Rupali Patwardhan, Sarah Schaack, Seanna McTaggart, Sebastian Becker, Shinichi Tokishita, Shiva Sinha, Shu Shang, Simon Webster, Snaebjorn Palsson, Stephanie Chan, Sumit Middha, Sun Kim, Susan Gordon, Susanne Paland, Timothy S. McClintock, Tiong Khong Loon, Todd Oakley, Tom Little, Toru Miura, Travis Garriott, Tutku Aykanat, W. Kelley Thomas, Way Sung, William Baldwin, Xin Hong, Xinguo Wang, Yang Bai, Yasuhiro Shiga, Yi Zou, Zuojian Tang.

Daphnia pulex genome assembly V1.1 and annotations are deposited at DDBJ/EMBL/GenBank under the accession ACJG00000000. ESTs (FE274839-FE425949) are in GenBank. Microarray platforms GPL11200-GPL11201 and data GSE25823 are deposited at NCBI GEO.

Acknowledgements – We thank Marvin Frazer, then head of U.S. DOE Life Sciences, for the inspiration and commitment to pursue the sequencing of this first crustacean genome. We thank Peter Cherbas, who directs the Center for Genomics and Bioinformatics (CGB), for his support and leadership in creating this new genomic model system. We thank Gregory Werner and his group at JGI for support of gene annotation tools. We also thank Roland Green, Tsetska Takova and their groups at Roche NimbleGen Inc. for providing early access and technical expertise to custom microarray technologies enabling the functional annotation of the genome sequence. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and in collaboration with the *Daphnia* Genomics Consortium (DGC). This project was also supported by the NSF Biocompolexity grant 0221837 to Joshua Hamilton (and Celia Chen, Carol Folt, Joseph Shaw, Michael Lynch and John Colbourne), and FIBR grant 0328516 to Michael Lynch (and John Colbourne, Justen Andrews, Curtis Lively, Elizabeth Housworth, Miriam Zolan, Jeffrey Boore, Carla Cáceres, Thomas Little and W. Kelley Thomas). NIH support IR24GM07827401A1 was granted to Michael Pfrender (and John Colbourne, Donald Gilbert, Dieter Ebert and W. Kelley Thomas). Euro Cores/EuroEEFG grant support (through DFG) grant LA 2159/6-1 was provided to Christian Laforsch, Georg Arnold and Thomas Frohlich, in partnership with Luc De Meester (PI; and Luisa Orsini, Ellen Decaestecker, Colin Janssen, Karel De Schamphelaere, Dieter Ebert, Cristoph Haag, Adam Petrusek, Mikko Frilander, John Colbourne, Andrew Beckerman, Thomas Little). Malay Kumar Basu, Liran Carmel, Eugene Koonin, Igor Rogozin, and Yuri Wolf were supported by Intramural funds of the US Department of Health and Human Services (NIH, National Library of Medicine). Todd Oakley was supported

by NSF grant DEB 1027279. Dietmar Kültz was supported by NSF grant IOS0542755 and NIH grant P42ES004699. Thanks to Keithanne Mockaitis and the CGB sequencing team (Jade Carter, James Ford, Zach Smith) for access to an early draft assembly of the *Daphnia magna* genome sequence. Sequencing infrastructure at the CGB was provided by a major grant from the the Lilly Endowment, Inc. Thanks to Matthew Hahn (Indiana University) for providing critical suggestions. The following people contributed DNA and RNA samples for this research: Jim Haney (University of New Hampshire), Rebecca Klaper (University of Wisconsin-Milwaukee), Thomas Little (University of Edinburgh), Norman Yan (York University), Jarkko Routtu and Dieter Ebert (University of Basel). Analyses, data curation and data distribution are primarily attributed to wFleaBase, developed at the Genome Informatics Lab of Indiana University with support to Don Gilbert from the National Science Foundation and the National Institutes of Health. Specialized shared databases were also created by Mark Blaxter (University of Edinburgh) and Hajime Watanabe (Okazaki National Research Institutes). Coordination infrastructure for the DGC is provided by The Center for Genomics and Bioinformatics 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. Computer support was provided by an allocation TG-MCB060059N through the TeraGrid Advanced Support, by the University Information Technology Services (UITS) and by The Center for Genomics and Bioinformatics computing group. We thank the computing group leaders Phillip Steinbachs (CGB), Craig Stewart and Richard Repasky (UITS). Ann Miracle advised on the successful timing for the initial submission of the White Paper proposal to the JGI. Our work benefits from, and contributes to the *Daphnia* Genomics Consortium. http://daphnia.cgb.indiana.edu

MATERIALS AND METHODS

I. Genome Sequence, Assembly and Mapping to Chromosomes

1. Strains for genome sequencing
2. Sequencing and assembly
3. Validating the draft genome assembly
4. Comparative genomic hybridization using multiplex microarrays
5. Chromosome studies

II. Gene Inventory

1. Manufacturing gene models and selection of the minimum set
2. Transcriptome sequencing 37 cDNA libraries
3. Proteome sequencing
4. NimbleGen genome tiling microarray experiments
5. Transcription profiling using NimbleGen multiplex microarrays
6. Annotating protein-coding genes
7. Annotating non-coding RNA and transposable elements

III. Attributes of a Compact Genome

1. Comparing genome structures
2. Comparative study of intron evolution

IV. Origin and Preservation of *Daphnia pulex* Genes

1. Assigning gene homologies
2. Studying the history of gene family expansions and losses
3. Studying the history of gene duplication
4. Measuring the distribution of duplicated genes using *Tandy*
5. Identifying lineage specific gene family expansions
6. Annotating and tracing the phylogeny of opsins

V. Implications of *Daphnia*'s Genome Structure

1. Finding non-allelic gene conversion events
2. Annotating and tracing the phylogeny of hemoglobins

VI. Evolutionary Diversification of Duplicated Genes

1. Estimating expression-level divergence among paralogs
2. Testing for genome structure effects on expression divergence

VII. Functional Significance of Expanded Gene Families

1. Charting metabolic pathways for co-expanding, interacting genes
2. Uncovering functional diversity of glycosphingolipid biosynthesis genes

VIII. Ecoresponsive Genes

1. Treatment of the transcriptome data with reference to the annotation

SUPPORTING TEXT

1. Chromosome Studies
2. Gene Homology among *Daphnia* Genomes
3. Micro-RNA and Transposable Elements

4. The 46 *Daphnia pulex* Opsins

SUPPLEMENTARY FIGURES

Figure S1. Reconstruction of the evolutionary history of sequenced arthropods.

Figure S2. Overview of the *Daphnia pulex* Genome Project.

Figure S3. Distributions of the cumulative scaffold and gap lengths for the JAZZ, Arachne, and PCAP assemblies.

Figure S4. Venn diagram highlighting the number of putative mis-assembled regions by using three different methods.

Figure S5. The number of detected breakpoints by GAV in the scaffolds with different lengths.

Figure S6. The karyotype of *Daphnia pulex* based on meiotic chromosomes prepared from testis.

Figure S7. Corroborating evidence for the existence of a minimal set of 30,907 predicted protein coding genes.

Figure S8. Cumulative frequency distribution of the ratio of non-synonymous over synonymous nucleotide substitutions among duplicated genes in the genome.

Figure S9. Evidence that genes residing in areas of low read coverage within the draft genome assembly are genuine.

Figure S10. *Daphnia pulex* reveals arthropod origin of two Hox cluster encoded microRNAs (iab-4 and mir-993).

Figure S11. Distribution of transposon *Pokey* in the ribosomal DNA of *Daphnia pulex*.

Figure S12. Age distribution of *Daphnia pulex* Long Terminal Repeats elements (LTRs).

Figure S13. Size distribution of introns in *Daphnia pulex, Caenorhabditis elegans*, *Drosophila melanogaster*, and *Mus musculus*.

Figure S14. Pair-wise percentage of intron conservation.

Figure S15. Ancestral reconstruction of intron gains and losses for arthropods and three other metazoans.

Figure S16. Estimated independent and parallel gain of introns in *Daphnia*.

Figure S17. Frequency of pair-wise genetic divergence at silent sites (K_s) among the 2-member gene duplicates in the *Daphnia pulex*, *Caenorhabditis elegans* and *Homo sapiens* genomes.

Figure S18. Frequency of pair-wise genetic divergence at silent sites (K_s) among gene duplicates in *Daphnia pulex*.

Figure S19. Position and size of Tandem Duplicated Gene (TDG) clusters within the genome assemblies of four model species.

Figure S20. Physical distances between neighboring members of large duplicated gene families.

Figure S21. Phylogenetic relationships of 39 of the 46 *Daphnia pulex* opsin genes and representative animal opsins.

Figure S22. Maximum-likelihood phylogenies of *Daphnia pulex* opsin genes.

Figure S23. Rates of gene conversion and number of intervening genes between duplicates in *Daphnia pulex* and *Drosophila* species.

Figure S24. Rates of gene conversion and divergence between duplicates in *Daphnia pulex* and *Drosophila* species.

Figure S25. Amino acid sequence alignment of di-domain hemoglobins (Hb) of *Daphnia pulex* and *D. magna*.

Figure S26. Nucleotide sequence alignment of di-domain hemoglobins (Hb) in coding regions of genes.

Figure S27. Nucleotide sequence alignment of intergenic regions between the stop codons of upstream genes and the TATA of the downstream genes of all *Daphnia pulex* (dpul) and *D. magna* (dmag) di-domain hemoglobins (Hb).

Figure S28. Differential expression (DE) profiles of 37 of the 46 *Daphnia pulex* opsin genes from eight microarray experiments.

Figure S29. Differential expression (DE) profiles of 11 *Daphnia pulex* di-domain hemoglobin genes from eight microarray experiments

Figure S30. Thirty-eight expanded and 54 contracted metabolic genes in arthropod genomes compared to vertebrates.

Figure S31. Expanded metabolic genes in the *Daphnia pulex* genome compared to other arthropods and vertebrates.

Figure S32. Distribution of the number of amplified genes with interactions, derived from 1,000 randomized metabolic networks.

Figure S33. Phylogenetic relationships of the three expanded gene families of the *Daphnia pulex* glycosphingolipid biosynthesis neo-lactoseries pathway of metabolism.

Figure S34. Differential expression (DE) pattern correlations among the *Daphnia pulex* gene-members of three lineage-specific gene family expansions from eight microarray experiments.

Figure S35. The phylogeny of duplicated fucosyltransferase genes compared to their differential expression profiles across 8 experimental conditions on microarrays.

Figure S36. Differential transcription of the genome from *D. pulex* on whole-genome tiling microarrays.

SUPPLEMENTARY TABLES

Table S1. Open-source web-portals for *Daphnia pulex* genome data, analysis results and bioinformatic tools.

Table S2. Summary of the *Daphnia pulex* genome assemblies using three assemblers.

Table S3. Analysis of shotgun reads from TCO and TRO derived libraries.

Table S4. Putative super-scaffolds based on focused paired-end read analysis.

Table S5. Scaffolds genetically mapped to chromosomes.

Table S6. Pair-wise comparison of genome assemblies: Arachne *versus* JAZZ and PCAP *versus* JAZZ using MUMmer.

Table S7. GAV (Genome Assembly Validator) is a machine learning approach to a combined evidence validation of genome assemblies.

Table S8. Chromosome size measurements.

Table S9. Results from the automated gene annotation procedures.

Table S10. The *Daphnia pulex* cDNA libraries and EST sequencing effort.

Table S11. Observed homology and transcription evidence for v1.1 annotated gene set of the *Daphnia pulex* genome.

Table S12. Supporting evidence for additional 7,965 TAR-predicted loci.

Table S13. List of identified proteins.

Table S14. List of identified peptides.

Table S15. List of 716 genes conserved as single-copy othologs across eukaryotic genomes.

Table S16. Fifty predicted *Daphnia pulex* miRNA genes.

Table S17. Comparative analysis of transposable elements in *Daphnia pulex*.

Table S18. Classification and distribution of Transposable Elements in *Daphnia pulex*.

Table S19. Comparison with gene-structure statistics for insects, nematode and mouse.

Table S20. Species used in the study of introns.

Table S21. Number and density of introns for nine species.

Table S22. Conservation of *Daphnia* introns.

Table S23: Conservation of intron positions between *Daphnia pulex* and other animals.

Table S24. Maximum Likelihood reconstruction of intron gain and loss events in arthropods and three other metazoans.

Table S25. Similarity of *Daphnia pulex* genes and 12 other genome-sequenced arthropods to human and other model eukaryote reference proteins.

Table S26. Gene families in *Daphnia pulex* with recognizable InterPro protein domains that have expanded relative to gene families in insects.

Table S27. Species used in the study of gene family expansions history (see Figure 1C).

Table S28. EvolMap reconstruction of gene gain and loss events in arthropods and four other metazoans.

Table S29. Gene duplication and duplicate gene birth rates in the *Daphnia pulex*, *Caenorhabditis elegans* and *Homo sapiens* genomes.

Table S30. Large fraction of *Daphnia pulex* duplicated genes.

Table S31. Gene families that are expanded and/or shared between *Daphnia pulex* and other aquatic (vertebrate) species.

Table S32. Part A. Forty-six *Daphnia pulex* opsin genes belonging to 6 major clades. Part B. Additional Metazoan Opsins in Figure S20.

Table S33. Summary of gene conversion features as a function of the number of genes within the genomes of *Daphnia pulex* and *Drosophila* species.

Table S34. Summary of genome-wide gene conversion features among *Daphnia* and *Drosophila* species.

Table S35. Summary of genome-wide gene conversion features as a function of the location of paralogs on scaffolds or Müller elements among *Daphnia* and *Drosophila* species.

Table S36. Summary of genome-wide gene conversion features as a function of the size of conversion tracts among *Daphnia* and *Drosophila* species.

Table S37. Summary of genome-wide gene conversion features as a function of the size of gene families among *Daphnia* and *Drosophila* species.

Table S38. Summary of genome-wide gene conversion features as a function of the distance of intra-element or intra-scaffold paralogs among *Daphnia* and *Drosophila* species.

Table S39. Homologous di-domain hemoglobin genes of *Daphnia pulex* and *Daphnia magna*.

Table S40. The number of paralog pairs that differ unambiguously in their expression patterns.

Table S41. Chi-square tests for associations between paralogs ($K_s < 2$) sharing expression patterns across 12 conditions tested on microarrays.

Table S42. The number of paralog pairs that have the same expression patterns and that have different expression patterns among 0 to 12 conditions.

Table S43. Metabolic pathways (classified by KEGG and highlighted in Figure 4) containing expanded metabolic genes in the *Daphnia pulex* genome compared to insects and vertebrates.

Table S44. Metabolic pathways (classified by KEGG and highlighted in Figure 4) containing expanded metabolic genes in the arthropod genomes compared to vertebrate genomes.

Table S45. Ninety-six (96) *Daphnia pulex* genes from three lineage-specific gene family expansions that are part of the glycosphingolipid biosynthesis neo-lactoseries metabolic pathway.

Table S46. Alignment of Enzyme 2.4.1.65 *Daphnia* proteins, with *Tribolium castaneum* and *Ixodes scapularis* orthologs, using MUSCLE.

Table S47. Alignment of Enzyme 2.4.1.206 *Daphnia* proteins, with *Tribolium castaneum* and *Ixodes scapularis* orthologs, using MUSCLE.

Table S48. Alignment of Enzyme 2.4.1.152 *Daphnia* proteins, with *Tribolium castaneum* and *Ixodes scapularis* orthologs, using MUSCLE.

Table S49. Counts of unique gene transcripts sampled from cDNA libraries partitioned into three ecological conditions.

Table S50. Differential expression (DE) of the genome of *Daphnia pulex* with four treatments measured on genome tiling path microarrays.

Supplemental References

MATERIALS AND METHODS

I. Genome Sequence, Assembly and Mapping to Chromosomes

1. Strains for genome sequencing

A natural isolate within the *D. pulex* species complex was picked for sequencing. The Chosen One (TCO) reproduces by cyclical pathenogenesis (capable of both clonal and sexual reproduction) and is easy to culture. The isolate was sampled from a naturally inbred population inhabiting a permanent pond in the Siuslaw National Forest, near the Pacific coast in Oregon, USA. Slimy-Log pond is situated south of Florence and Dunes City, in Douglas County, on the east side of HWY 101, at milepost marker 201 (GPS coordinates N 43.830013, W -124.148152). Sequences from mitochondrial genes suggest that the isolate belongs to an incipient lineage of *D. pulex*, endemic to an area west of the Cascade Mountains, called *D. arenata* [S1]. Allozyme and microsatellite genotyping indicated that gene diversity within this population is ~4% [S2]. Of eight randomly chosen individuals, TCO possessed the lowest nucleotide heterozygosity (~0.14%) at 17 sequenced loci. This level of nucleotide polymorphism is comparable to variation found in the sequenced human genome [S3] and is suitably homozygous for the assembly of shotgun derived sequences into contigs. The actual nucleotide heterozygosity of the genome is 0.1% per site.

A second isolate was also sequenced, albeit at 1x coverage of the genome, to map and study polymorphisms. The Rejected One (TRO) is a hybrid clone of *D. pulex* found in ponds and of the lacustrine species *D. pulicaria*. The nucleotide heterozygosity of TRO is 1.44% per site to study molecular evolutionary patterns.

The isoclonal animals were grown to large numbers in filtered culture medium, and then treated with 500 mg/L of Tetracycline to reduce bacterial contamination and with 4.5 micron copolymer microsphere beads (Duke Scientific cat# 7505A; Palo Alto, CA) to clear the gut. High molecular weight DNA was isolated by Genomic-tips using the manufacturer's protocol for animal tissues (Qiagen, Valencia, CA).

2. Sequencing and assembly

Three size-specific genomic DNA libraries were created using standard protocols for paired-end shotgun Sanger sequencing on ABI 3730xl and MegaBACE 4000 machines. From a total of 2,711,298 sequences, 1,225,940 reads (45%) were obtained from a 2,000-3,000 bp insert plasmid library, 1,272,122 reads (47%) were obtained from an 6,000-8,000 bp library and 228,191 reads (8%) were obtained from a 35,000-40,000 bp insert fosmid library. The total number of sequenced nucleotides used in the assembly is 1,211 Mb, of which 95.7% are ascribed to the *D. pulex* genome. Over 2.4 Mb are ascribed to the *Daphnia* metagenome [S4]. In addition, there were 1,065,732 reads that were ultimately not used in the assembly, containing 1,006 Mb of untrimmed sequence.

The draft genome assembly v1.1 was built using the JAZZ assembler [S5] from 1,645,566 quality-filtered sequence reads. The JAZZ assembly is composed of 44,403 contigs and 26,848 scaffolds of which 5,191 belong to the nuclear genome. This assembly includes 17,555 gaps averaging 3,300 bp (ca. 39 Mb in total). Two additional assemblies were created using the ARACHNE [S6] and PCAP [S7] assemblers. The results are reported without filtering. Compared to JAZZ, the ARACHNE assembler produced an equivalent number of scaffolds, yet from twice as many contigs. Although the ARACHNE contigs include only 20 Mb of additional nucleotides, the

ARACHNE scaffolds sum to 396 Mb. This discrepancy is attributed to 3.25 times more gaps, which total an estimated 186.8 Mb of missing data (Table S2). By contrast, the PCAP assembler produced 2.3 times more scaffolds than JAZZ, yet both sets sum to the same length (Figure S3).

We next improved the genome assembly by the manual alignment of trimmed paired-end reads from both TCO and TRO to the sequence scaffolds to build super-scaffolds. Custom scripts identified paired-end reads that aligned uniquely to separate TCO scaffolds. Strict criteria were imposed so to not introduce errors: alignments met a minimum e-value threshold of 1×10^{-100} and scored better that the next best alignment by > 50 orders of magnitude. Results after filtering the data are summarized (Table S3).

Remaining mate-pairs located on the same scaffold provided an actual DNA insert length distribution for each gDNA library. We found that some clones measured distances that were much larger than the predicted insert sizes. Therefore, a second filtering step was applied by removing all mate-paired sequences that spanned > 2x their predicted distance. The modified means and standard deviations for each library were then used to determine whether paired-ends that aligned to different scaffolds were sufficiently close to the scaffold terminals to serve as bridges. The set threshold was three standard deviations from the modified average read length for each gDNA library. If both paired-end reads were within the set cutoff from the ends of scaffolds, the reads were considered appropriate candidates for bridging scaffolds.

This above strategy identified 151 instances where at least one set of unique paired-end reads joins two scaffolds. After verification of the results, we propose a final set of 118 super-scaffolds (Table S4). In 51 cases, a super-scaffold is supported by more than one independent set of paired-end reads. Further support is provided for seven cases where the joined scaffolds are found on the same chromosome, base on independent genetic analysis [S8](see Table S5).

The super-scaffolds represent a significant improvement of the overall assembly (Figure S3). The N50 for the super-scaffold assembly is 83 compared to 103 for the current assembly. Furthermore the number of super-scaffolds longer than 2.5 Mb has nearly tripled (14) compared to the original assembly (5).

3. Validating the draft genome assembly

Eukaryotic draft genome assemblies contain errors that often appear in regions with low read or clone coverage, regions containing chimeric or recombined sequence reads, regions that have compressed distances due to repeated elements or have wrongly oriented paired-end reads [S9]. We validated the overall quality of the *D. pulex* genome sequence assembly using two methods. First, we compared the assembly created by JAZZ [S5] to competing assemblies built by the ARACHNE [S6] and PCAP [S7] assemblers (Table S2). Comparative results were obtained by matching shared contiguous regions between assemblies using MUMmer [S10]. JAZZ produced 44,403 contigs having a total length of 186,524,647 bp. We find that 94% and 98% of the JAZZ contigs matched with the ARACHNE and PCAP contigs, respectively (corresponding to 98% and 95% of their total contig lengths)(Table S6). By contrast, ARACHNE and PCAP produced many more contigs than JAZZ (80,844 and 74,521) with greater total lengths (~209 Mb and ~234.5 Mb, respectively)(Table S2). To detect inconsistent regions between the JAZZ assembly and the two reference assemblies, blocks of fixed length (e.g. 2,000 bp) in the JAZZ assembly were classified into three categories (Table S6): (1) unmatched blocks without alignments to the contigs in the reference assembly; (2) uniquely matched blocks that align to a unique and contiguous region in the reference assembly; and (3) overlapping blocks containing two overlapping regions matched to two different contigs in the reference assembly. This third

category lists putatively mis-assembled regions, which are called breakpoints in the contigs. Two sets of breakpoints (blocks) were identified by referencing each of the two assemblies, after filtering out imperfect matches within the MUMmer output if they did not have a unique region of a certain number of bases. We used 500 bp and 1,000 bp blocks to define regular and stringent criteria.

Our second method applied machine learning to a combined evidence validation of genome assemblies (called GAV)[S11]. The machine learning model was trained to predict breakpoints within a 2,000 bp block of assembled sequence using features deduced from the placement of reads and mate-pairs that cover this block, such as the read and clone coverage, clone length statistics, and repeat content. The training data sets included blocks that positively contained breakpoints and blocks that were positively error-free (Table S7A). These were confirmed by the EST alignments to the genome sequence. The training procedure follows. (1) ESTs that were not well aligned to the genomic (contig) sequences were filtered out based on the matching length (L), score (S), and e-values (V). By default, we used L=200, S=100, $V=1 \times e^{-10}$. (2) The mate-pairs (reads from the 5' and 3' ends of cDNA clones) were individually and unambiguously aligned onto the contigs. (3) When the 5' and 3' ESTs from a cDNA clone had incorrect orientation, the corresponding block was classified as a mis-assembled (negative) region of the contig. (4) Otherwise, unaligned regions in the cDNA clone were checked when aligned to the contigs. If the size of an unaligned region was greater than 50 bp, the block covering the boundary of the unaligned region was classified as a mis-assembled (negative) region. (5) We also checked the distances between the location of 5' and 3' ESTs. If the distance was greater than a cutoff (10,000 bp), we classified the block covering the boundary of the EST as a mis-assembled (negative) sample. (6) The blocks covered by the remaining ESTs were classified as correctly assembled (positive). In total, 116,714 positive blocks and 10,232 negative blocks were obtained, which represent 4,536 contigs and 920 scaffolds (Table S7A).

Since the 5,191 scaffolds of the current JAZZ assembly were chosen for the annotation of the *D. pulex* genome, these were further validated. Using the procedures described above, a consensus set of likely mis-assembled blocks (of length 2,000 bp) was predicted. We identified 1,889 breakpoints when using the ARACHNE assembly as reference, and 3,304 blocks when using PCAP as reference. GAV predicted 3,053 putatively mis-assembled blocks (Table S7B). Shared predicted breakpoints among the three sets are shown in Figure S4. Since each of the genome validation methods have inherently high false positive rates, concordance in their independent results produced a more reliable count of likely assembly errors. For instance, among the predicted mis-assembled regions by GAV, the best performance of the program produced 60% false positives [S11]. Although the program's performance seemed poor, its false negative rate was negligible [S11]; this exercise was therefore helpful to guide the necessary experimental validation. Finally, Figure S5 demonstrates a correlation between the length of scaffolds and the number of break points (i.e., the longer scaffolds tend to contain more breakpoints). Based on these analyses, sequence assembly errors are minimal, ranging between 0.1% and 0.5% of the total assembly (Table S7; Figures S4-5).

We investigated the genomic features residing in the assembly gaps by first identifying 19,733 paired-end sequences that were not included in the assembly (7,652 from 3 kb insert libraries; 8,397 from 7 kb insert libraries; and 3,684 from 35 kb insert libraries) where one end unambiguously aligned to a scaffold region, and the other end of the sequenced DNA fragment failed to unambiguously align, and fell within a gap, based on the insert sizes of the fragments. 17.5 Mb of DNA within 6,075 gaps were thus surveyed. The *D. pulex* paired-end reads that "dangle within gaps" are annotated as follows, using RepeatMasker

[http://www.repeatmasker.org] and RepBase [http://www.girinst.org/repbase/] database of arthropod repeats, and Gmap [S12] for EST and transcript finding:

- 1.16% of DNA within gaps is composed of simple repeats, 1.55% is composed of low complexity regions, and 2.50% is composed of transposons. These values are slightly higher than the 0.38 % simple repeats, 0.77 % low complexity, and 0.69 % transposons for the full assembly. These small increments are unlikely to have impacted the assembly.
- 22% of the *Daphnia* genes (see section II.1 below) have high-identify paralogs within gaps, which is equal to the number of paralogs found elsewhere in the assembled genome. These paralogs are found in 3,598 of the 6,075 surveyed gaps (59%).
- ESTs also mapped to the dangling reads at the same rate as found in the assembled genome. Of 151,075 ESTs, 5% are found in these reads – with average 90% identity – compared to 92% found in full assembly at 95% identity. Therefore, ESTs align to genomic DNA at nearly equal rates for dangling reads residing in gaps (0.0006 EST/base) and for assembled sequences of the genome (0.0008 EST/base).

Overall, we conclude that gaps contain repeated sequences. Given the number of high-identity paralogs arranged within 59% of the surveyed gaps, we surmise that, in particular, high-identity gene paralogs contributed to creating gaps in the *D. pulex* assembly

4. Comparative genomic hybridization using multiplex microarrays

In collaboration with Roche NimbleGen Inc. we designed and manufactured a multiplex (12-plex) long-oligonucleotide (60 nt) *D. pulex* microarray that measures gene expression and can also be used for comparative genome hybridizations. Each glass slide contains 12 identical arrays prepared using a Maskless Array Synthesizer [S13]. Each array consists of 137,000 temperature-balanced probes; 22,076 genes are represented by three unique probes, 13,232 genes are represented by two unique probes, 357 genes are presented by a single probe, while the remaining probes are designed from transcriptionally active regions whose gene models are not yet described. The array also contains control probes and random probes designed to reflect the genome nucleotide composition by Markov modeling.

DNA samples from 24 cultures of TCO were obtained using a CTAB method [S14] then quantified using a Quant-iT™ PicoGreen® dsDNA protocol [S15]. High molecular weight DNA (1 µg) was sheered using the Sonicator 4000 (Misonix, Farmingdale, NY) to generate 500–2,000 bp fragments. The fragmented gDNA sample was assessed by capillary electrophoresis using Bioanalyzer 2100 (Agilent Technologies, Colorado Springs, CO) then labeled using the Roche NimbleGen labeling kit. Briefly, 1 µg fragmented gDNA in 40 µl water was primed with 40 µl of 1-O.D. CY-labeled random nonomer primer at 95°C for 10 minutes, then immediately cooled to 4°C for 10 minutes. The reaction was followed with 100 U Klenow fragment (3>5 exo-) and 10 µl of 10 mM dNTP mix to a final volume of 100 µl, incubated at 37°C for 2 hours, and terminated with 0.5 M EDTA. CY-labeled gDNA was purified by isopropanol precipitation in the presence of sodium chloride. Concentration and purity of the resuspended Cy/DY labeled gDNA in water was determined using NanoDrop ND-1000 (Thermo Fisher Scientific, Waltham, MA).

Hybridization, post-hybridization washing and scanning were done according to NimbleGen User's Guide for CGH Analysis v.5.1 (16 Mar 2009) with modifications for the 12-plex array format. Images were acquired using a GenePix 4200A scanner with GenePix 6.0 software (Molecular Devices, MDS Analytical Technologies). The data from the images were extracted using the software NimbleScan v2.4 (Roche NimbleGen Inc., Madison, WI).

The data were imported into an in-house analysis pipeline using Bioconductor for the analysis [S16]. The signal distributions of all probes, including random probes, were adjusted across the 24 replicates to the same median.

5. Chromosome studies

The *D. pulex* karyotype (Figure S6) is based on the preparation of meiotic chromosomes as described previously [S17]. Prepared slides were placed on a heat block at 65°C overnight, incubated in 2×SSC at 60°C for 1h, and rinsed in 0.9% NaCl. For G banding, slides were dipped in 0.05% trypsin for 10 sec, rinsed in Gurr's buffer (Gibco, Carlsbad, CA), and stained with Giemsa (1 ml Giemsa [Gibco] buffered with 50 ml Gurr buffer) for 12 min. Finally, slides were rinsed in distilled water, air dried and analyzed by blight field observation. For DAPI banding, slides were stained with DAPI mounted in an antifading solution, Vectashield (Vector Laboratories, Burlingame, CA), and analyzed by fluorescence observation. Observations were made on a Nikon Ecripse 80i microscope equipped with a motorized Z axis. Images were captured with Photometrics HQ using Metamorph software and processed with Adobe Photoshop software. Measurements were performed using Scion image software.

II. Gene Inventory

1. Manufacturing gene models and selection of the minimum set

The minimum gene set refers to Dappu version 1.1 gene models. These models were predicted using several methods: Fgenesh [S18], Genewise [S19], SNAP [S20], PASA [S21] and Gnomon [S22](Table S9). These gene prediction methods include a combination of *ab initio* modeling, homology-based modeling using protein seeds from similar sequences in other genomes, and modeling based on cDNA sequence alignments to the genome assembly. Whole genome tiling path microarrays, peptide sequencing, and comparison with *D. magna* genome sequence were used as additional lines of evidence. In addition, genes were also manually curated.

The annotation pipeline typically produced multiple overlapping gene models, which were created by different gene predictors at each locus. To select the best representative gene model, we employed a heuristic approach, based on a combination of protein homology and EST support. Homology information was based on the best alignments produced by BLASTp searches [S23] from the NCBI protein database. Only alignments with scores >50 and coverage greater than 25% of the length of the gene models were considered valid models with homology support.

EST support was based on the correlation coefficient (CC), a measure commonly used to estimate the accuracy of predicted gene models relative to known, experimentally validated gene models [S24]. For this annotation project, an average CC value was computed from all ESTs that mapped to a gene model. The CC values ranged from -1 to +1, with +1 assigned to a perfect match between the ESTs and the predicted gene model, and -1 representing a complete disagreement. Negative correlations indicated potentially poor quality gene models. Therefore, models with negative correlations and poor homology support (alignment coverage both for gene model and its protein homolog <50%) were initially discarded from the minimum gene set.

Each gene model was assigned scores based on the following formula: $S = Sblast \times (cov1 \times cov2 + CC)$; where Sblast is the BLASTp score of alignments between a gene model and a protein homolog, cov1 and cov2 are the alignment-coverage for the model and homolog, respectively ($0 \leq cov1,cov2 \leq 1$), and CC is an average correlation coefficient between the model and all overlapping ESTs. For a given locus, the model with the highest score was

selected, and all other models that had greater than 5% overlap with the selected model were excluded from the final minimum gene set.

Ab initio models with no detectable homologs were also excluded from the minimum Dappu v1.1 set. Reducing the stringency of this gene selection project predicted a much larger count, potentially exceeding 40,423 genes. A protein similarity search against a draft genome sequence for *D. magna* at 8-fold coverage identifies 2,319 (23%) of 10,015 *ab initio* gene models, and 3,653 (46%) of 7,965 gene models proposed by TARs (section II.4) that are all presently excluded from our minimal set of genes. Moreover, of the >11,000 *D. pulex* peptide sequences detected by tandem mass spectrometry (section II.3), 880 peptides map to 95 *ab initio* gene models that are absent from the minimum set.

Multiple methods that follow were used to validate the Dappu version 1.1 gene builds.

2. Transcriptome sequencing 37 cDNA libraries

Twenty non-normalized cDNA libraries were generated from RNA extracted from a *D. pulex* isolate TRO. The libraries represent transcriptomes under a combination of 13 ecological conditions and three developmental stages (Table S10). The animals were cultured within large, aerated, 200 liter container of filtered lake water by feeding a concentrated monoculture of green algae (*Scenedesmus acutus*). Total RNA was isolated using Trizol reagent (Invitrogen Life Sciences, Carlsbad, CA) and was subsequently purified using the RNeasy protocol (Qiagen, Valencia, CA). The cDNA libraries were constructed and sequenced using previously described methods [S25], except that paired-end sequences were now obtained. This effort produced 70,765 reads from a total of 50,070 clonal plasmids. This method resulted in a gene discovery rate of 41% to 85% among the libraries and an average rate of 64%.

Sixteen additional cDNA libraries were constructed using normalization procedures that improve the sampling of uniquely identified genes among conditions (Table S10). Total RNA was isolated from the TCO isolate using Trizol reagent (Invitrogen Life Sciences, Carlsbad, CA) and was subsequently purified using the RNeasy protocol (Qiagen, Valencia, CA). The cDNA libraries were produced using the Creator SMART (Clontech, Mountain View, CA) system by following the manufacturer's instructions. After the cDNA synthesis but prior to cloning, the cDNA pool was normalized using the Trimmer-Direct cDNA normalization kit (Evrogen, Moscow, Russia), amplified then ligated into the pDNR-LIB vector. The vector-cDNA ligants were bacterial transformed into TOP10 competent cells (Invitrogen Life Sciences, Carlsbad, CA), grown onto selective 2×YT agar plates overnight and individual colonies were archived by freezing within 15% glycerol 2×YT selective media. These libraries are available to the research community by the Indiana University Center for Genomics and Bioinformatics. Sequencing reactions were performed by priming at the 5' end of cDNA using vector primer pDNRlib30-50 (TAT ACG AAG TTA TCA GTC GAC G) and by priming at the 3' end using vector primer M13rev (AAA CAG CTA TGA CCA TGT TCA C) with ABI BigDye chemistry and the 3730xL sequencer. Vector and poor quality sequences were trimmed from the sequencing reads and ESTs were assembled into contigs using ESTPiper [S26]. This effort produced 89,140 reads from a total of 59,904 clonal plasmids. This method resulted in a gene discovery rate of 75% to 87% among the libraries and an average rate of 81%. EST sequences have been deposited in GenBank, accession numbers: FE274839-FE425949.

The ESTPiper program assembled 113,931 ESTs out of 148,410 sequences that passed quality assurance thresholds producing a unigene set of 14,891 sequences. The assembly to the *D. pulex* genome sequence scaffolds began first by using BLAT [S27] to find overlapping and

mate-paired EST clusters, then by using PASA [S21] to merge sets of compatible overlapping EST alignments to identify alternative splice variants. The following parameter options were applied: blat min. identity = 95%; blat max. intron = 750 Kb; clustering min. coverage = 80%; clustering min. overlap = 40 bp; clustering max. magnification = 10 bp. A PASA database was constructed for *D. pulex* (Table S1) that provides web access to EST assembly summaries and details, EST validation and correction reports for gene predictions, providing a useful reference for expert gene annotators.

3. Proteome sequencing

We sequenced over 11,000 peptides using two approaches.

1D Nano-LC Orbitrap approach – Animals were freeze-dried and solubilised in SDS Buffer (0.5 M Tris pH 6.8, 5% SDS, glycerol, milli-Q water, Bromophenol Blue, 10 mM DTT). After centrifugation at 100,000×g , 100 µg protein was subjected to separation by SDS-PAGE on a 12.5% maxi gel using the BioRad Protean II Electrophoresis system (BioRad, Veenendaal, Netherlands) using 60 V in the stacking layer, increasing up to 80 V during the separation. The gel was stained using Gelcode® blue stain reagent (Pierce, Rockford, USA) overnight and subsequently washed with milli-Q water. The lane was subsequently excised into 20 gel pieces and reduced with 6.5 mM DTT (Roche Diagnostics) followed by alkylation with 54 mM iodoacetamide (Sigma-Aldrich, St. Louis, USA) for one hour, to be then digested with trypsin at an enzyme: substrate ratio of 1:50 (w/w). Nanoflow liquid chromatography was performed on an Agilent 1100 HPLC binary solvent delivery system (Agilent Technologies, Waldbronn, Germany) with a thermostated wellplate autosampler coupled to an LTQ-Orbitrap mass spectrometer (Thermo Electron, Bremen, Germany). 30 mm × 100 µm Aqua C_{18} (Phenomenex, Torrance, CA) trapping column and a 200 mm × 50 µm Reprosil-Pur C_{18}-AQ (Dr. Maisch GmbH, Ammerbuch, Germany) analytical column . Peptides were trapped at 5 µl/min in 100% A (0.1 M acetic acid in water) on the Aqua C_{18} column for ten minutes. After flow-splitting down to ~ 100 nl/min, peptides were transferred to the analytical column and eluted with a gradient of 0-40% B (80% Acetonitrile/0.1 M Acetic Acid) in 40 minutes in a 60 minute gradient. Nanospray was achieved using a coated fused silica emitter (New Objective, Cambridge, MA) (o.d., 360 µm; i.d., 20 µm, tip i.d. 10 µm). A 33 M′Ω resistor was introduced between the high voltage supply and the electrospray needle to reduce ion current. The LTQ-Orbitrap mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS/MS. The two most intense peaks above a threshold of 500 were selected for collision induced dissociation (CID) in the linear ion trap at normalized collision energy of 35%. In the LTQ-Orbitrap full scan MS spectra (300-1500 m/z) were acquired with a resolution of 60,000 at 400m/z after accumulation to a target value of 500,000.

2D Nano-LC LTQ approach – Growth of daphniids (TCO isolate), protein preparation, SDS gel fractionation of 50 µg protein and in-gel digestion with Trypsin were performed as described in detail [S28]. The 2D-nano-LC separation of peptides derived from 10 SDS gel slices was performed on a multi-dimensional liquid chromatography system (Ettan MDLC, GE Healthcare, Piscataway, NJ). Chromatographic parameters for the first dimension were: 50 × 0.32 mm SCX column (BioBasic, Thermo Electron, Bremen, Germany), flow rate 6µL/min with 6 discrete salt plugs of increasing salt concentration (10, 25, 50, 100, 500 and 800 mM NH_4Cl in 0.1% formic acid and 5% ACN). The eluted peptides were bound on a RP trap column (C18 PepMap 100, 5µm, 300µm i.d. 5mm, LC Packings) and subsequently separated on the second-dimension RP column (C18 PepMap 100, 3µm, 75µm i.d. 15cm, LC Packings) with a 72min linear gradient (A: 0.1% formic acid, B: 84% ACN and 0.1% formic acid) at a flow rate of 260nL/min. Mass spectrometry was performed on a linear ion trap mass spectrometer (LTQ, Thermo Fisher,

Waltham, MA) online coupled to the nano-LC system. For electrospray ionization a distal coated SilicaTip (FS-360-50-15-D-20, New Objective, Woburn, MA, USA) and a needle voltage of 1.4 kV was used. The MS method consisted of a cycle combining one full MS scan (Mass range: 300-2000 m/z) with three data dependent MS/MS events (35% collision energy). The dynamic exclusion was set to 30 s.

Database searches and statistical data evaluation – MS/MS spectra of both approaches were converted to DTA files using Bioworks (Thermo, San Jose). Perl scripts were used to convert all spectra into a single file and searched using MASCOT search engine (Matrix Science, London, UK, Version 2.2.01) against *D. pulex* gene model databases (v1.1 or All Models) with cysteine carbamidomethylation and Methionine oxidation as a fixed variable modifications, respectively. A peptide mass tolerance of 5 ppm for Orbitrap spectra and 2 Da for LTQ data was used. As fragment mass tolerance 0.8 Da was selected and Trypsin was chosen as proteolytic enzyme allowing one missed cleavage. All data were loaded into Scaffold (version 02.01.00, Proteome-Software, Portland, OR) and was used to probabilistically validate peptide and protein identifications. Peptide and protein identifications were accepted when reaching 90% and 95% probability, respectively, requiring a minimum of two peptides per protein.

4. NimbleGen genome tiling microarray experiments

We used a set of two custom-designed Roche NimbleGen high-density-2 (HD2) whole genome tiling microarrays, each with 2.1 million isothermal long-oligonucleotide probes (50-75 nt in length) that sequentially overlap 30 bp, on average (NCBI GEO accession numbers GPL11200-GPL11201). Included are 225,000 markov modeled random probes sharing base compositions equivalent to the *Daphnia* sequences represented by the experimental probes. These random probes are used to set appropriate thresholds that measure significant hybridization signals over the background. All experimental probes were designed from unique regions of the genome sequence using the NimbleGen ArrayScribe software and the quality assurance tests of the probes were conducted by CGB in-house algorithms. Experiments conducted on this tiling array are used to (1) validate the frozen gene sets of the current genome annotation, (2) improve the predicted gene structures by empirically determining UTRs and intron-exon boundaries, identifying missing upstream, internal, and downstream exons and alternative transcripts, (3) propose gene structure models in transcribed regions containing no predicted genes and (4) delineate transcriptionally active regions of the genome from intergenic, intronic and genic regions. Signal to background ratios were determined by first calling probes that fluoresced at intensities greater than 99% of the random probes' signal intensities; therefore only 1% of fluorescing experimental probes should be false positives. The arrays reliably produced high signal to background ratios; $\log_2$ ratios of eight were observed for signal over background.

We conducted two-color competitive hybridizations that measure differential expression from three replicates, each using RNA from independent biological extractions of (1) adult males *vs* adult females, (2) 4th instar juveniles responding to kairomones by the dipteran predator *Chaoborus americanus vs* controls, (3) 11 day old animals exposed to four metals separately for 24 hours *vs* controls, and (4) four week old animals who were exposed for 21 days to cadmium *vs* controls.

Comparing the sexes – We used *Daphnia pulex* isolate TCO for comparing adult male and female transcriptomes. Animals were reared in filtered lake water at 20°C and a 12:12 light/dark cycle at a density of approximately 1 individual per 5 ml. Animals were fed *Scenedesmus* algae at approximately 0.1 mg ml^{-1} each day and split into two groups of 20. One group was exposed

to 400 nM methyl farnesoate in methanol (60 µl L-1), which is known to reliably induce male production [S29], while the other group of 20 individuals were untreated. Progeny were raised under conditions described above in common beakers, with about 25 individuals per beaker and inspected by microscopy to verify healthy appearance and the development of animals of both sexes. After 14 days, adult males and females were sacrificed. Total RNA was isolated using Trizol (Invitrogen) and RNeasy columns (Qiagen), including a DNase treatment performed on-column. Quality of total RNA preparations was assayed by spectrophotometry and by the Bioanalyzer 2100 system (see Bioanalyzer section of [S30]). Three biological replicates were compared. Two female replicates were labeled with Cy-3 (green) dye, therefore two male replicates were labeled with Cy-5 (red) dye, while the third replicate consisted of a dye flip.

Exposure to kairomones – We used *Daphnia pulex* clone R9 (isolated from arctic Canada by Dr. Larry Weider) for our kairomone experiments. This clone was shown to respond to chemical cues from *Chaoborus* by producing distinct neckteeth [S31]. We conducted the experiments in two separate labs under different culture conditions and slightly different induction protocols. The three biological replicates for each experimental condition were then mixed before the RNA was extracted. This allowed us to focus our search on genes that were up or down-regulated simultaneously under all induction conditions. In both set-ups, we cultured the *Daphnia* and conducted the experiments in artificial medium (consisting of local tap water, ultra pure water and trace elements) under fluorescent light in climate controlled rooms at 21°C. *Scenedesmus acutus* was used as food and offered at non-limiting concentration to stimulate offspring production in the cultures.

In one lab, we simultaneously raised cohorts and placed a mixture of 150 differently aged adult mothers into 3L borosilicate glass beakers. We had a control and an induction treatment. In the induction treatment, 60-70 fourth instar *Chaoborus flavicans* larvae were placed into a net cage hanging into the experimental beakers. The net prevented direct predation but allowed all chemical cues to pass. The *Chaoborus* larvae were fed with first instar *D. pulex* from the cultures because the kairomone production depends on actively feeding *Chaoborus* larvae. Dead larvae were replaced and prey remnants were removed daily. We collected all offspring produced by the mothers in two day intervals, ensuring that the *Daphnia* were in the first two juvenile instars where they are still inducible. Offspring released during the initial two days were not used because their induction time had been two short. We verified neckteeth production by checking subsets of the offspring under a dissecting microscope. All juveniles in the induction treatment carried neckteeth. We harvested all offspring produced during the next 10 days (500 – 150 animals every second day per beaker). The control treatments had identical setups with the exception that the net cages contained no predators. Both treatments had four independent replicates. The animals were directly transferred to Trizol and frozen at -80°C.

Experiments in the second lab deviated in some minor aspects. Animals were raised in 1.5L beakers containing age-synchronized females and net cages. 15 larvae of *Chaoborus flavicans* were placed into the net cages for the induction. After releasing their offspring, the mothers were removed and the offspring stayed in the treatments. Offspring were harvested after molting to the second instar. Induction was checked under a dissecting microscope. Both treatments had three independent replicates. The animals were directly frozen in artificial medium at -80°C. The RNA was isolated and treated as above. Two kairomone treatment samples were labeled with Cy-3 (green) dye, therefore two control replicates were labeled with Cy-5 (red) dye, while the third replicate consisted of a dye flip.

Exposure to metals – We used *Daphnia pulex* isolates TCO and PA33 (from Portland Arch nature preserve in Lafayette, Indiana) for comparing the transcriptome of stage-specific adult

females challenged by metals to that under no stress. The experiment followed a protocol described in an earlier study [S32]. Animals were reared in 3.5L borosilicate glass beakers (25 per beaker) held at a constant temperature (20 ± 1°C) and photoperiod (16:8 light-dark). The animals were maintained in nanopure water reconstituted to moderate hardness [S33] and renewed weekly. They were fed *Scenedesmus* algae daily at a concentration of 75,000 cells/mL. Our pre-experimental procedure consisted of maintaining cultures of neonates (< 24 hours old) for one generation prior to the metal exposure to control for maternal effects [S34]. These animals are referred to as 'brood females', which were synchronized with respect to time of maturity for producing neonates for the metal experiments.

We conducted a chronic (16-day) exposure experiment to cadmium. Test solutions were prepared immediately prior to use with culture media from stocks made with $CdCl_2$ (analytical grade, Sigma Chemical, St. Louis, MO, USA) dissolved in deionized water. Three independently replicated *Daphnia* microarray experiments used 24 hour old animals exposed to nonlethal concentrations of cadmium (0.5 µg Cd/L) and control conditions in batches of 50 *Daphnia* per 3.5L exposure chamber. Earlier experiments showed that this concentration inhibits reproduction by ~30%. The animals were directly frozen in artificial medium at -80°C. The RNA was isolated and treated as above. Two cadmium treatment samples were labeled with Cy-3 (green) dye, therefore two control replicates were labeled with Cy-5 (red) dye, while the third replicate consisted of a dye flip.

In another tiling array experiment, adult *Daphnia* (17-24 d) were acutely exposed (24-h) to one of five metals (arsenic, 1384 µg/L; cadmium, 20 µg/L; copper, 1 µg/L; nickel, 200 µg/L; zinc, 200 µg/L) or control conditions were identical, but lacked metals. The metal concentrations in these tests were demonstrated to be non-lethal over the acute exposure period. Arsenic and copper experiments were conducted with TCO. Copper, nickel, and zinc experiments were conducted with PA33. All *Daphnia* were exposed in batch with 25 individuals housed per 3.5L. Batch number was optimized to provide adequate sample mass for molecular evaluation (e.g., 1 adult *Daphnia* equals 1 µg of total RNA). Each exposure included four replicate beakers per treatment and control. Culture conditions followed those previously described. RNA was extracted from each sample and pooled in equal-molar amounts from the five treatments and controls to form two groups (e.g., metal, control). Replicates within these groups were independent, as pools were randomly constructed from individual biological replicates obtained for each exposure condition.

For both experiments, three biological replicates were compared. Two metal exposure replicates were labeled with Cy-3 (green) dye, therefore two control conditions replicates were labeled with Cy-5 (red) dye, while the third replicate consisted of a dye flip.

RNA sample processing and analysis of data – Beginning with at least 0.5 µg of total RNA, a single round of amplification using MessageAmpTM II aRNA kit (Ambion) produced more than 100 µg for all other tissue types. Starting with 10 µg of cRNA, double strand cDNA synthesis was carried out using the Invitrogen SuperScript Double-Stranded cDNA Synthesis kit using random hexamer primer followed by DNA labeling using 1 O.D. CY-labeled random nonomer primer (either Cy3- or Cy5-coupled) and 100U Klenow fragment (3>5 exo) per 1 µg double-stranded cDNA (see NimbleGen labeling protocol for gene expression contained in the following PDF available from the NimbleGen website: exp_uerguide v3p2.pdf). Each treatment and control was differentially labeled and a dye-swap was included among the replicate experiments. Dual-color hybridization (15 µg of both Cy-labeled samples), post-hybridization washing and scanning were done according to the manufacturer's instructions (exp_userguide_v3p2.pdf). Images were acquired using an Axon GenePix 4200A scanner

(Molecular Devices, Sunnyvale CA) with GenePix 6.0 software. The data from these arrays were extracted using the software NimbleScan 2.4 (Roche NimbleGen, Inc., Madison, WI).

Transcriptional active regions (TARs) were defined by stringing together overlapping probes showing fluorescence above a 1% false positive rate (FPR). First, replicate arrays were quantile-normalized [S35] and to each probe the median value of the replicate probe values was assigned. The fluorescence signal of 225,453 random probes, designed to reflect the genome nucleotide composition by Markov modeling, was used to determine a FPR threshold. Probes were considered positive if their fluorescence signal was higher than the 99th percentile of the fluorescence signal of the random probes. (Fluorescence signal of 275,000 probes from 3,889 scaffolds likely to be from bacterial DNA were also assessed. Only 1.8% of those mostly bacterial probes had signal above that 1% random probe FPR cutoff.) Contiguously transcribed elements, TARs, were generated similarly to the approach developed in [S36]. Positive probes were joined into a TAR if they were adjacent (maxgap=0, no intermittent non-positive probe) and a TAR's length had to be at least 45 bp (minrun=45, mid-point first positive probe to mid-point last positive probe, resulting in at least 3 adjacent positive probes for a TAR).

The exons or genes were deemed to be transcribed only when greater than 80% or their tiled length was expressed. Genes validated by tiling array or EST data are shown in Table S11.

The data analysis to measure differential expression of genes and of unannotated TARs was performed using the statistical software package R [S37] and Bioconductor [S16] with additions and modifications. The signal distributions across chips, samples and replicates were adjusted to be equal according to the mean fluorescence of the random probes on each array. All probes including random probes were quantile-normalized across replicates. Expression-level scores were assigned for each predicted gene based on the median $\log_2$ fluorescence over background intensity of probes falling within the exon boundaries. This following analysis protocol was used for estimating differential expression of genes and other genome features from tiled expression data. (1) We created a "tile-expression" table containing normalized $\log_2$ expression scores for each oligonucleotide probe, with columns for each treatment and replicate, as well as the designated genome location (or address) of each probe. (2) We next created a "tile-gene-mapping" table, in the same sorted order as the tile-expression table, which has columns of gene IDs for each exon, intron, tar-region, in rows matching the address of each probe. (3) We calculated the per-tile, per-treatment differential expression (DE) levels with LIMMA R package [S38]. This balanced-design DE calculation is of the same type that LIMMA is designed to produce. (4) Using in-house algorithms, we combined the per-tile DE results using the tile-gene mapping table to produce statistics for each gene, gene-intron, tar-region of interest that include M and A expression estimates, t-statistic, and probability. The data are deposited at NCBI GEO under the accession GSE25823.

5. Transcription profiling using NimbleGen multiplex microarrays

We employed the 12-plex gene expression microarray described above (section I.4) for additional higher-throughput gene expression experiments. Our protocol on the use of this microarray platform for two-color hybridizations – comparing one conditions versus another – is described in a technical report [S30].

To investigate the evolution of gene expression, we gathered twelve microarray datasets that were produced using the same protocol by the same person (J. Lopez, CGB). We compared the gene expression patterns of four to six replicates of *D. pulex*: coping with 0.5 µg/L Cd, with 1.5 mgC/L of a 1:1 mixture of *Microcystis* and *Ankistrodesmus*, with a Cd/microcystis mixture, and

with 5.33 g/L $NaCl_2$. For each condition we compared the expression response of adapted and non-adapted isolates to control exposures. We also exposed a non-adapted isolate to acid stress (pH 6), and compared young and geriatric isolates. Results from each of these 12 experiments are presented more fully within companion studies (added to [S39]).

After hybridizations and scanning, the data from each experiment were extracted using NimbleScan v2.4 software (Roche NimbleGen, Inc., Madison, WI) and imported into an in-house analysis pipeline using Bioconductor for normalization and analysis [S16]. All probes including random probes were quantile-normalized across chips, subarrays, samples and replicates. Differential expression was assessed using LIMMA and EBarrays [S38, S40] using the median signal of probes representing genes. EBarrays uses a parametric mixture model to calculate the posterior probability of differential expression for arbitrarily complex experimental designs. This method was applied to each experiment. To determine the significance of expression differences, and adjust for multiple testing, we calculated the False Discovery Rate using the Benjamini-Hochberg method [S41] for each gene using the Bioconductor LIMMA package. The data are deposited at NCBI GEO under the accession GSE25823.

6. Annotating protein-coding genes

All predicted protein-coding gene models were functionally annotated by homology to annotated genes from the NCBI non-redundant set and classified according to Gene Ontology [S42], eukaryotic orthologous groups [S43], KEGG metabolic pathways [S44] and phylogenomic gene clustering [S45]. The automated annotation is followed by a distributed community-wide manual curation. The JGI Portal provides tools for web-based manual curation that enables a search for the gene of interest, validation of predicted gene structures, correcting and *de novo* model building with the correct structure, and correcting and/or providing additional details on functional annotation.

Manual curation is focused on either specific genes searchable by keyword or BLAST or groups of genes from metabolic and regulatory pathways (KEGG browser), functional categories of eukaryotic clusters of Orthologous Groups (KOG, via KOG browser) and molecular functions, biological processes or cellular components of Gene Ontology (GO via GO browser). At every locus, curators assess the quality of the predicted gene models using available supporting evidence on the DNA level displayed on the genome browser (ESTs, homology, genome conservation, etc.), or on the protein level (protein and alignments, domains, completeness), or through additional custom analysis (e.g., multiple alignment). After these assessments, the best available model is selected for the final minimum gene set (gene catalog v1.1). In the absence of models of sufficient quality, the models are edited or created *de novo* to be included in the gene catalog. Annotation data were submitted for1,688 manually curated genes and 523 novel or structurally modified genes. Gene annotations are deposited at DDBJ/EMBL/GenBank under the accession ACJG00000000.

7. Annotating non-coding RNA and transposable elements

Automated searches for non-protein-coding loci added more characterized loci to the v1.1 gene builds. To estimate the repeat copy number for the rRNA arrays, we mapped by using BLAT all homologous reads from the TCO shotgun genome dataset to a reference sequence for the ribosomal RNA genes. The average coverage of shotgun reads for the rRNA repeat in this analysis was 4,120×. Given that the average genome-wide coverage is 8.7×, we estimate that the number of copies for the rRNA repeat is ~468. A similar analysis of the TRO shotgun reads suggests a repeat copy number of ~500.

We located tRNA genes using the Aragorn [S46] and rRNAscan-SE [S47] algorithms, which generated counts of 3,983 and 5,440 loci respectively. The combined analysis identified an overlapping set of 3,798 tRNAs. These annotated tRNA gene models are mapped to the genome sequence using Gbrowse at wFleaBase (Table S1).

Micro-RNA (miRNA) loci in the *D. pulex* genome (Table S16) were identified using a pipeline that uses Support Vector Machine models, homology and an orthology procedure [S48].

Transposable element (TE) content in *D. pulex* was determined using a two-step process. Consensus sequences were identified using various programs and used to build a library which was subsequently used to mask the genome to estimate the proportion of the genome comprised of TEs. Long terminal repeat (LTR) retrotransposons were located using MGEScan-LTR, a de novo identification method based on string pattern matching and profile hidden Markov model [S49]. MGEScan-LTR identified full-length elements having LTRs at both ends, and clustered them into families by using threshold parameters of 80% identity of reverse transcriptase (RT) protein sequences. The program mainly found elements in *Gypsy*, *Copia*, and *Bel/Pao* clades. In order to identify DIRS elements, protein domain searching was used with RT and tyrosine recombinase (YR) as queries. Non-LTR retrotransposons were identified using MGEScan-nonLTR, a probabilistic model for finding the protein domains for RT and endonuclease [S50]. Stop codons and frameshift mutations were allowed in this search. The elements identified were subsequently clustered into families by using the threshold parameters of 80% identity of RT protein sequences. DNA transposons were identified using a combination of complementary approaches including protein homology, RepeatScout – a de novo repeat identification tool [S51], and the classification tool, Repclass [S52]. At the final step, a library of consensus (representative) sequences from families was assembled and used with RepeatMasker to estimate the proportion of the genome comprised of TEs, including full-length copies, fragments (including solo LTRs), as well as non-autonomous families. The results of RepeatMasker estimate the proportion of the genome represented by each superfamily (Table S17-18) after filtering short fragments (length < 20% of query element for DNA transposons and non-LTR retrotransposons and length < 1,000 bp for LTR retrotransposons).

We visualized the number and genomic organization of an important transposable element within the repeated (and consequently unassembled) rDNA genes by fluorescence *in situ* hybridization (FISH; Figure S11). Preparation of chromosome spreads was performed as described previously [S17] with slight modification. Briefly, testes of adult males fixed in 4% paraformaldehyde were extracted and dissected in PBS, incubated in PBS containing 0.5% Triton-X, then briefly incubated in water. The tips of testes were gently torn in 4% paraformaldehyde, and squashed under a coverslip. After freezing the sample on dry ice with the coverslip facing up, it was removed.

DNA fibers were prepared from oocyte nuclei. Ovaries of adult *D. pulex* females were extracted in PBS, and an oocyte was isolated with forceps. The oocyte was placed in NDS [1% (wt/vol) sodium lauroyl sarcosinate, 0.5M EDTA, 10 mM Tris] on a slide and incubated for 10 min. DNA fibers were mechanically spread on the slide using the edge of a coverslip, and the slides were put on a heat block at 65°C to dry. The slides were then washed briefly in PBS and fixed in ethanol.

Labeling of probe DNAs and hybridization were performed as described previously [S17]. PCR product of the *D. pulex* IGS was labeled using the Bio-Nick labeling system (Invitrogen, Carlsbad, CA) for labeling with biotin-14-dATP. Pokey element from *D. pulex* was labeled using the DIG-nick translation mix (Roche) for labeling with digoxigenin (DIG)-11-dUTP. Hybridization

mixture [50 % (v/v) formamide, 10% (v/v) dextran sulfate, 100 ng/µl salmon sperm DNA, and 0.1-0.2 µg labeled probe DNA in 2×SSC] was applied to the specimen, covered with a coverslip, and sealed with rubber cement. After the rubber cement solidified, the slide was heated for denaturation on a heat block at 80°C for 6 min, and incubated for hybridization at 37°C in a humid chamber for 72 hrs. After hybridization, the rubber cement was peeled away and the slide was immersed in 2×SSC to float the coverslip off. Subsequently, the slide was washed once for 15 min in 50% formamide dissolved in 2×SSC at 37°C, twice for 10 min in 2×SSC, once in 4×SSC for 5 min at room temperature, and then blocked with 4% Block Ace (Dainippon Sumitomo Pharma, Osaka, Japan) in 4× SSC for 15 min at 37°C. Hybridization of biotin-labeled probes was detected with goat anti-biotin antibody (Vector Laboratories, Burlingame, CA), followed by staining with Alexafluor 488 rabbit anti-goat IgG antibody (Molecular Probe, Invitrogen, Carlsbad, CO). Hybridization of digoxigenin-labeled probes was detected with mouse anti-digoxigenin (Roche Diagnostics GmbH, Mannheim, Germany), followed by staining with Alexafluor 594 rabbit anti-mouse IgG antibody (Molecular Probe, Invitrogen, Carlsbad, CO). Each antibody was diluted in 4×SSC containing 1% Block Ace at the concentration suggested by the manufacturer. Incubation for detection was 1 hr at 37°C, followed by washing for 10 min in 4×SSC, for 15 min in 4×SSC containing 0.1% Triton X-100, and for 10 min in 4×SSC at room temperature. Staining was done for 45 min at 37°C, followed by washing for 10 min in 4×SSC, for 20 min in 4×SSC containing 0.1% Triton X-100, for 20 min in 4×SSC, and for 5 min in 2×SSC at room temperature. Finally, the specimens for chromosome FISH were counterstained with DAPI mounted in an antifading solution, Vectashield (Vector Laboratories, Burlingame, CA). The specimens for fiber-FISH were mounted in Vectashield. Observations were made on a Nikon Ecripse 80i microscope equipped with a motorized Z axis. Images were captured with Photometrics HQ using Metamorph software.

III. Attributes of a Compact Genome

1. Comparing genome structures

Gene structures were measured for EST-validated gene models of *D. pulex* and compared to gene structures of six insects plus two non-arthropods (*Acyrthosiphon pisum*, *Apis mellifera*, *Nasonia vitripennis*, *Tribolium castaneum*, *Anopheles gambiae*, *Drosophila melanogaster*, *Mus musculus*, *Caenorhabditis elegans*) (Table S19). PASA [S21] was used first for EST assembly and for the production of cDNA-gene models. PASA also provided a method of validating gene models from the EST assemblies. The structure statistics were produced by processing gene exon locations with Perl and R language scripts that tabulate exon, intron and coding exon locations per gene. The data and software are deposited at [S53]. A table of arthropod gene structure statistics is updated with new genome data, as available at [S54, S55](Table S1).

2. Comparative study of intron evolution

Clusters of probable orthologous genes were constructed for nine animal species, including six arthropod genomes, two genomes of vertebrates, and the only available cnidarian genome (Table S20). Orthologous relationships were established by comparing the complete sets of protein sequences from these animals using a modification of the previously described method [S56]. If there was more than one gene from a particular species in any putative orthologous set, the ortholog with the highest similarity to the rest of the proteins in the cluster was chosen [S57]. Therefore, each of the clusters contained exactly one sequence from each species. Clusters that included sequences with obvious annotation errors (e.g., incorrectly assembled genes) were discarded. When applied to the six arthropod species, this approach yielded 3,936 clusters of likely orthologous groups. Adding the remaining three species yielded 2,946 clusters.

Sequences from each orthologous cluster were aligned using MUSCLE [S58]. The protein sequence alignments were converted back to the corresponding nucleotide sequence alignments, and intron positions were mapped onto the alignments [S59]. Only those positions without gap within five amino acids on either side were included in the calculations to prevent errors caused by misalignment. The intron presence-absence matrices were then constructed from such verified intron positions for each species, and intron gain and loss events were inferred using a maximum likelihood (ML) method [S60](Table S24).

IV. Origin and Preservation of *Daphnia pulex* Genes

1. Assigning gene homologies

For the comparative study of the *D. pulex* repertoire of protein-coding genes, we used Smith-Waterman alignment algorithm as implemented in Paralign (Sencel Bioinformatics, Oslo, Norway) to search for homologous genes in *Tribolium castaneum* (beetle), *Drosophila melanogaster* (fruitfly), *Pediculus humanus* (louse), as well as *Strongylocentrotus purpuratus* (urchin) *Gallus gallus* (chicken), *Xenopus tropicalis* (frog) and *H. sapiens* (human). Using these all-against-all gene comparisons, we identified orthologous gene relations, i.e. gene lineages originating from the last common Bilaterian ancestor of these species, using the OrthoDB procedure [S61]. It employs a clustering approach of best reciprocal hit triangles with an e-value cutoff of $1 \times e^{-3}$, and tuples with cutoff of $1 \times e^{-6}$, that are expanded to include all more closely-related within-species homologs and require all member sequences to overlap by at least 30 amino acids. This procedure has been scrutinized as part of several genome projects [S62, S63, S64, S65], and the extensive manual examination of orthologous groups in *Daphnia* [S66, S67, S68, S69, S70, S71, S72, S73, S74] and in other species [S75, S76, S77, S78] has confirmed their accuracy.

An interactive data-mining tool was created to explore orthologous gene sets among the proteomes of all sequenced arthropods [S79] including *D. pulex*, *Ixodes scapularis* (tick), *Acyrthosiphon pisum* (pea aphid), *P. humanus* (louse), *Aedes aegypti*, *Anopheles gambiae*, *Culex pipiens* (mosquitoes) *Apis mellifera* (honeybee), *Nasonia vitripennis* (wasp), *T. castaneum* (beetle) and three drosophiliids: *D. melanogaster, D. pseudoobscura, D. mojavensis*. An all-against-all protein similarity searches using BLAST was performed [S23]. Small (<40 amino acid) proteins and alternative transcripts were removed to only use the most similar gene variants; the discarded sequences included 6,500 alternate transcripts for *D. melanogaster*, 1,300 from A. *aegypti*, and fewer than 800 from all others. The similar genes were clustered using the standard methods outlined for OrthoMCL [S80, S81], which can be summarized as follows. Significance criteria were applied with recommended options: a similarity e-value $\leq e^{-05}$, protein percent identity ≥40%, and MCL inflation of 1.5 (influencing the granularity of the clustering). Reciprocal best similarity pairs between species, and reciprocal better similarity pairs within species (i.e., recently arisen paralogs, or proteins that are more similar to each other within one species than to any protein in the other species called in-paralogs) were added to a similarity matrix. The matrix was normalized by species and subjected to Markov clustering (MCL; [S82]) to generate ortholog groups including recent in-paralogs. An additional round of MCL clustering was applied to link related gene groups.

Finally, the Superfamily annotation [S83] was explored to verify patterns of gene family expansions observed by the above methods. Superfamily is based on a collection of hidden Markov models representing structural protein domains at the SCOP superfamily level. The results of all three investigations are available online (Table S1).

Results from these methods were verified to be consistent with the gene tree procedure PhIGs [S45]. PhIGs conducts a true phylogenetic analysis using maximum likelihood. Briefly explained, PhIGs performs these steps: (1) an all-by-all Blast search of the inferred amino acid sequences of each gene model of each considered genome, (2) extension to a full-length alignment of each significantly similar pair using MUSCLE [S58], (3) scoring of the similarity among each pair, (4) building a graph with each sequence as a node and the scores of the pairs as edges, (5) specifying the deepest ingroup versus outgroup relationship, (6) building clusters of gene families by noting the distance between each set of ingroup-outgroup gene pairs then doing a single-linkage clustering of all genes of ingroup organisms that have smaller distances, (7) successively moving through each descendent node of the tree of organisms, in each case specifying the new set of ingroup-outgroup relationships and repeating the clustering, (8) creating a multiple sequence alignment of each cluster, (9) performing a series of quality control measures, considering such things as total length of the multiple sequence alignment and eliminating highly gapped positions using GBlocks [S84], (10) creating a maximum likelihood evolutionary tree of each gene cluster. The complete gene sets from 14 genomes used for this analysis are: the protist *Monosiga brevicollis*, the cnidarian *Nematastella vectensis*, *Homo sapiens*, the teleost *Takifugu rubripes*, the urochordate *Ciona intestinalis*, the nematode *Caenorhabditis elegans*, the mollusk *Lottia gigantea*, the polychaete *Capitella capitata*, the oligochaete *Helobdella robusta*, the dipterans *Drosophila melanogaster*, *Anopheles gambiae*, and *Aedes aegypti*, the coleopteran *Tribolium castaneum*, and *Daphnia pulex*. The PhIGs results can be downloaded from [S55].

2. Studying the history of gene family expansions and losses

The gene families of hypothetical ancestral species were reconstructed by a step-wise detection of BRH – here also called the symmetrical best alignments (sym-bets) – for each of the ancestral species. This comparison of gene families among the ancestral species of the phylogeny provides a hypothesis for the timing of gene duplication and loss events throughout evolution. We used Evolmap [S85] to elucidate these events, which is an algorithm that reconstructs sym-bets and localizes the gene duplications and losses to the most parsimonious branch of the phylogenetic tree by assuming a known species history and by applying the Dollo parsimony criterion. We applied Evolmap on 11 species (Table S27) using *Nematostella vectensis* as the outgroup for the assumed species phylogeny from Figure 1C.

3. Studying the history of gene duplication

To characterize the evolutionary pattern and rate of gene duplication, we compared the protein coding genes (Dappu v1.1, n=30,940) to one another using a modified installation of Genome History [S86], which measures substitution patterns between gene copies in the context of gene family assignments. Our study included other genomes for comparative insights. The entire gene catalogue from *C. elegans*, and *H. sapiens* were downloaded from Ensembl [S87] For genes with multiple splice variants, the largest gene was chosen. Transposable element genes were excluded to the extent that they could be identified.

Genome History (GH) detects and compares gene duplicates within a genome by using a set of user-specified parameters and input. The following protocol was followed:

1. All predicted protein sequences were compared to each other using WU-gapped-BLASTp. Self-alignments were discarded and alignments better than e^{-10} proceeded to next step.
2. Gene matches were aligned using ClustalW [S88] with restrictions set at a minimum alignment length of 100 amino acids and percent identity greater than 40%. These strict

settings minimized false relationships due to highly conserved motifs and narrowed the focus of this study to recent gene duplicates (K_s < 1).

3. Each aligned gene pair was then back-translated using the nucleotide gene file. For each pair, K_a (substitutions / replacement-site) and K_s (substitutions / silent-site) were calculated using the maximum likelihood, codon-based model [S89].

Birth rates of gene duplicates were calculated using the number of single-pair duplicates in the youngest cohort (K_s < 0.01), the baseline number of single copy genes and the synonymous substitution rate (K_s), providing units of duplications/gene/K_s. Birth rates of nematodes and humans were comparable to those found in earlier studies [S90]. *D. pulex* appears to have a higher rate of gene duplication than other animals studied to date (see Table 8.1 in [S91]).

While the observed number of new duplicates can be used to estimate a birth rate, it should be considered a downwardly biased estimate, since observed duplications may represent a subset of events that rose to high frequency in the population, and were not purged by selection. Additionally, an accurate gene birth rate must also account for gene losses over the measured interval (K_s = 0.0-0.01), which can be inferred, assuming steady-state birth/death rates, from an estimate of instantaneous mortality rate using the slope of the regression of duplicate numbers at time t (n_t) on synonymous substitution rate (K_s) [S91]. Birth rates estimates that account for losses give slightly higher values (5-20% higher), but do not affect the phylogenetic pattern of estimated rates (*D. pulex* > *H. sapiens* > *C. elegans*).

4. Measuring the distribution of duplicated genes using *Tandy*

Tandem duplicated genes can be nearly identical (>95% identity), arranged in very close proximity to one another (within the length of introns), produce regular signals of genome structure evolution and may be linked to interesting biology. Yet, software that relies on alignment with gapping produce poor gene models from repeated high-identity exons. Gapped alignments often mistakenly merge exons from neighboring genes into gene models. Therefore, *Tandy* software was developed to address problems of accurately predicting genes when arranged within tandem duplicated gene (TDG) clusters [S92]. The *tandy* approach compares exons, and secondarily predicted genes and proteins, to locate all duplicates in a region. Gene predictors typically call exons with greater success than their calls of full gene models because exon matches are made without gaps.

After identifying all predicted exons, *tandy*'s algorithm marked runs of duplicate exons. These marked exons were then combined and split into duplicate gene models based on a heuristic method that uses (a) inter-gene *versus* intron distances, (b) runs of exon sets (e.g. exons 1, 2, 3 of a gene model that are repeated), and (c) gene start/stop exons and strand inversions. *Tandy*'s final output was a GFF feature file of duplicated regions, of gene models and of the exon matches per gene model. Duplicates were then classified based on their relative distance from one another (<15 Kb), based on the number of intervening genes, based on gene predictions and several quality measures.

Tandy was applied to produce comparative results using the well-studied genomes of *C. elegans*, *D. melanogaster*, 11 other *Drosophila* genomes, and *D. pulex*. Recent improvements add protein predictions to identify duplicates. Although these have a higher error rate than exon predictions, when one protein of duplicate set is well modeled, it can find other duplicates. The *Tandy* results were also used as evidence for gene prediction software to indicate gene boundaries.

5. Identifying lineage specific gene family expansions

Groups of orthologous genes were delineated by the OrthoMCL method [S79] described above (section IV.1). Lineage-specific gene family expansions were defined as orthologous groups with multiple copies in *Daphnia* whose numbers are significantly greater than those of insects and tick ($p < 0.05$) based on 2,000 random permutations of exact probability, without correction for multiple testing (Table S26). To identify independent gene-family expansions in *D. pulex* and among the three mosquito species, the same test was repeated for each of these four species against the distribution of gene copy numbers of the remaining arthropod taxa.

6. Annotating and tracing the phylogeny of opsins

Sequence similarity searches against the *D. pulex* v1.1 gene set were performed by BLASTp [S23], using protein sequences of each *D. melanogaster* opsin gene of interest from FlyBase [S93] as "bait". The searches retrieved top best matches until *D. pulex* models outside the subfamilies of interest were obtained. Each *D. pulex* gene identified from this search was manually annotated with reference to the draft genome assembly and assigned to a subfamily by inclusion in maximum likelihood phylogenies.

We performed three separate phylogenetic analyses to understand the evolution of these *Daphnia* opsins (Figures S21-22). First, we analyzed diverse representatives of the major opsin clades, including ciliary, rhabdomeric (Gq) and RGR/Go opsins. In this analysis, we also included all opsins recently described from the branchiopods *Triops longicaudatus*, *T. granarius*, and *Branchinella kugenumaensis* [S94], plus opsin sequences from two crustaceans, a copepod *Tigriopus californicus* and an ostracod *Vargula tsujii*, which are included in Figure S21. Accession numbers are given in Table S32. To determine opsin sequences from these two crustaceans, we first used Trizol (Invitrogen) to extract total RNA from the copepod *Tigriopus californicus* provided by Ron Burton of the Scripps Institution of Oceanography, and from the ostracod *Vargula tsujii* collected from baited traps set near Cabrillo Beach, San Pedro, CA (33.706,-118.279). For the copepod, we first performed degenerate RT-PCR with a 48C annealing temperature using primers LWF1a (TGGTAYCARTWYCCICCIATGAA) and OPSRD (CCRTANACRATNGGRTTRTA), then performed a hemi nested reaction on this PCR product diluted 1:10 with primers LWF1 and Scylla (TTRTAIACIGCRTTIGCYTTIGCRAA). For the ostracod we used primer SLF [S95] for degenerate 3' RACE. We sequenced the initial products to enable design of species-specific opsin primers. These gene specific primers allowed for successful 5' and 3' RACE reactions and subsequent cloning and bidirectional sequencing of fragments representing an entire opsin for each species. For this phylogenetic analysis, we aligned opsin proteins using MUSCLE [S58], then estimated the most likely tree using RaxML [S96], while assuming the WAG+I+Γ model. We performed bootstrapping with 100 pseudoreplicates (Figure S21). This phylogeny is rooted with ciliary opsins as the outgroup, following [S97].

We also studied *Daphnia* opsin evolution using two analytical approaches matching those of a companion paper [S66] on the evolution of other multiple gene families involved in vision and eye development (Figure S22). The first approach produced a maximum likelihood analysis of rhabdomeric-clade *Daphnia* opsins (Figure S22A), plus close related genes found when using the *Daphnia* opsins to search Uniprot databases [S98]. The tree is rooted with arthropsin according to Figure S21. In addition, Figure S22B presents a maximum likelihood analysis of rhabdomeric-clade *Daphnia* opsins, plus closely related genes from 19 metazoan genomes, and rooted with arthropsin (see [S66] for methodological details of these companion analyses).

V. Implications *Daphnia*'s Genome Structure

1. Finding non-allelic gene conversion events

To determine how much concerted evolution has shaped the patterns of divergence among duplicated genes throughout the *Daphnia* genome, we compared gene conversion features and rates of gene conversion in *D. pulex* to those of five species of *Drosophila*. The original data set comprised 14,653 paralogous *D. pulex* genes from 2,259 gene families. These genes were used to make 66,501 pair-wise alignments of the coding sequences, which were subsequently processed to remove regions of low similarity, including gaps. The latter step is required to eliminate regions with very high divergence in the alignments, which could elevate the rate of false positives. Furthermore, as this filtering process can shorten the alignments to a large extent and possibly introduce some bias in the data set, only alignments that retained 50% or more of their original length after this step were further analyzed. The final set included most of the original data (13,330 genes grouped in 55,362 pair-wise alignments).

Gene conversion among *D. pulex* paralogous genes was investigated using the program Geneconv v.1.81 [S99], which was run using all default settings, except for the addition of the option to display pair-wise p-values and the option to include monomorphic sites in the calculation. The latter option allows the program to take into account constant sites and is required to examine alignments containing only two paralogs. The significance level is determined based on 10,000 permuted datasets. All fragments identified with $p < 0.05$ were regarded as gene conversion events. The initial Geneconv output included 11,659 pairs from 6,943 genes. Of these genes, many were present in 10 or more converted pairs. We removed those pairs because such multiple conversion events between paralogs are highly improbable. The same threshold was applied for our analysis of gene conversion in *Drosophila* species. Rates of conversion were calculated as the ratio between gene pairs with conversion over the total number of screened pairs per species. The genetic divergence (number of synonymous substitutions per synonymous site or K_s) between paralogs was estimated by the maximum-likelihood method implemented in the program codeml from the package PAML [S100]. To correct the genetic divergence in converted pairs, we multiplied the original K_s value by the ratio between the alignment length and the length of the alignment minus the conversion tract. Several aspects of gene conversion were compared between *D. pulex* and five *Drosophila* species: *D. melanogaster*, *D. yakuba*, *D. pseudoobscura*, *D. virilis* and *D. grimshawi*.

2. Annotating and tracing the phylogeny of hemoglobins

Sequence similarity searches for hemoglobin genes against the *D. pulex* v1.1 gene set were performed as described above in finding opsin genes. Each *D. pulex* gene identified from this search was manually annotated with reference to the draft genome assembly. The *D. pulex* genome is found to contain 11 recognizable di-domain Hb genes (Table S39). Eight of the *D. pulex* Hb genes (named Dpul-Hb1 to Dpul-Hb8) are organized in tandem within a 23.6 kb region on scaffold 4 (chromosome 7 based on the single Dp112 marker of the genetic map). Their arrangement along the same coding DNA strand is interrupted only by a non-protein encoding gene between Hb4 and Hb5. The eight clustered genes plus Dpul-Hb9 on Scaffold 17 are composed of seven exons, whereas Dpul-Hb10 and Dpul-Hb11consist of six exons, where the second intron is deleted from the ancestral gene structure. Although incomplete, a gene that may have encoded a single domain Hb chain is identified on scaffold 67 (dappu-109652).

An earlier study reported on the partial genomic sequence of a *D. magna* Hb gene cluster containing four Hb genes [S101]. To study the origin and evolution of duplicated Hb genes, and the consequences of their structural arrangements along two distant branches of the *Daphnia*

phylogeny, we further analyze the *D. magna* Hb gene cluster. The nucleotide sequence of the cluster was determined by first screening for clones containing Dmag-Hb1 to Dmag-Hb4 from a lambda Zap genomic library using a DIG-labeled DNA fragment, which was located on the intergenic region between Dmag-hb4 and Dmag-hb5. We then determined the nucleotide sequences of a 6.6 Kb genomic region containing Dmag-Hb2 and Dmag-Hb3 by chromosome walking. Finally, the genomic clone encoding Dmag-Hb1 was screened by using a DIG-labeled DNA fragment that was generated by DNA amplification of the upstream region of Dmag-Hb2. We determined the nucleotide sequence of a 3.6 Kb DNA fragment containing Dmag-Hb1. A total of seven di-domain Hb genes were thus discovered (newly named Dmag-Hb1 to Dmag-Hb8); genes that were previously labeled dhb1 to dhb4 correspond to Dmag-Hb6, Dmag-Hb8, Dmag-Hb5, and Dmag-Hb4, respectively. The seven genes are clustered in the same direction within a length of about 23.5 kb. Other than the obvious absence of Dmag-Hb7 from the *D. magna* cluster, elements in synteny between the two species are seemingly preserved from a duplication history that predates the split between the *Ctenodaphnia* and *Daphnia* subgenera.

The Hb gene cluster is used as a model for analyzing the evolutionary processes associated with tandem gene duplications. Specifically, we hypothesized that TDG clusters, like the Hb gene cluster, are subject to concerted evolution. Three alignments were created for our phylogenetic investigations comparing divergence among protein coding regions and intergenic regions of the Hb clusters. The first alignment of the deduced amino acid sequence of the 18 *Daphnia* Hbs and two nematode genes (*Ascaris suum* and *Pseudoterranova decipiens*) were produced using ClustalW [S102]. Major adjustments were then made according to the conserved amino acids in known functional domains among arthropods and vertebrates (Figure S25). All gaps and amino acids corresponding to gap position were deleted and the amino acid sequences were then converted to the nucleotide sequences. As a result, 882 nucleotides were aligned, of which 534 were variable among the *Daphnia* genes and 749 were variable when outgroup Hbs were included (Figure S26). A second nucleotide sequence alignment by ClustalW was produced for intergenic regions between the stop codon of the upstream gene and the TATA box of the downstream gene, except for upstream sequences of Hb1. All gap positions were removed from the alignment. The 837 nucleotides were aligned, of which 833 were variable (Figure S27). A gene phylogeny for the coding regions was constructed using MrBayes v3.1.2 [S103] by applying the GTR and a site-specific rate model for each codon position. The four Markov Chain Monte Carlo (MCMC) chains were run for 3,000,000 generations and 15,100 trees were sampled with their posterior probabilities. A 50% majority consensus rule tree was estimated. By contrast, a phylogenetic tree for the intergenic regions was constructed by using GTR base substitution model with a gamma rate substitution. The MCMC chains were run for 3,000,000 generations. A 50% majority consensus rule tree was estimated from 15,100 trees.

VI. Evolutionary Diversification of Duplicated Genes

1. Estimating expression-level divergence among paralogs

Identification of duplicate genes – The paralogs used for this study were those identified by *Tandy* (section IV.4) and by our analysis of genome history (section IV.3), which also produced the estimate of sequence divergence at silent sites (K_s) among all pairs of duplicates. Duplicated genes were grouped into gene families by the Markov clustering and MCL clustering methods described above (section IV.1).

Gene expression data – Two datasets were examined for this study, each taken from the multiplex microarray experiments described above (section II.5). The first set of analyses investigated variation of expression among duplicates of individual gene families. The M values ($\log_2$ treatment – $\log_2$ reference) from eight of the twelve experiments were used to calculate

the Pearson product-moment correlation using the statistical package JMP (SAS Institute Inc.). Prior to filtering, correlations were measured for 46,343 pairs of paralogs with $K_s < 5$, of which 35,770 pairs were assigned to 1,393 annotated gene families. Hierarchical clustering of the genes [S104] was based on their M values across experiments and required that a significant expression-level difference was observed for at least one experimental condition. Clustering was performed using the program Cluster v2.11 and visualized using TreeView v1.6 (rana.lbl.gov/EisenSoftware.htm). Plots comparing the correlation coefficients for paired orthologs as a function of their relative ages (measured by K_s) were also produced using JMP (SAS Institute Inc.).

The second set of analyses investigated variation of expression among all duplicated genes within the *D. pulex* genome across all 12 experiments. The microarray probes used for detecting expression differences between paralogs were filtered to only include probes for genes which differed in sequence from the sequence of the closest related paralog by greater that 5% of the nucleotides. This threshold was chosen based on the reported specificity of long oligonucleotides on this NimbleGen microarray platform [S13]. By consequence, of the original 80,142 probes on the array that are designed to query the expression of 29,569 genes, our analysis was restricted to 14,323 probes interrogating 6,241 genes with paralogs in the genome: 3,059 genes are represented by three probes, 1,964 genes are represented by two probes, and 1,218 genes are presented by a single probe. Log_2 signal to background ratios were determined for each probe under each experimental condition, by first calling probes that fluoresced at intensities greater than 99% of the random probes' signal intensities; therefore only 1% of fluorescing experimental probes should be false positives. Probes with negative ratios were discarded from measurements of differential expression for each of the 12 contrasting conditions. The data file is available at [S55].

Our approach followed the general statistical method of Gu et al [S105], who defined a pair of duplicated genes as having "similar" or "different" expression patterns across experimental conditions based on whether their expression scores differed at $p \leq 0.05$ using an analysis of variance. Using custom scripts written for the R statistical package [S37, S55], we employed a similar ANOVA model where all the replicate probes for the two genes formed the error term, and the mean difference of the two genes was the measured effect.

In brief, we define a distinguishable expression pattern by a significance criterion ($p < 0.05$) using ANOVA for the simple statistical model of "aov(Yab ~ Xab)", for matrices Yab differential expression M values and Xab gene factors, with replicates. A supplemental file [S55] reports these ANOVA results (effect(M), se(M), pr(M), and df(2)), for each paralog gene-pair, along with their K_a and K_s values. We use pr(M) < 0.05 as criterion that expression differs between paralogous genes, for zero to twelve treatments. The tested hypothesis is one investigating the number of paralogous pairs in each K_s category that reach the criterion of a distinguishable expression pattern, which is tested for significance with Fisher's exact test for count data presented in Table S42. This method is reliable for as few as two probes for one gene and one probe for the other, although a greater number of replicate probes produced more significant results. The relation between the maximum observed difference in the expression response of paralogs to a shared experimental condition and their number of synonymous substitutions per synonymous site (K_S) was measured by a linear regression model using the R package [S37]. Because large K_s values are unreliable estimates of age, we restricted our analysis to $K_s < 3$.

2. Testing for genome structure effects on expression divergence

To test for genome structure effects on the evolution of gene expression, we compared the expression patterns of duplicated genes that are (1) arranged within TDG clusters, (2) that have signatures of gene conversion (section V.1), and (3) that are dispersed in the genome. The observed numbers of paralogs within each class that shared the same expression patterns, or that had different expression patterns in at least one of the 12 conditions tested on the microarrays were tested against expectations that there are no differences using Chi-square tests.

VII. Functional Significance of Expanded Gene Families

1. Charting metabolic pathways for co-expanding, interacting genes

Homologous genes are defined by the metazoan Non-supervised Orthologous Groups (meNOGs), which are obtained from the eggNOG database [S106]. The meNOGs are built upon 363,805 proteins from the following 18 metazoan species: *Homo sapiens, Pan troglodytes, Macaca mulatta, Mus musculus, Rattus norvegicus, Canis familiaris, Bos taurus, Monodelphis domestica, Gallus gallus, Xenopus tropicalis, Tetraodon nigroviridis, Takifugu rubripes, Danio rerio, Ciona intestinalis, Anopheles gambiae, Drosophila melanogaster, Apis mellifera, Caenorhabditis elegans.* The meNOGs assemble 241,305 proteins into 23,033 orthologous groups. These groups are then subdivided into 4,404 subgroups of genes having a 1-to-many relationship (i.e., gene duplications occurred within a single species), 3,721 subgroups of many-to-many gene relationships (i.e., gene duplications occurred in multiple species) and 14,908 subgroups of genes with 1-to-1 relationship (i.e., single genes are found in each genome). The initial meNOG dataset was extended by the addition of *D. pulex.* The 30,907 *D. pulex* proteins were aligned to the 363,805 meNOG proteins using the PARALIGN software [S107] and the Swiss-Waterman algorithm. *Daphnia pulex* proteins were assigned to the meNOGs by reciprocal best matches above a sequence similarity threshold of 180 bit scores. Thus, 13,816 *Daphnia* proteins were assigned to 7,413 meNOGs. Ortholog groups were previously annotated with enzyme (EC numbers) and to metabolic pathways on the basis of the KEGG database [S44]. Therefore, enzyme annotations were transferred to 1,908 *Daphnia* genes. The data file is available at [S55]

Expanded and contracted enzymes were identified by the Fisher exact test. The test was based on the distribution of the number of genes corresponding to enzymes among the subset of equally distributed species between vertebrates (*H. sapiens*, *M. musculus*, *G. gallus* and *T. nigroviridis*) and arthropods (*D. melanogaster, A. mellifera, A. gambiae*). For example, we identified 89 copies of the *Daphnia* gene encoding enzyme EC2.4.1.152 (fucosyl transferase). By contrast, the total number of genes encoding this enzyme in other species is 13 (2 in *H. sapiens* , 1 in *M. musculus*, 1 in *G. gallus*, 3 in *T. nigroviridis*, 2 in *D. melanogaster,* 2 in *A. mellifera* and 2 in *A. gambiae*). The Fisher exact test statistic and the corresponding p-value were calculated based on expectations derived from comparing 1,908 total genes in *Daphnia* to 7,876 genes in all other species. Finally, a Bonferroni correction was applied to account for multiple testing using 563 as the total number of unique *Daphnia* enzymes. We tested for expanded and contracted enzymes comparing arthropods *vs* vertebrates (see Figure S30 and Table S44 for detailed information) and *Daphnia vs* all the other genomes (see Figure S31 and Table S43 for detailed information). Thirty-eight enzymes showing significant deviations from expected numbers (p-value < 0.05) were finally mapped onto the overview metabolic network [S108] to observe functional relationships (Figure 4).

Among these 38 enzymes encoded by amplified genes, we identified the fraction of interacting genes (i.e., sharing metabolites) within the whole metabolic network (in total 563 enzymes and 478 interactions). As a result, 19/38 enzymes interact within small subnetworks (Figure 4 panels A-G). To assess the significance of the observed number of interacting amplified

genes, we first applied a binominal test. The probability distribution required for the binominal test was generated from 1,000 sets of randomly selected 38 genes. As a result, we proved that the nineteen (half of 38) is a significantly greater number of genes than numbers that are observed by chance. Second, we additionally performed a network permutation analysis. That is, we generated 1,000 randomized whole metabolic networks using node permutation (i.e., relabeling all nodes), and checked the number of interactions among the same set of 38 amplified genes. As a result, the number of interacting genes within the amplified genes in the "real" network is significantly higher than that in randomized networks ($p < 0.03$ in the null distribution derived from the 1,000 randomized networks (Figure S32).

2. Uncovering functional diversity of glycosphingolipid biosynthesis genes

To test whether evolutionary preservations of duplicated genes may be functionally interdependent, we compared the average similarity of expression patterns for interacting genes from lineage-specific expanded families within shared metabolic pathways (i x i matrix) to that for non-interacting genes from families of different pathways (i x j matrix). The differential gene expression patterns (only $\log_2$ fold change > 0.5 were considered) of 275 duplicated genes across 12 experimental conditions and belonging to 38 metabolic pathways (Table S43) were used to calculate 37,675 pair-wise estimates of expression similarity based on their root mean square difference (RMSD). Thus, a RMSD near 0 is indicative of genes that are alike in their expression patterns, whereas a RMSD = 1+ is indicative of genes whose expressed patterns are different. There are mostly only single comparisons of interacting gene families (i x i) within the same pathway (same KEGG map ID in Table S43), but many possible pairs of gene families to chose for different-pathway comparisons (i x j). To reduce chance bias of selecting high scoring pairs from this large null-hypothesis i x j matrix, different-path comparisons were limited to gene families having a similar number of paralogs.

The hypothesis being tested is whether a greater similarity in expression is observed for best-matched genes belonging to two different families within the same pathway, than observed for best-matched genes belonging to two families from a different pathway. We therefore implemented sampling without replacement for each enzyme (gene family) pairing, calculating RMSD for all possible gene pairs, then selecting the most alike pairs until all genes from the smallest enzyme group are matched. Thus, duplicated genes are sampled only once from each enzyme group. Significant differences between the averages calculated for all i x i and i x j gene pairs were tested using the t-statistic. The input files, custom perl program "dpx-msrevpathq.pl" and results files for this "PathXDiverge" analysis for *D. pulex* gene expression patterns across metabolic pathways are available at [S55].

To further test our hypothesis and provide a specific example, we contrasted the phylogenetic history of interacting and co-expanded gene families of the glycosphingolipid biosynthesis pathway of metabolism to their similarity in expression patterns across eight microarray experiments. Amino acid sequence alignments were obtained using MUSCLE [S58] for 96 genes from among three families (Tables S45-48). Phylogenetic gene trees were constructed by the maximum likelihood method using the PHYLIP ProML algorithm [S109] with corrected distances by the Jones-Taylor-Thornton model of molecular evolution [S110]. Correlation coefficient plots and hierarchical clustering of genes, based on their differential expression patterns, were conducted as described above (section VI.1). Of particular interest was the functional association of genes within the largest expanded metabolic gene family (fucosyltransferase; enzyme 2.4.1.152) and the nine members of the expanded glycosyltransferase gene family (enzyme 2.4.1.65), because both enzymes are required to catalyze biochemical reactions for the production of branched glycans along the glycosphingolipid biosynthesis pathway [S111]. To test

these associations, we partitioned the variance in differential gene expression (DE) from microarray experiments with a nested ANOVA and REML estimator using JMP 8.0 (SAS Institute Inc.). We used the estimated variance component to calculate the ratio of among group variation to total variation. This ratio is the statistic D_{st} that estimates group differentiation based on the quantitative differential expression data and varies from 0 to 1, similar to F_{st} [S112]. Unlike F_{st}, D_{st} is a measure of phenotypic, not genetic variance. The test was based on calculating the variance in the expression patterns of duplicated genes sharing memberships within (1) phylogenetically distinct clades (>95% identity at amino acids) relative to the variance in expression patterns observed among genes having independently evolved, and (2) groups of genes clustered with unrelated interacting genes based on the hierarchical clustering. We used the Delta method [S113] to estimate the significance of D_{st}.

VIII. Ecoresponsive Genes

1. Treatment of the transcriptome data with reference to the annotation

Sequences obtained from the cDNA sequencing project (section II.2) were classified as transcribed genes under biotic ecological conditions, abiotic ecological conditions, and standard non-ecological conditions, based on the libraries from which the gene transcripts were sampled. The biotic ecological challenges include exposure to bacterial infection, predators, hormones and varying diets (Table S10; TRO 12-20, TCO 9, 14). The abiotic ecological challenges include animals exposed to environmental toxicants, elevated UV, hypoxia, acid, salinity and calcium starvation (TRO 1-4, 6-9, TCO 4-8, 10-13, 15). Standard non-ecological conditions include animals at various stages of life history within a controlled laboratory environment (TRO 5, 10-11, 21, TCO 1-3). The transcribed gene counts with and without homology to proteins from other species were tabulated and tested against expectations that these were equally distributed among the three classes of ecological conditions using Chi-square tests. Chi-square tests were also performed for transcribed genes from the tree classes found within and outside of tandem duplicated gene (TDG) clusters.

Differentially expressed Transcriptional Active Regions (TARs) obtained from the whole genome tiling path microarray studies (section II.4) were classified as overlapping with annotated exons (gene), residing within predicted introns of annotated gene models (intron), or located outside of currently annotated gene models (unknown). For each of the four tested treatments, counts of the tiles with up-regulation, down-regulation and no differential expression in each genome feature were tabulated. Chi-square tests were conducted against the null expectation that the pattern of regulation of tiles in each genome feature would be proportional to the number of tiles in each feature within each category of regulation (up-, down-, and no differential).

SUPPORTING TEXT

1. Chromosome Studies

The chromosomes of *Daphnia* are extremely small. Past karyological observations have therefore been restricted to counting the diploid chromosome numbers [S114, S115]. Recent advancements in cytological techniques and instrumentation have permitted some successes at characterizing the morphology of *D. pulex* chromosomes (Figure S6).

Because most chromosomes are uniformly short, they are only roughly arranged according to size. Yet three size classes are apparent (Table S8). Chromosome 1 is obviously the largest, measuring 5.6-6.6 µm or 25% of the total. Chromosomes 2-4 form the second class, containing 30% of the total nuclear DNA, while chromosomes 5-12 constitute the third and shortest class for the remaining 45%. Heterochromatic (A-T rich) regions are observed only on the four largest chromosomes. Two internal regions are identified in chromosome 1 and both terminal regions of chromosome 2 are banded; single broad bands are observed on chromosomes 3 and 4.

A first genetic linkage map for *D. pulex* was already published using 185 microsatellite markers [S8]. This investigation measured the segregation of polymorphisms within 129 (F_2) selfed progeny from a *D. pulex* hybrid (F_1) obtained by crossing two genetically divergent isolates from populations in Oregon. The map spans 1,206 Kosambi cM and shows an average inter-marker distance of 7 cM. Linkage groups range in size from 7 to 185 cM and the number of markers per linkage group varied from 4 to 27. The map reveals linkage groups corresponding to the 12 chromosomes and covers approximately 82% of the genome.

We consolidated the genetic map data with the genome scaffolds to assign these sequences to each of the 12 chromosomes for the purpose of validating the genome assembly, identifying gaps and to begin defining the recombinational landscape. Mapped microsatellite marker sequences were unambiguously identified on the genome scaffolds by sequence similarity searches (Table S5). Of the 5,191 scaffolds from the present assembly, only 73 are placed onto chromosomes. Work is underway to obtain better coverage and consolidation of the *D. pulex* genetic and physical maps, while substantial progress is made at discovering the recombination map of the 10 *D. magna* chromosomes [S116].

Telomeres in Arthropoda are so far known to range from simple TTAGG telomeric repeats – with relatively uniform and short ~3 kb sub-telomeric regions for the long arm telomeres in the honey bee *Apis mellifera* [S117] – to much longer arrangements including multiple retrotransposon insertions within the TTAGG repeats in the silkmoth *Bombyx mori* and flour beetle *Tribolium castaneum* [S62, S118], to the unusual situation in Diptera, which have lost both telomerase and TTAGG repeats and depend entirely on regular insertions of particular retrotransposons (e.g. [S119]). We identified and manually annotated a single full-length ortholog of insect telomerase [S117, S120] in the *D. pulex* genome [NCBI Acc. Num for DpulTERT].

We searched the 228,190 fosmid clone end reads for tandem repeats of TTAGG with lengths of 1,000 bp. We found several hundred matches, most with long stretches of TTAGG repeats, although sometimes interspersed with TTAGGG repeats, which is the ancestral arthropod repeat. Almost all of these are plus/minus orientation, indicating that ends of chromosomes in *D. pulex* indeed consist of long stretches of TTAGG repeats (otherwise we would expect equal numbers of plus/plus and plus/minus matches). Examination of the mate pairs of these fosmid end-reads,

which should therefore be 30-40 kb internal to the TTAGG repeats, revealed almost entirely repetitive sequences.

One particular 136 bp satellite repeat was very common amongst these mate pairs and appears to form long repeat stretches that immediately border the TTAGG repeats, so was named TELSAT1 (consensus sequence is TTTTTCTAAGTATTGTCATCAGCGCCACCTGGTGGCAAGTTTTGGAACTAAATTTTATTATGATCGCATCGTGTTCAGCGTTAAATTCTGATCAAGAATATGTTTGTTTCAAATGGTTCTGAGCAGTAGAAGTGCC). Examination and alignment of all 86 junctions between TELSAT1 and TTAGG repeats within the full set of sequence reads revealed that TELSAT1 repeats only occur in front of TTAGG repeats in direct tandem orientation, although rarely they are interspersed within the TTAGG repeats. There are 28 unique junctions of TELSAT1 repeats with TTAGG repeats, all joined from different positions within the TESAT1 repeat to the GG of a TTAGG repeat. A few of these junction sequences are singletons that might be interspersed within the TTAGG repeats, leaving around 24 unique junctions with multiple reads representing them, which likely are the 24 telomeres on the 12 *D. pulex* chromosomes.

To identify unique sequences upstream of the TELSAT1 repeats, a second search of the fosmid end reads was conducted with multiple TELSAT1 repeats, and the mate pairs of plus/minus matches were examined. Most of these sequences are composed of more TELSAT1 repeats, indicating that these repeats form sub-telomeric clusters over 40 kb in length. The few others include another 193 bp satellite named TELSAT2 (consensus sequence is TTTCCCTGTTACAGGATATGTTCATCGATGTCCAATACACTATTTAAAGTCATTAAAATCAATGAATCTATTAAGACATTCATGATGGAAAAGAAATAGAAATAAGAGTTGATAGAAAATCTTCCAGGAACTGAAAATCACAACTTCAATGAATTTAAAATGACGATTCTGATTGTTTTACAAATTTCAAGGG). Efforts to progress beyond these TELSAT2 repeats led only to multiple other repetitive regions, thwarting efforts to connect these sub-telomeric regions to unique scaffolds in the assembly. In summary, the *D. pulex* telomeres appear to consist of terminal TTAGG repeats of a few kb, with long stretches of TELSAT1, TELSAT2, and other repeats in the sub-telomeric regions.

2. Gene Homology among *Daphnia* Genomes

TCO genes were partitioned among four classes of models, based on supporting evidence. Searches for homologs were conducted by measuring nucleotide similarities using BLASTn [S121] between TCO and TRO genomes. We estimated the levels of sequence divergence between these two strains range between 3% and 5%. The first class of models consisted of TCO v1.1 gene predictions with both homology to non-daphniid proteomes and EST evidence. We found 17,411/18,233 (95.5%) genes models with significant alignments ($e < 10^{-5}$) to TRO sequence. The second class of TCO gene predictions consisted of models without homology to other sequenced proteomes, yet having EST or paralogs (i.e., lineage-specific genes). We found that 9,733/12,707 (77%) of these gene models had significant sequence alignments to TRO sequences. The third class consisted of TCO *ab initio* gene predictions that were not included in the Frozen Gene Set v1.1 because they lacked supporting evidence. Here, 6,576/10,015 (65.7%) had clear homologs in the TRO genome. Finally, the fourth class consisted of extra gene predictions inferred from transcriptional active regions (TARs) where tiling array data suggested significant expression levels in areas without ESTs or gene prediction models (Table S12). Based on BLASTn scores, 6,684/7,897 (84.6%) TARs had homology between TCO and TRO.

We also searched for homologs of *D. pulex* genes within the *D. magna* genome that is currently being sequenced. *Daphnia magna* is a member of the subgenus *Ctenodaphnia* and resides primarily in Eurasia, whereas *D. pulex* is mainly in North America and its lineage split

from the *D. pulex* ancestor ca 150-200 MYA [S122], although younger estimates are obtained from nuclear genes [S123]. We currently have a draft genome assembly from 8 x coverage sequencing using the Roche-454 genome sequencer. Due to the possibly deep evolutionary history between these species, we used tBLASTn to detect homology between the two genomes (cut-off set at $e < 10^{-5}$). Using the same four categories as above, we found evidence of homology for 1) 16,486/18,233 (90.4%) "best" predictions, 2) 4,969/12,707 (39.1%) of lineage-specific genes, 3) 2,319/10,015 (23.1%) of weak-evidence predictions and 4) 2,787/7,897 (35.3%) of TARs.

3. Micro-RNA and Transposable Elements

We located 50 micro-RNA (miRNA) loci in the *D. pulex* genome (Table S16) using a pipeline that uses Support Vector Machine models, homology and an orthology procedure [S48]. All loci are preserved in insects, most are single copy genes except for three loci: dpul-mir-2, dpul-mir-7, dpul-mir-87.

MicroRNAs are short (21 – 24-nt) non-coding RNAs that bind to complementary sites, usually located in the 3'-UTR of target mRNAs, and regulate protein translation. We discovered three miRNA-producing loci are evolutionary conserved within sequenced insect and *Daphnia* Hox clusters. Locus dpul-iab-4 resides in the Bithorax complex between the Abd-B and Abd-A genes, while dpul-mir-993 and dpul-mir-10 reside in the Antennapedia complex between Pb and Dfd, and between Dfd and Scr genes, respectively. Recent reports demonstrated that the iab-4 gene produces two distinct miRNAs that are encoded on opposite DNA strands [S124]. They inhibit endogenous UBX expression to induce Ubx-like haltere-to-wing transformations [S125, S126, S127]. Surprisingly, the structural arrangements important for wing development are preserved in the *D. pulex* genome (Figure S10). Knowledge on the general functional conservation of miRNA is restricted by the limited diversity of available arthropod genomes. For example, Shiga et al. [S128] reported several alternatively spliced variants of *D. magna* Antp and Ubx mRNAs, including bi-cistronic transcripts of both genes, yet no protein expression was observed from the fused Ubx/Antp transcripts. Ubx mRNA was shown to be a direct target for iab-4 microRNAs in *Drosophila melanogaster* [S124], implying that regulation of protein expression from fused transcripts might be mediated by functions of microRNAs in *Daphnia*.

In annotating transposable elements, 1,712 intact or fragmented elements are identified from five superfamilies of non-LTR retrotransposons, including the L2 superfamily, which is abundant in *D. pulex* but otherwise found only in the *Anopheles gambiae* genome. Representatives of 10 superfamilies of DNA transposons, including the *Helitron* and *Maverick* subclasses, are also found in *D. pulex*. Many have full-length open reading frames indicating they may have been recently active. Finally, as expected, the *Daphnia* specific DNA transposon *Pokey* [S129] is inserted in multiple copies throughout the large subunit ribosomal RNA gene of the single ribosomal DNA (rDNA) array, in addition to occurring at other genomic locations. The distribution of *Pokey* in the rDNA array is visualized using fiber-FISH (Figure S11) because sequence assemblies of the tandemly arrayed rDNA units are not possible.

4. The 46 *Daphnia pulex* Opsins

Animals use proteins of the opsin family of seven-transmembrane G-protein-coupled receptors to detect light (e.g. [S130, S131]). Three major lineages or subfamilies of opsins in animals are generally recognized: the ciliary opsins represented most prominently by the vertebrate visual opsins, the rhabdomeric opsins represented by the insect visual opsins, and the retinochrome- or Go-like opsins represented by RGRopsin, peropsin, and neuropsin in chordates (e.g. [S97,

S132, S133, S134]). Some opsin evolution experts split the latter group into multiple subfamilies in recognition of their considerable divergences (e.g. [S131]). The classification of this third subfamily remains unsettled, and some authors rank these as protein families within an opsin superfamily. Nevertheless, substantial evidence suggests that all three subfamilies predate the major split of bilateral animals into the protostomes and deuterostomes: (i) the chordate melanopsin are relatives to the previously protostome-only rhabdomeric opsins [S132, S135]; (ii) the insect pteropsin and an annelid ciliary opsin are protostome representatives of the ciliary opsins [S133, S134, S136]; (iii) vertebrate members of the retinochrome-like subfamily resemble squid retinochrome (e.g. [S137, S138, S139]), as does the opsin 2 gene in scallops [S140]. More recently, older animal phyla are revealing additional opsin lineages and evolutionary complexity, including a clade named 'cnidops' known only from Cnidarians (e.g. [S97, S141]).

The *Daphnia* compound eye consists of eleven ommatidia and the two eyes are fused during ontogeny into a single anterior and dorsal organ. *Daphnia* also have a single ocellus. Like some other arthropods (reviewed in [S142]), each *Daphnia* ommatidium of the compound eye has eight photoreceptor cells (see [S143]). Attempts to study the wavelength specificity and sensitivity of these individual light-detecting units proved difficult. But in pioneering work, [S144] used intracellular recordings to identify photoreceptor cells that respond specifically to blue, green, and red light. Smith and Macagno [S143] confirmed these capabilities using extracellular recordings from entire ommatidia and also demonstrated UV sensitivity.

By manual annotation of the *D. pulex* genome sequence, we identified 46 opsin genes (Table S32). *Daphnia pulex* has the greatest number of opsins of which we are aware described to date for any animal (Figure S21), although the genomes of the cnidarians *Hydra magnipapillata* and *Nematostella vectensis* rival *D. pulex* if counting the numerous cndiarian sequences that are presumably pseudogenes [S97]. Our phylogenetic analysis along with genes from the three known subfamilies of animal opsins revealed that most *D. pulex* opsins originated by gene duplications among four lineages, including a novel rhabdomeric opsin lineage we name arthropsins. Arthropsins are highly diverged from other known opsins. Their phylogenetic position, coupled with absence from all other available animal genome sequences implies multiple independent losses of this kind of opsin, whose functions are unknown. This large repertoire of opsins, along with previous studies revealing multiple photoreceptors and opsins in other crustaceans (e.g. [S95, S145, S146]), indicates that a remarkable diversity of opsins mediates light-sensitive behavior in these arthropods.

Expansion 1, Arthropsins – We were surprised to discover an entirely new and putatively ancient lineage of opsins in the *D. pulex* genome, which we call arthropsins. Arthropsins form a sister group to all known members of the rhabdomeric clade, confidently outside even the vertebrate 'melanopsin' rhabdomeric lineage.

Because of the unexpected position of arthropsins, we looked for evidence of rapid rates of evolution because fast evolution could cause positively misleading topological results [S147]. We performed all possible three-taxon, maximum likelihood relative rate tests between arthropsin genes and all other genes, using a ciliary opsin outgroup (Takifugu TMT, GenBank AAM90677). These relative rate tests were implemented in HyPhy [S148], assuming a WAG + F model of protein evolution and a critical value using Bonferroni correction for multiple comparisons. There was no evidence of elevated rates of molecular evolution in arthropsin genes based on ML relative rate tests, which does not support long-branch artifacts determining clade position: 970 out of 988 arthropsin comparisons were non-significant; 18 comparisons significantly rejected the null hypothesis of equal rates of evolution between an arthropsin gene and another gene; 16

of these comparisons involved Amphiop4 or Amphiop5, showing that Amphiop4 and 5 genes evolved significantly slower than arthropsin genes. These results do not indicate arthropsins genes are fast evolving. Rather, Amphiop4 and 5 genes are slow, as indicated by significantly slower rates in 213 of 254 relative rate comparisons involved Amphiop4 or 5. Two comparisons showed that arthropsin genes evolved significantly slower than Squid Retinochrome. Taken together, there is no evidence of rapid evolution in the arthropsin genes, and no reason to suspect LBA in the placement of the clade. Other possible explanations for this placement, including convergent evolution of rhabdomeric-clade synapomorphies remain to be explored. Indeed, arthropsins share several diagnostic amino acids with the rhabdomeric opsins, including the SHP (or SSP) motif at the terminus of TM7, which contrasts to the XNX motif shared by all ciliary, peropsin and RGR opsins. Moreover, the cytoplasmic loop 3 (CL3) domain of arthropsins is longer than that of the ciliary opsins, in keeping with all other rhabdomerics. The sequence of this loop is divergent from the other rhabdomeric opsins, however, whereas it is highly conserved within all the arthropod visual opsins (e.g. [S146]).

Besides being ancient, arthropsins have undergone their own expansion within the *D. pulex* genome, including two presumably old lineages (based on their low ~50% amino acid identity), each with multiple sub-lineages. In the absence of functional information, the only obvious features that distinguish the arthropsins from the other rhabdomeric lineages is that they all have relatively long C-termini, comparable in length to the pteropsins and some other ciliary opsins. Arthropsins1-5 also have a few additional amino acids in CL3, making this loop longer than those of other known opsins. We name these genes arthropsins to indicate their presence in at least one major arthropod lineage. We hypothesize that others will be discovered in other crustaceans, perhaps some insect lineages, as well as other arthropods, or other protostomes.

Expansion 2, Pteropsins – Pteropsin is a protostome lineage of ciliary opsins, which are otherwise primarily known from vertebrates [S133, S134, S136]. Arendt et al. [S133] defined the ciliary and rhabdomeric lineages based in part on their recognition of both kinds of opsins in the annelid *Platynereis dumerilii*, which is a protostome. In both insects and annelids, this ciliary opsin is expressed in the brain rather than in visual organs, and hence is likely to serve a non-visual role in light detection, perhaps in entraining circadian rhythms [S134, S149]. Although duplication of pteropsin is known from *Anopheles gambiae* mosquitoes (AgOp11 and 12, [S136]), *D. pulex* again reveals multiple, sometimes old (based on as low as 54% amino acid identity), duplications of this lineage. Among these nine duplicated genes we discovered the only obvious pseudogenes among the total set of 46 opsin genes; specifically, Pteropsin2 has multiple frameshifts and a mutated intron/exon boundary, while Pteropsin5 has a small frameshifting deletion in exon 7. The *D. pulex* pteropsin genes share all five introns that insect pteropsin genes share, including the three that group them with the vertebrate ciliary opsins, as well as two idiosyncratic introns not seen in any other opsin gene (data not shown). The expansion of the *D. pulex* pteropsins also led to some proteins with unusual features. These include insertions of 5-15 amino acids in CL2, which includes a string of 5 or 6 glycines in Pteropsin5-8. Similarly an insertion of 4-25 amino acids is present in EL3 in Pteropsin4-9.

Expansion 3 – Short wavelength and unknown wavelength opsins – Daphnia pulex have four opsins that fall within a paraphyletic grade at the base of rhabdomeric opsins. This grade also includes opsins from other arthropods with experimentally determined short wavelength sensitivities, including Drosophila UV (rh3) and blue (rh5) opsins. Two of the *Daphnia* opsins are similar to UV and blue opsin clades, respectively. Most insects have single orthologs of the blue and UV opsins, therefore, these findings are unremarkable [S143, S144]. In addition, Kashiyama et al [S94] found *Triops* and *Branchinella* to have single orthologs sister to known UV opsins (they did not detect the blue ortholog we report in *Daphnia*). The other two *D. pulex* opsins in

this grade are homologous to the Rh7 opsin in *D. melanogaster* (also called “the unknown wavelength opsin”). The *Daphnia* genes share only 49% amino acid identity.

Expansion 4 – Medium- and long-wavelength opsins – Daphnia pulex have numerous opsins from two major lineages of presumably medium and long-wavelength opsins. Lineage A is already known from a crab [S150] and *Triops* [S94]. The crab opsins are maximally sensitive to green light around 480 nm [S150]. Lineage B is composed of only *D. pulex* genes and other branchiopod genes. The two lineages cluster confidently with the long-wavelength opsins of insects and of other arthropods. However, the divergence of *D. pulex* genes from the other crustacean opsins is curious, because the better known long-wavelength lineage in insects has clear orthologs in crustaceans [S146] including *Procambarus clarkii* [S151] and in a chelicerate *Limulus polyphemus* [S152]. Presumably, genes from this better-known long-wavelength opsin lineage were lost during evolution leading to branchiopods. In turn, the *D. pulex* opsins in lineages A and B are sufficiently ancient to also be present in other branchiopods.

We speculate that these two opsin lineages underlie the green and red wavelength photoreceptor cell sensitivities identified by [S144] and Smith and Macagno [S143], with the Lineage A genes mediating green sensitivity and the Lineage B genes mediating red sensitivity. Furthermore, the expansion of these two lineages to total 25 genes is unprecedented in animal genomes, although expansions to six genes have been reported for the long-wavelength opsins of *A. gambiae* mosquitoes [S136] and Oakley and Huber [S95] reported up to eight opsins in two ostracods. Unlike the insect expansions, which are all relatively recent and apparently largely species-specific, these two crustacean long-wavelength lineages are very old (less than 50% amino acid identity for all A-B comparisons) and each have diversified in both ancient and recent times; multiple young duplications encode almost identical proteins.

The remarkable repertoire of opsins encoded by the *D. pulex* genome indicates that their visual capabilities, while long recognized as being sophisticated, might be even more so. Early work demonstrated sensitivity to at least four different wavelengths, corresponding to UV, blue, green and red light [S143, S144]. In his intracellular recordings from single photoreceptor cells within ommatidia, Schehr [S144] observed that R6 and R8 have peak sensitivity around 450 nm, R2, R3, and R5 are most sensitive around 510 nm, and R1 around 590, so the R4 and R6 cells are candidates for the UV receptor cells. Smith and Macagno [S143] noted that the long wavelength specificities were less easily defined when observed extracellularly for entire ommatidia. It is therefore possible that each cell expresses a different opsin, or sometimes even multiple opsins. In addition, Smith and Macagno [S143] noted that spectral sensitivities showed slight variations between dorsal and ventral ommatidia. It is therefore also possible that the particular opsin expressed in a particular photoreceptor cell is different in different ommatidia. Detailed *in situ* hybridization studies of the expression patterns of these opsins, particularly the many LOPA/B genes, will help address these questions.

SUPPLEMENTARY FIGURES

A. Introduction

Figure S1. Reconstruction of the evolutionary history of sequenced arthropods by maximium likelihood methods. Branch lengths are actual sequence divergence corrected for multiple substitutions at 131 aligned and concatenated universal single-copy orthologs totaling 23,748 amino acids. All nodes of the phylogeny are supported by the bootstrap value of 100%. The *Daphnia* lineage is firmly positioned at the base of the insect clade, together forming the Pancrustacea, confirming current knowledge of the phylogeny and showing that the overall molecular evolutionary rate in the *Daphnia* lineage is not extraordinary. **Common names:** *Acyrthosiphon pisum*, pea aphid; *Pediculus humanus*, human louse; *Apis mellifera*, honey bee; *Nasonia vitripennis*, jewel wasp; *Aedes aegypti*, yellow fever mosquito; *Culex quinquefasciatus*, southern house mosquito; *Anopheles gambiae*, African malaria mosquito; *Drosophila melanogaster*, fruit fly; *Tribolium castaneum*, flour beetle; *Daphnia pulex*, waterflea, *Ixodes scapularis*, blacklegged tick.

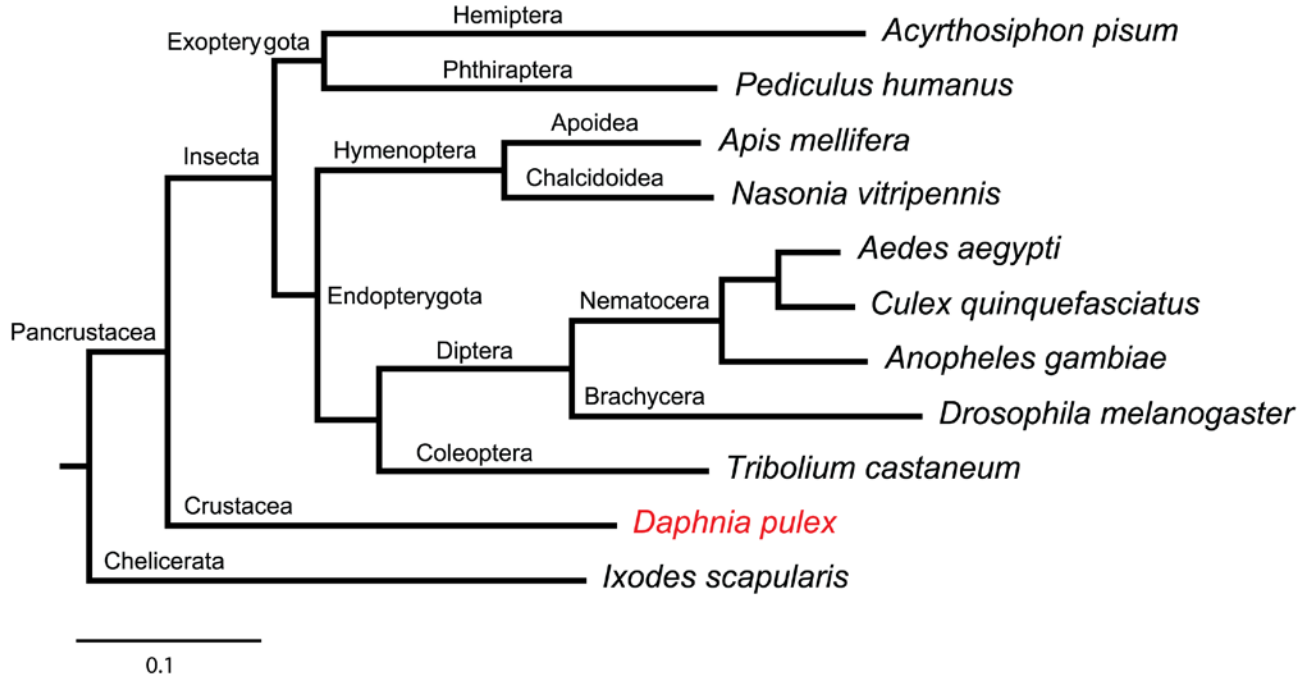
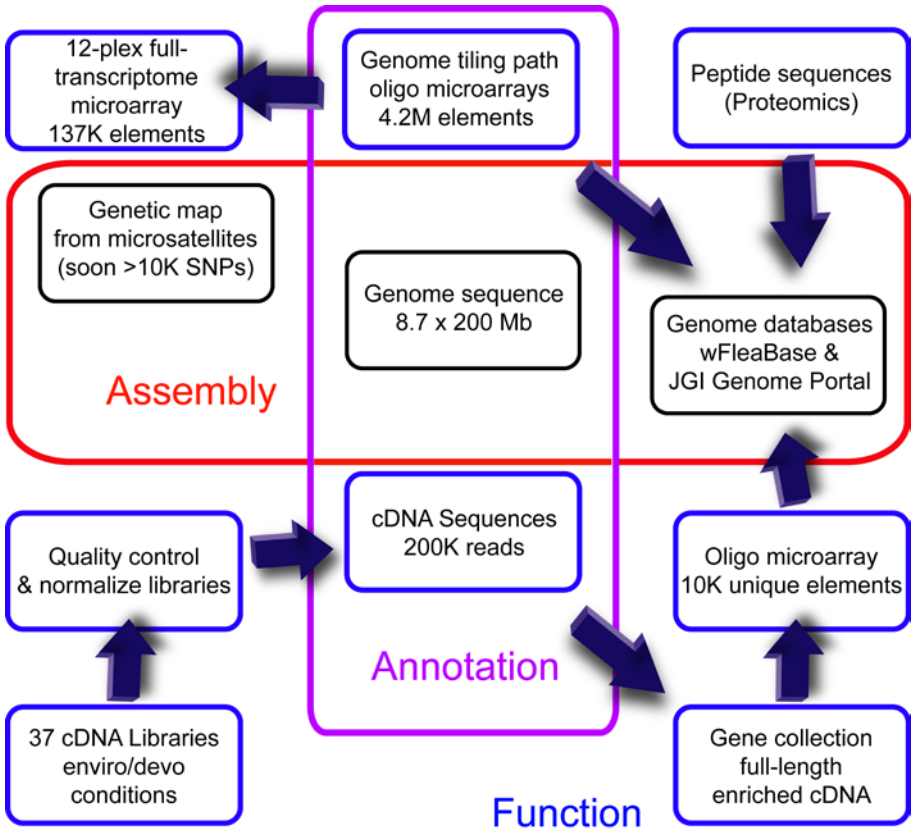


Figure S2. Overview of the *Daphnia pulex* Genome Project. This multi-institutional project is divided into three sections. **(1)** Sequencing and assembly was done at the Joint Genome Institute (JGI) using DNA prepared at the University of New Hampshire (UNH) and at Indiana University (IU). Two additional assemblies and a genetic map were used to validate the results at IU. **(2)** Automated gene calls were made using algorithms implemented at the JGI, IU and at the National Center for Biotechnology Information (NCBI). Empirical gene annotations were made possible by sequencing 200,000 ESTs at the JGI from cDNA libraries created at UNH and IU to sample genes expressed under a variety of ecological settings and developmental stages; additional RNA was obtained from consortium members at the University of Wisconsin-Milwaukee, the University of Edinburgh and the University of Basel. Genome tiling path microarray experiments were carried out at Roche NimbleGen Inc. and at IU; additional RNA was obtained from Ludwig-Maximilians-Universität (LMU). **(3)** Discoveries of gene products and functions were made based on functional genomic experiments using in-house spotted oligo and Roche NimbleGen 12-plex microarrays at IU and proteomics at Utrecht University, University of California Davis and LMU; additional RNA for microarray experiments was obtained from Utah State University. All results are integrated at two publicly available databases: wFleaBase at IU [S153] and at the JGI's Genome Portal [S154]. Finally, over 100 investigators of various disciplines received manual annotation training at IU and via telephone and video conferences from JGI and IU, then trained others to ultimately contribute a series of manuscripts describing *Daphnia*'s genome biology [S39]. Arrows indicate the flow of information across the three sections.



B. Genome Sequence, Assembly and Chromosomes

Figure S3. Distributions of the cumulative scaffold and gap lengths for the JAZZ, Arachne, and PCAP assemblies (with 26,848, 23,643 and 61,858 scaffolds, respectively). The JAZZ assembler produced the best results, likely because of differences in its algorithm. For JAZZ, the unhashability threshold was set to five times the estimated sequence depth (i.e., 40). This is the threshold before a given sequence string is deemed too frequent to be used to seed alignments. The mismatch penalty was set to -30.0, which would tend to assemble together sequences that were more than 97% identical. Other scoring and penalty options were set to their default values. Default parameters were used for Arachne and PCAP. PCAP produced the largest number of scaffolds and placed more reads than the other two assemblers. A majority of the scaffolds, however, contain only a single contig. Super-scaffolds are also displayed, based on manual gap-bridging. The inset plots the earliest cumulative rate of assembly for scaffolds 1-1,500.

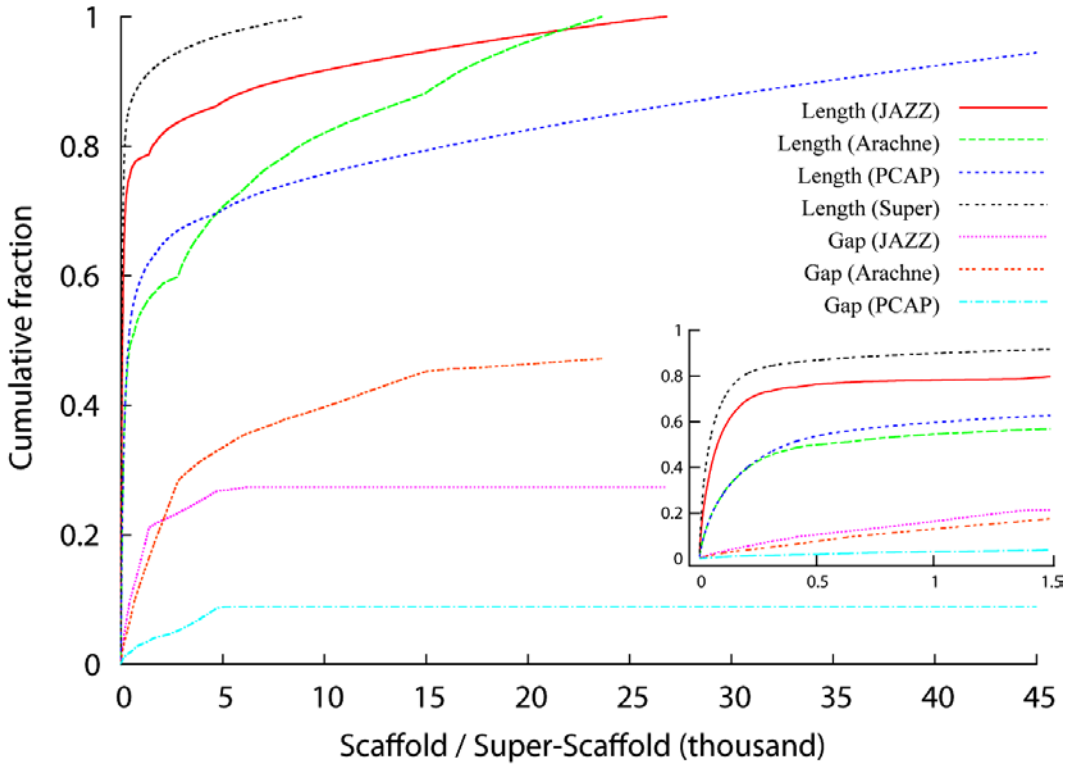


Figure S4. Venn diagram highlighting the number of putative mis-assembled regions by using three different methods: GAV (Genome Assembly Validator), which identified 3,053 putatively mis-assembled regions; comparative validation using Arachne assembly as the reference, which identified 1,889 putatively mis-assembled regions; and the comparative validation using PCAP assembly as the reference, which identified 3,304 putatively mis-assembled regions. Notably, only a small number (84) of regions were reported by all three methods, whereas most of these regions were reported by only one method (2,519 for GAV, 891 for Arachne, and 2,168 for PCAP).

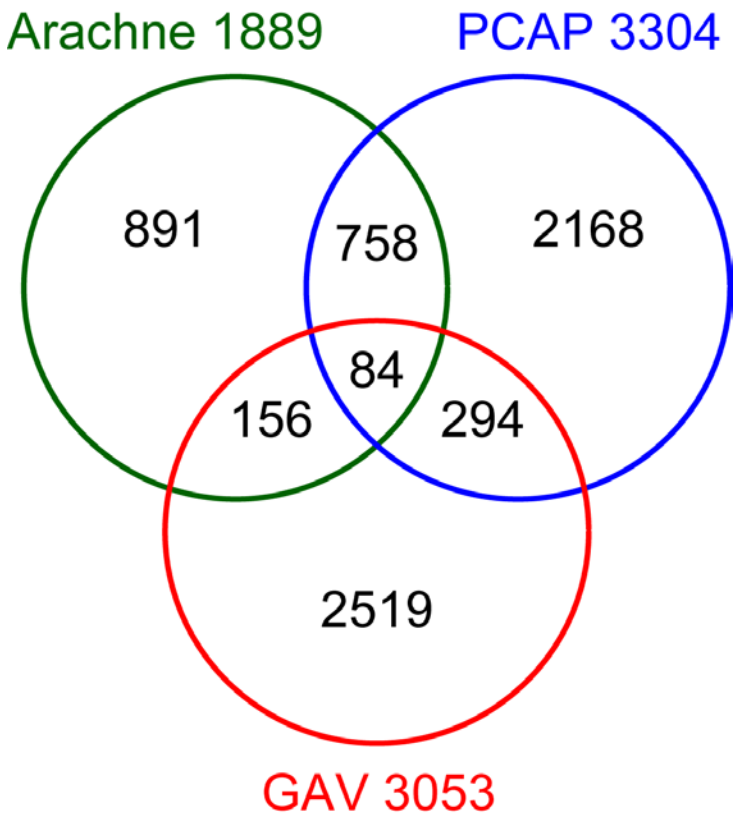


Figure S5. The distribution of detected breakpoints by GAV among the scaffolds with varying lengths. As expected, long scaffolds tend to contain more potentially mis-assembled regions.

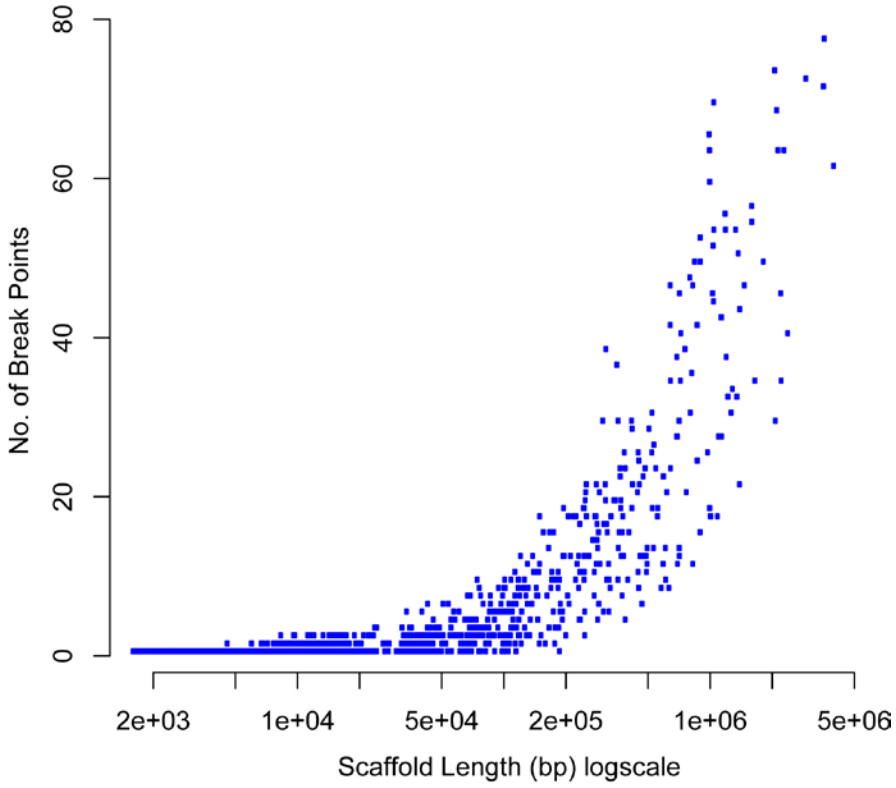
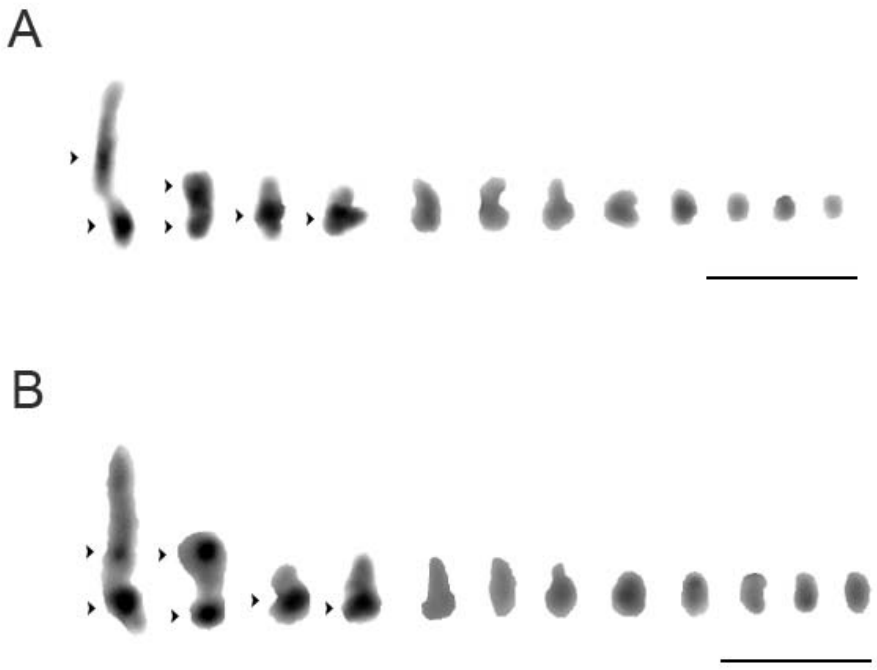


Figure S6. The karyotype of *Daphnia pulex* based on meiotic chromosomes prepared from testis. DAPI banding (A) and G banding methods (B) were used to reveal heterochromatic, dense DNA and/or higher AT content bands. Chromosomes are aligned according to their length. Arrowheads indicate the conspicuous heterochromatic bands on four large chromosomes. Bars represent 5 μm. Chromosome and banding measurements are listed in Table S8; we estimate that chromosomes 1-4 contain half of the genome's DNA and that 25% of the genome is heterochromatic.



C. Largest Gene Inventory

Figure S7. Corroborating evidence for the existence of a minimal set of 30,907 predicted protein coding genes. **(1)** Expression of 10,578 genes was detected by cDNA sequencing (ESTs aligning to >90% of the gene model); **(2)** Expression of 13,445 genes was detected by tiling path microarray experiments as transcriptional active regions (TARs aligning to >80% of the gene model). **(3)** Paralogs were found for 13,105 loci ($p < 10^{-20}$). **(4)** Homologs of 18,765 genes were detected within a draft assembly of 8-fold coverage of the *D. magna* genome sequence. **(5)** Homology was found for 19,641 genes in other sequenced genomes ($p < 10^{-5}$). **(6)** Peptides were sequenced matching 1,273 genes (not shown). At least 26,649 loci (86%) are conservatively supported by at least one line of evidence. Homology and transcriptional evidence for the v1.1 annotated gene set is listed in Table S11.

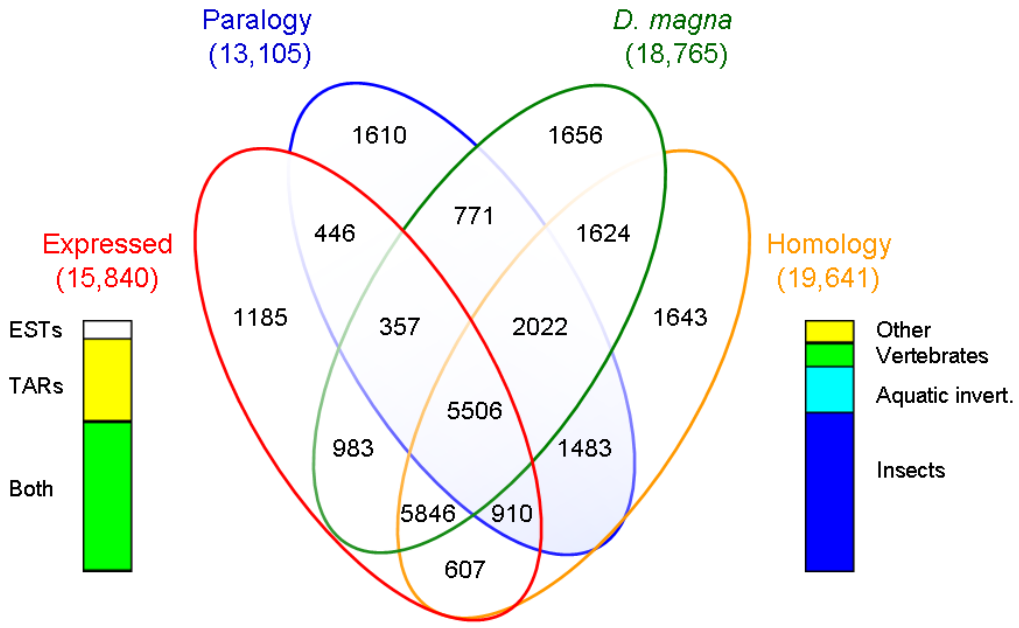


Figure S8. Cumulative frequency distribution of the ratio of non-synonymous (K_a) over synonymous nucleotide substitutions (K_s) among duplicated genes in the genome of *Daphnia pulex* and three reference genomes. This analysis purposefully adds evidence (one of six independent assessments presented) of the validity of the large number of annotated/duplicated genes in the sequenced genome, by utilizing the prediction that no evidence of purifying selection is expected from mistakenly annotated duplicated genes. Measurements are obtained from 34,550 pairs of *D. pulex* duplicated genes sharing a minimum of 60% amino acid identity, compared to 9,562 pairs for *Homo sapiens*, 5,048 pairs for *Caenorhabditis elegans* and 1,367 pairs for *Drosophila melanogaster*. Median K_a/K_s for *D. pulex* is 0.38, while 90% of the measurements are below 0.83. *Daphnia pulex* had the lowest proportion of paralogs < 0.5 K_a/K_s (a metric for purifying selection), but not much different than for human. This pattern is caused by *Daphnia* and *Homo* having a disproportionate number of very recent duplicates compared to *Caenorhabditis* and *Drosophila* (Figure S17), since K_a/K_s cannot be calculated when there is no divergence between paralogs.

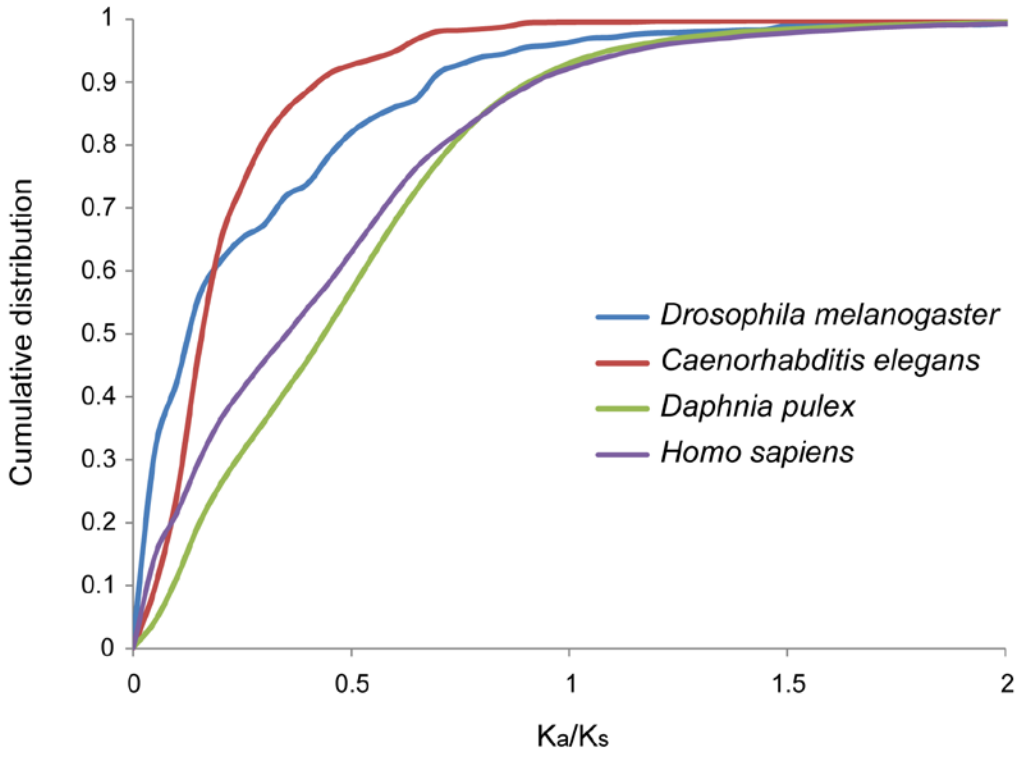


Figure S9. Evidence that genes residing in areas of low read coverage within the draft genome assembly are genuine. Results from the comparative genomic hybridization of TCO labeled DNA on a custom 12-plex microarray manufacture by Roche NimbleGen Inc. containing 3 unique probes for 21,133 predicted genes, 2 unique probes for 8,307 predicted genes and 1 unique probe for 129 unique genes representing 96% of the total predicted gene set. The experiment was replicated 24 times; no correlation is found between read coverage and the mean fluorescing units of probes representing genes.

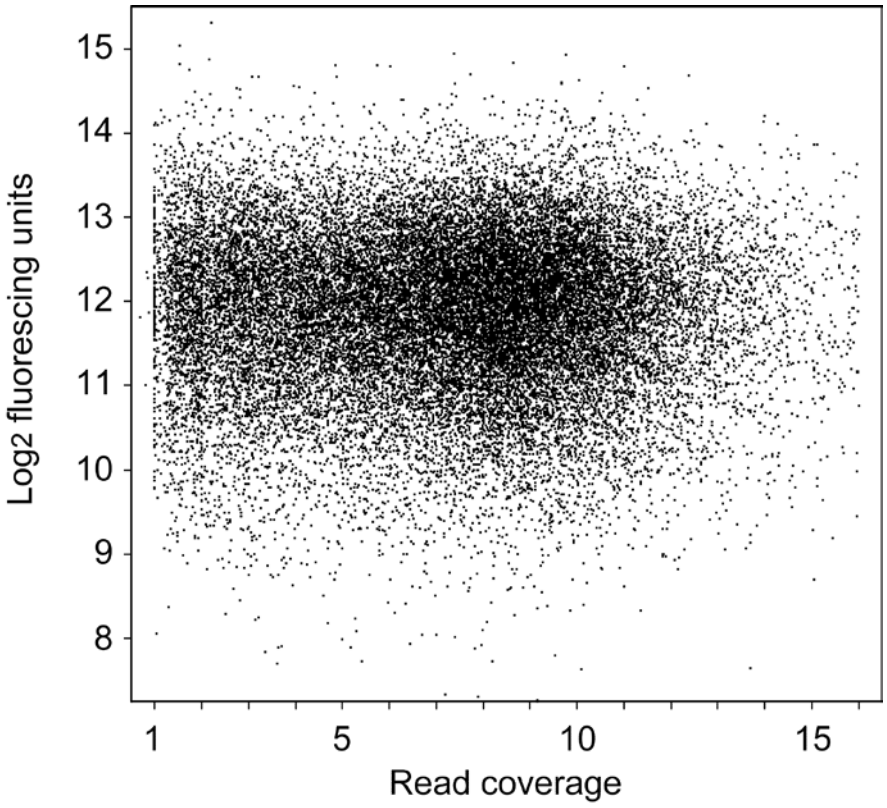
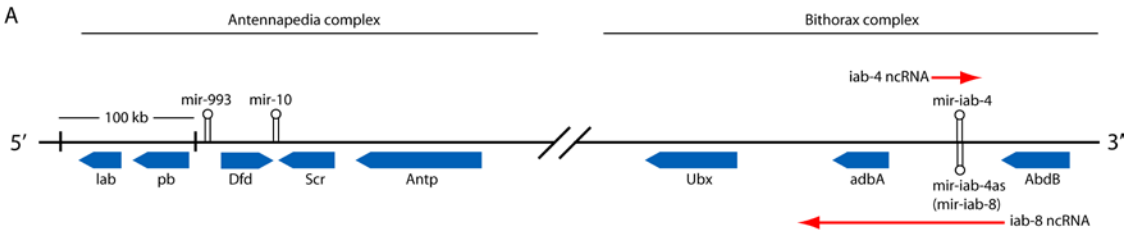
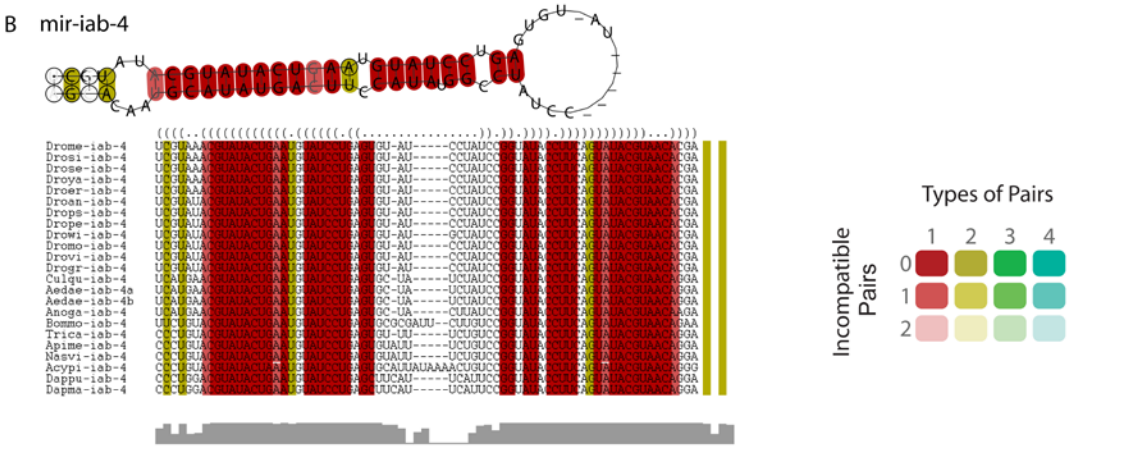
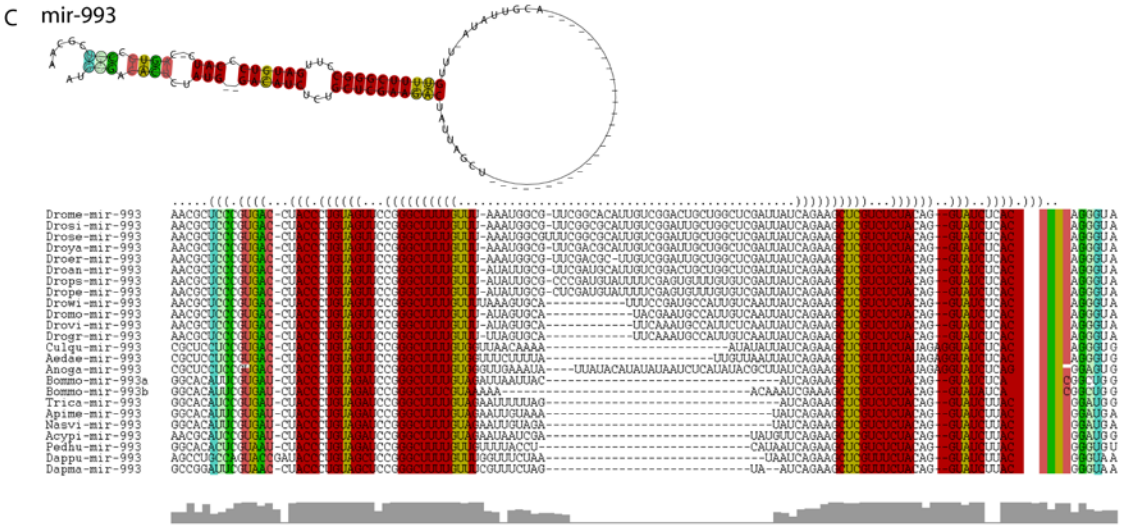


Figure S10. *Daphnia pulex* reveals arthropod origin of two Hox cluster encoded microRNAs (iab-4 and mir-993). **(A)** *Drosophila melanogaster* Hox cluster arrangement. Alignments and secondary structures for conserved microRNAs **(B)** iab-4, **(C)** mir-993 and **(D)** mir-10. The color hue code of the alignment indicates the number of different consistent nucleotide pairs occurring for a given base pair. The saturation of the color indicates the number of sequences that are not consistent with the base pair, in the sense that they have nucleotides at the relevant positions that do not form one of the six standard RNA base pairs. Abbreviations: Drome, *Drosophila melanogaster*; Drosi, *D. simulans*; Drose, *D. sechellia*; Droya, *D. yakuba*; Droer, *D. erecta*; Droan, *D. ananassae*; Drops, *D. pseudoobscura*; Drope, *D. persimilis*; Drowi, *D. willistoni*; Dromo, *D. mojavensis*; Drovi, *D. virilis*; Drogr, *D. grimshawi*; Culpi, *Culex quinquefasciatus*; Aedae, *Aedes aegypti*; Anaga, *Anopheles gambiae*; Bommo, *Bombyx mori*; Trica, *Tribolium castaneum*; Apime, *Apis mellifera*; Nasvi, *Nasonia vitripennis*; Acypi, *Acythrosiphon pisum*; Dappu, *Daphnia pulex*; Dapma, *Daphnia magna*; Lotgi, *Lottia gigantean*; Danre, *Danio rerio*; Musmu, *Mus musculus*.


A
Antennapedia complex
Bithorax complex
iab-4 ncRNA
mir-iab-4
mir-993
mir-10
100 kb
5′
3′
lab
pb
Dfd
Scr
Antp
Ubx
adbA
AbdB
mir-iab-4as
(mir-iab-8)
iab-8 ncRNA


B mir-iab-4
Types of Pairs
Incompatible Pairs
1 2 3 4
0
1
2


C mir-993

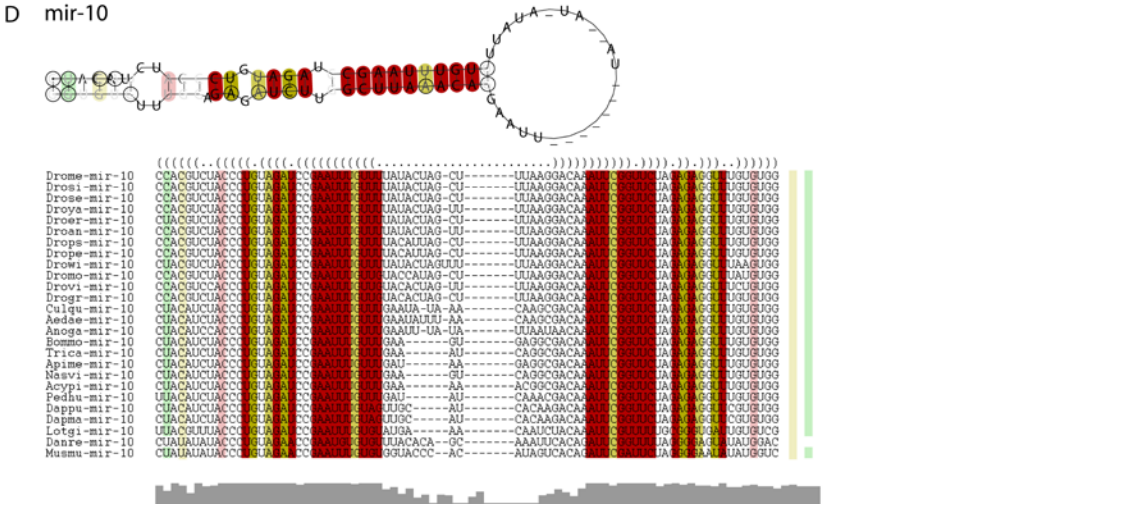

D mir-10

Figure S11. Distribution of transposon *Pokey* in the ribosomal DNA of *Daphnia pulex*. **A.** DAPI-stained mitotic chromosomes. **B.** rRNA gene clusters (green) revealed by fluorescence *in situ* hybridization (FISH). The intergenic spacer (IGS) was used as probe DNA. Red represents counterstained DNA. **C.** IGS (green) and *Pokey* (red) visualized on a stretched chromatin fiber by fiber-FISH. *Pokey* elements are clearly dispersed along the whole length of the rDNA array. Bars represent 2 μm.

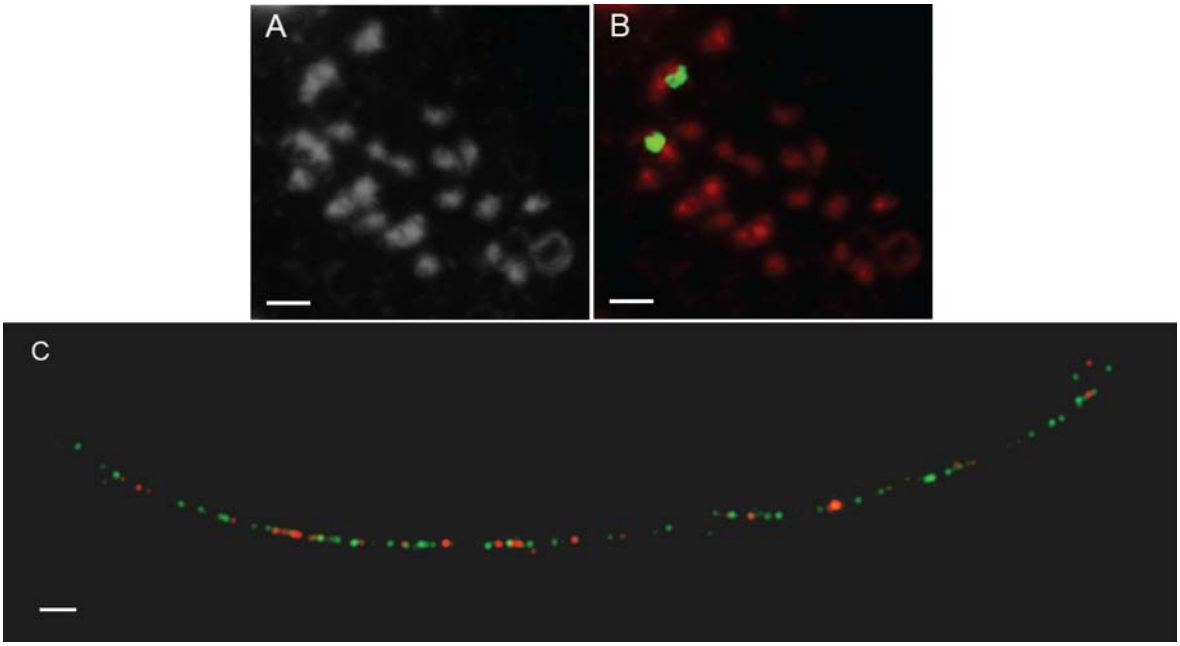
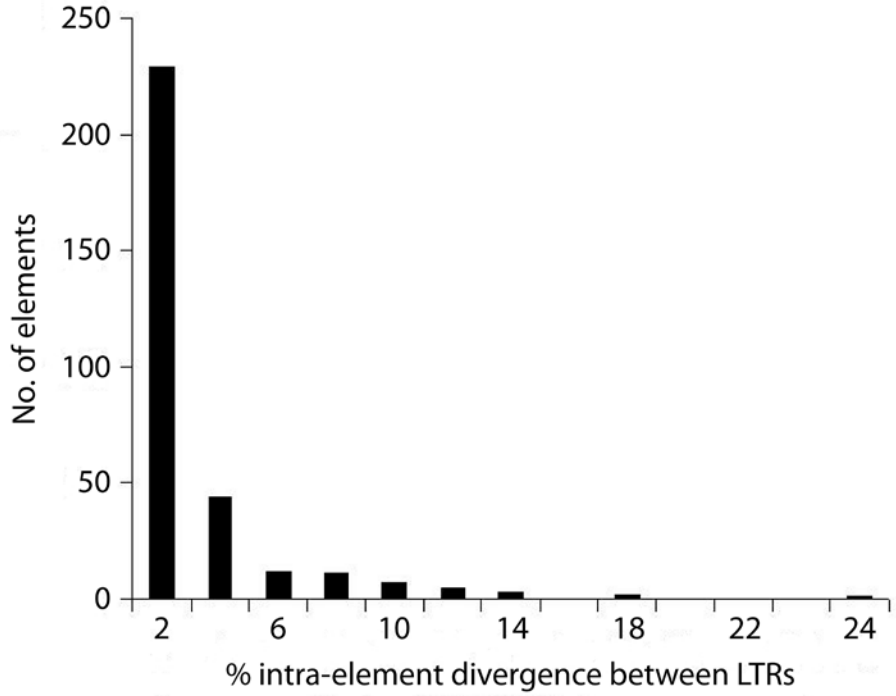


Figure S12. Age distribution of *Daphnia pulex* Long Terminal Repeats elements (LTRs) as pair-wise divergence at nucleotide positions at the termini.



D. Attributes of a Compact Genome

Figure S13. Size distribution of introns in *Daphnia pulex, Caenorhabditis elegans* (smaller, gene rich genome), *Drosophila melanogaster* (relatively small arthropod genome), and *Mus musculus* (large, gene rich genome). **A.** Distributions of the cumulative density and intron size comparing the four species. **B.** Density distributions of intron size for the four species. **C.** Same density distribution as in panel B, observed by scaling down the y-axis values to show bimodal distributions in genomes except for the *D. pulex* genome.

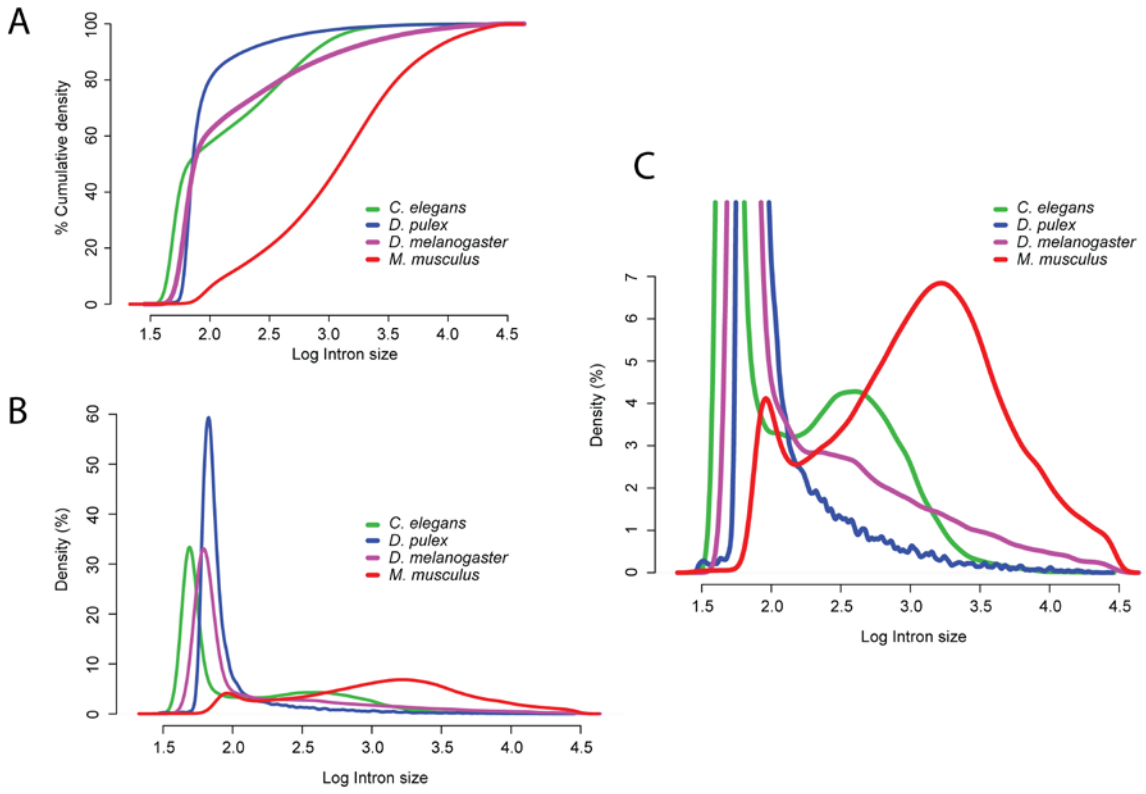


Figure S14. Pair-wise percentage of conservation of intron position. The numbers were obtained by dividing the number of shared introns by the total number of introns in the given two species and converting the result to percentage and clustering using the UPGMA algorithm. Scale bar represents percent divergence. These results were validated by a subsequent analysis using clusters that only contained EST validated *D. pulex* introns. Validated introns were identified by the application of PASA (Program to Assemble Spliced Alignments), which produced 114,128 valid alignments from 166,289 high quality ESTs representing 15,827 genes. A link to the PASA analysis and exploratory tools for the *Daphnia* project is listed in Table S1. The number of annotated *Daphnia pulex* introns supported by ESTs is 33,386. The result from this validation test is the same and is presented in Table S23.

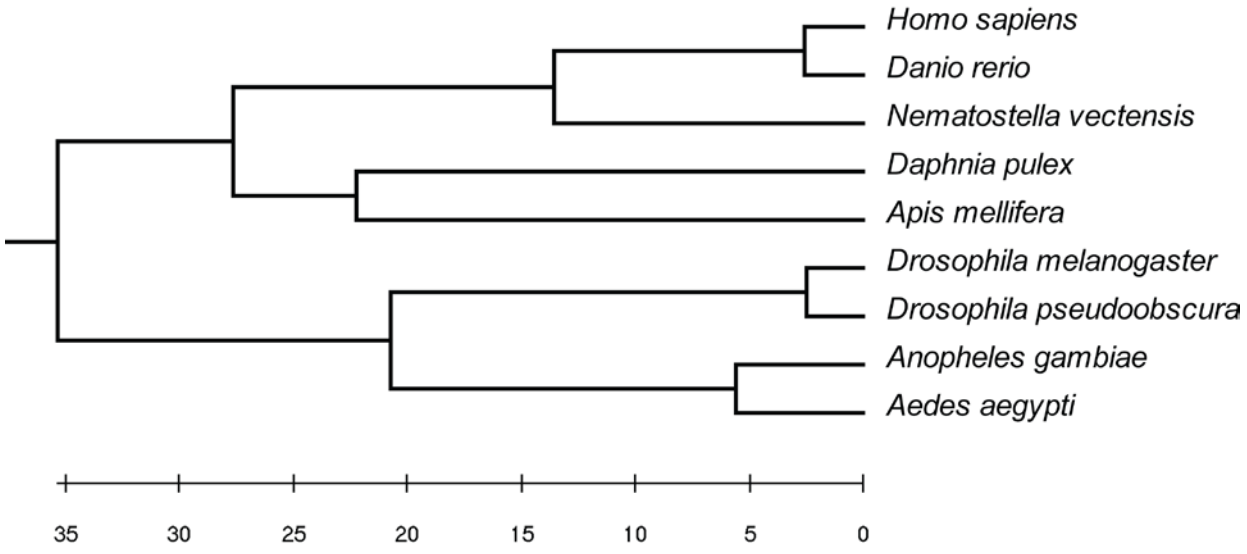


Figure S15. Ancestral reconstruction of intron gains and losses for arthropods and three other metazoans using Maximum Likelihood methods. The *Daphnia* lineage shows a burst of new introns. The node sizes are proportional to the intron content. The green bars indicate intron gain events and scaled by the maximum gain (in *Daphnia*). The red bars indicate intron loss events and scaled by the maximum loss (in Arthropoda ancestor). **Abbreviations:** Anaga, *Anopheles gambiae*; Aedae, *Aedes aegypti*; Drome, *Drosophila melanogaster*; Drops, *Drosophila pseudoobscura*; Apime, *Apis mellifera*; Dappu, *Daphnia pulex*; Homsa, *Homo sapiens*; Danre, *Danio rerio*; Nemve, *Nematostella vectensis*, which is used as the outgroup species.

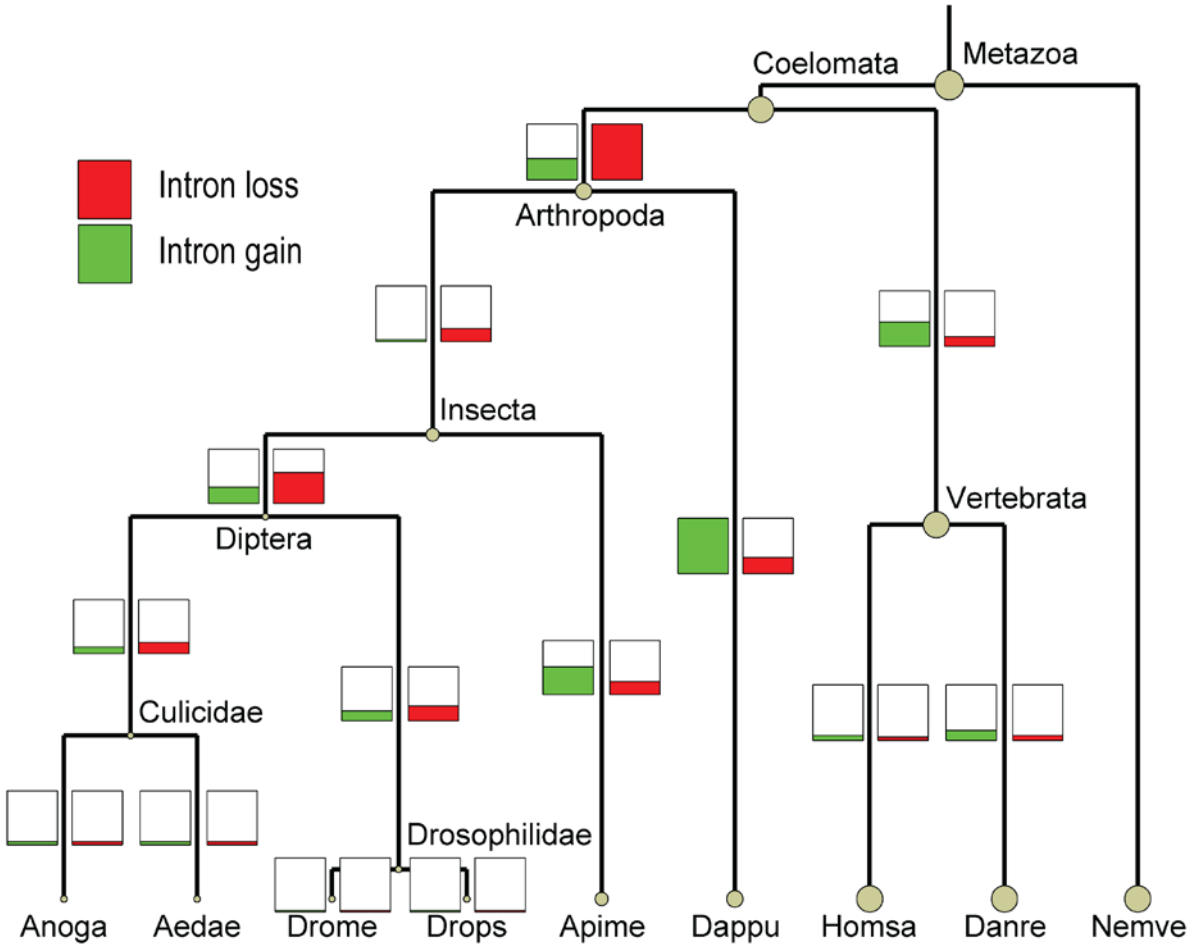
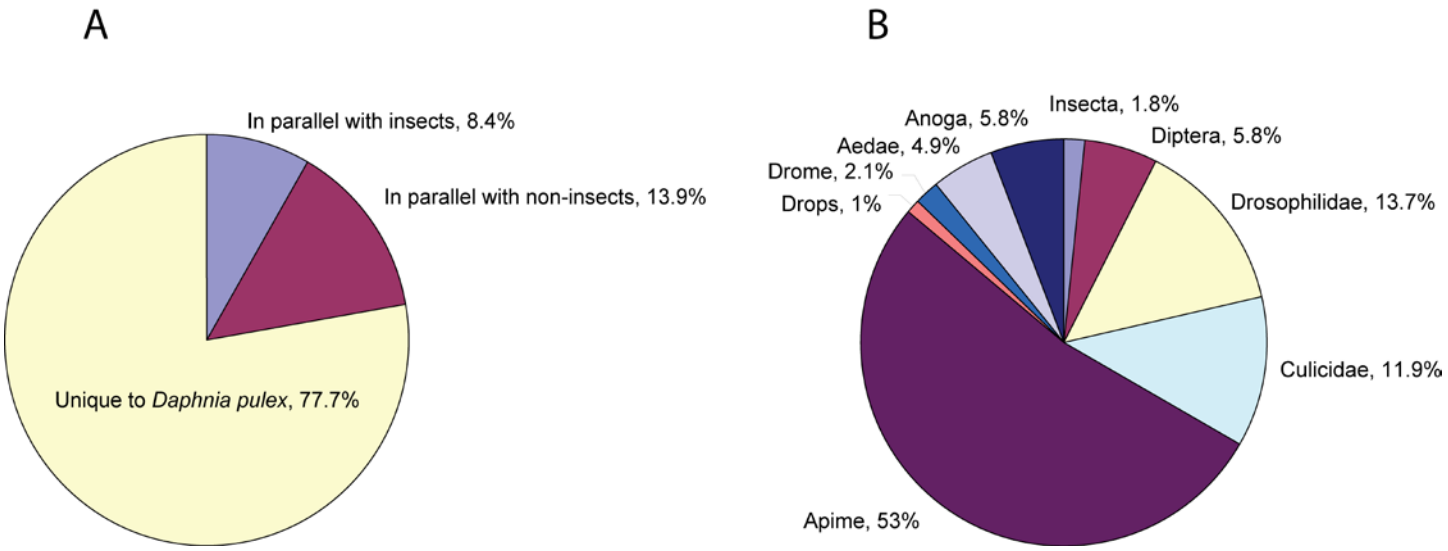


Figure S16. Estimated independent and parallel gain of introns in *Daphnia*. **A.** Estimated independent intron gains in *D. pulex* and parallel gains with arthropods and non-arthropod animals. **B.** Estimated parallel gains in *D. pulex* and different arthropod lineages. **Abbreviations:** Anaga, *Anopheles gambiae*; Aedae, *Aedes aegypti*; Drome, *Drosophila melanogaster*; Drops, *Drosophila pseudoobscura*; Apime, *Apis mellifera*.



E. Origin and Preservation of *Daphnia pulex* Genes

Figure S17. Frequency of pair-wise genetic divergence at silent sites (K_s) among the 2-member gene duplicates in the *Daphnia pulex*, *Caenorhabditis elegans* and *Homo sapiens* genomes. The comparisons are made between genes with greater than 100 aligned amino acids and with percent identity better than 40%. Here, 1,437, 949 and 962 pair-wise comparisons are made for the three genomes, respectively.

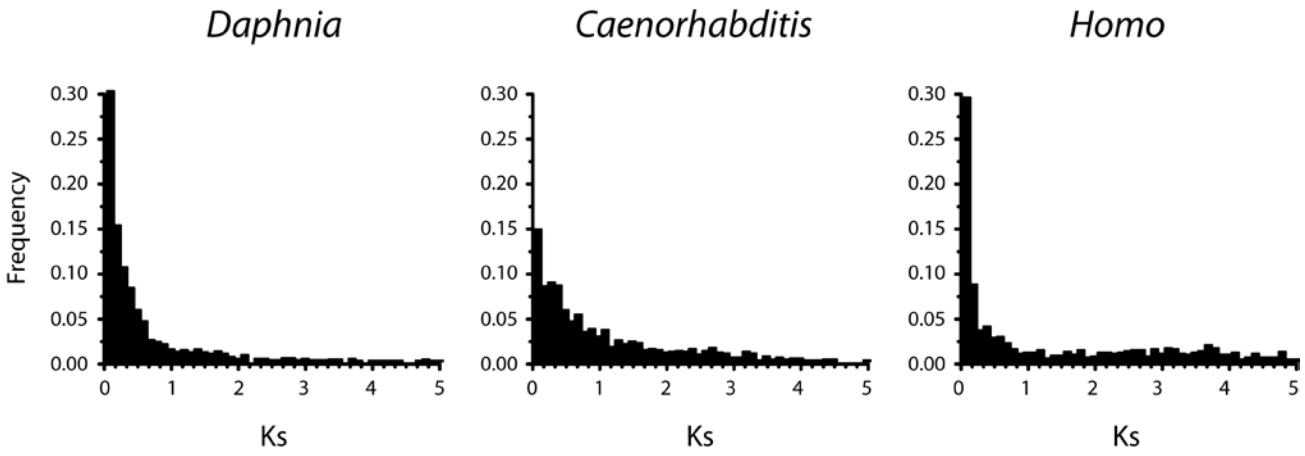


Figure S18. Frequency of pair-wise genetic divergence at silent sites (K_s) among gene duplicates in *Daphnia pulex*. Panels include genes with greater than 100 aligned amino acids and with percent identity better than 40%. **A.** 66,502 pairs including gene duplicates with evidence of gene conversion. **B.** 60,444 pairs excluding gene duplicates with evidence of gene conversion. We find that our estimate of the age distribution of duplicated genes is unaffected by gene conversion.

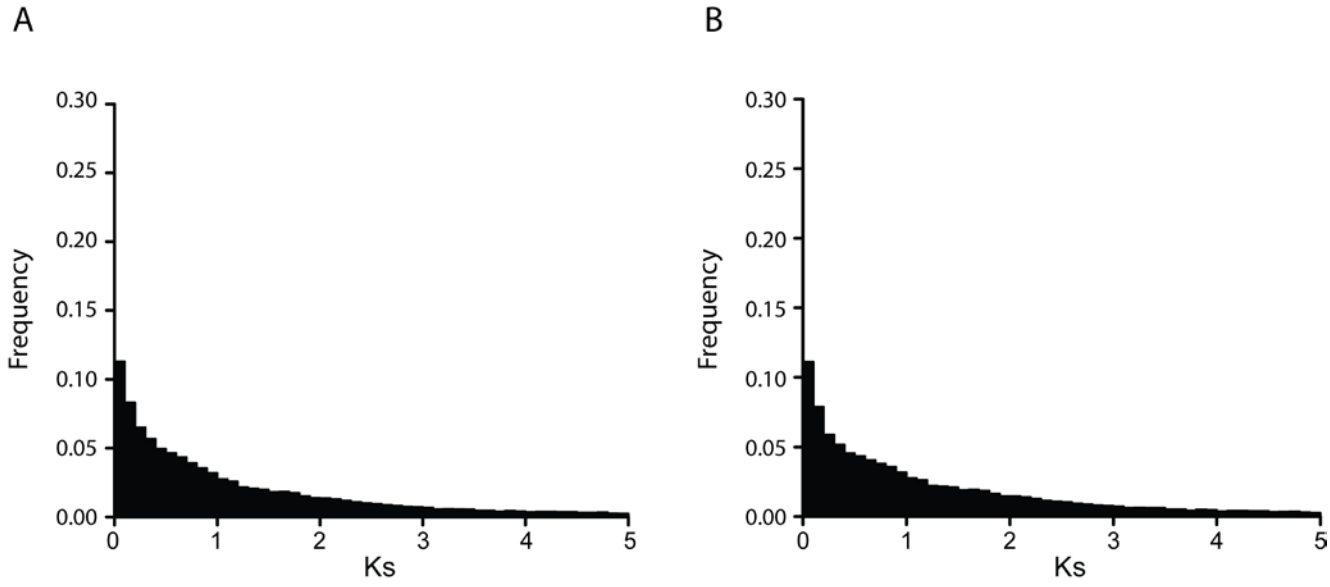


Figure S19. Position and size of Tandem Duplicated Gene (TDG) clusters within the genome assemblies of four model species. Clusters are identified using the custom algorithm called Tandy (described in Methods section). The bottom axis plots all genes binned in groups of 50 and ordered from largest scaffold/chromosome to the smallest. The peak heights along the y-axis represent the percentage of genes that are simple tandem gene repeats (red) and mixed tandem gene repeats (green) within that 50 genes window in the respective genomes. In *Caenorhabditis elegans*, the largest TDG clusters have biased genomic distributions, as previously reported [S155]. In *Daphnia pulex*, TDG clusters are larger on average (scaled to 23,791 genes following the removal of small scaffolds under 80 Kb, compared to 20,062 in *C. elegans*), yet are more evenly distributed among the genome scaffolds. The *Drosophila melanogaster* and *Mus musculus* genomes also contain TDG clusters, yet these are comparatively less prominent.

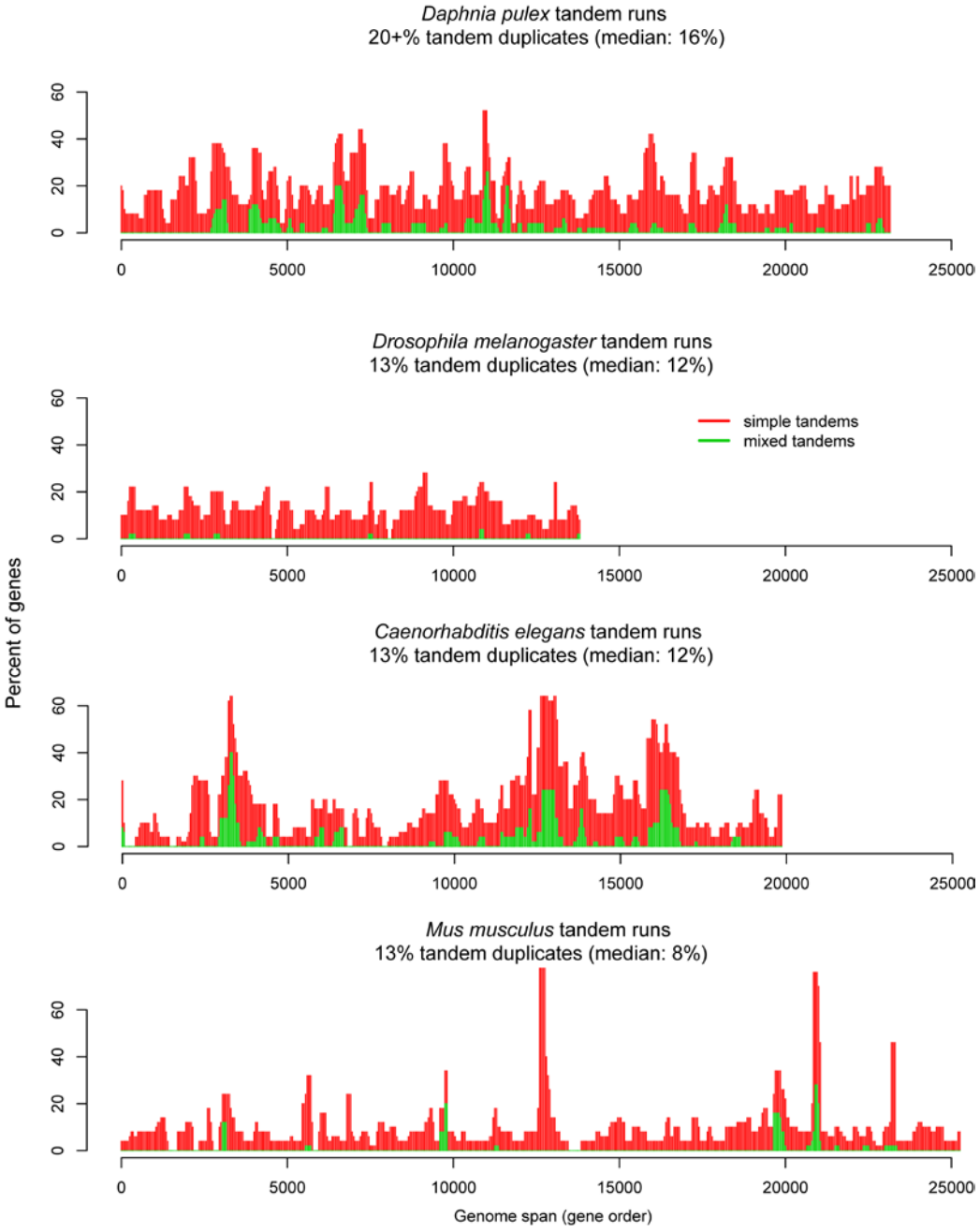


Figure S20. Physical distances between neighboring members of large duplicated gene families composed of 10-80 genes within the *Daphnia* and three reference genomes. *Daphnia*'s duplicated genes are most densely arranged into clusters. Observations are binned within intervals of 0-5 Kb, 5-10 Kb, 10-20 Kb, 20+ Kb and duplicates distributed among different sequence assembly scaffolds (unlinked). The last two bins are scaled 2x for the y-axis values. Shaded fractions designate inverted duplicates (shaded portions of bar graph). Nearby tandem duplicates show a lower inversion rate than other species. *Daphnia*'s genome shows an excess in unlinked (across-scaffold) duplicate genes as well as very near 1,000 – 2,000 bp tandem genes. As this draft genome assembly has thousands of small scaffolds, the unlinked duplicates may be found to be nearby tandems with further assembly refinement. The small scaffolds likely failed to assemble in part due to tandem duplicate gene regions. **Abbreviations:** Caeel, *Caenorhabditis elegans*; Dappu, *Daphnia pulex*; Drome, *Drosophila melanogaster*; Musmu, *Mus musculus*.

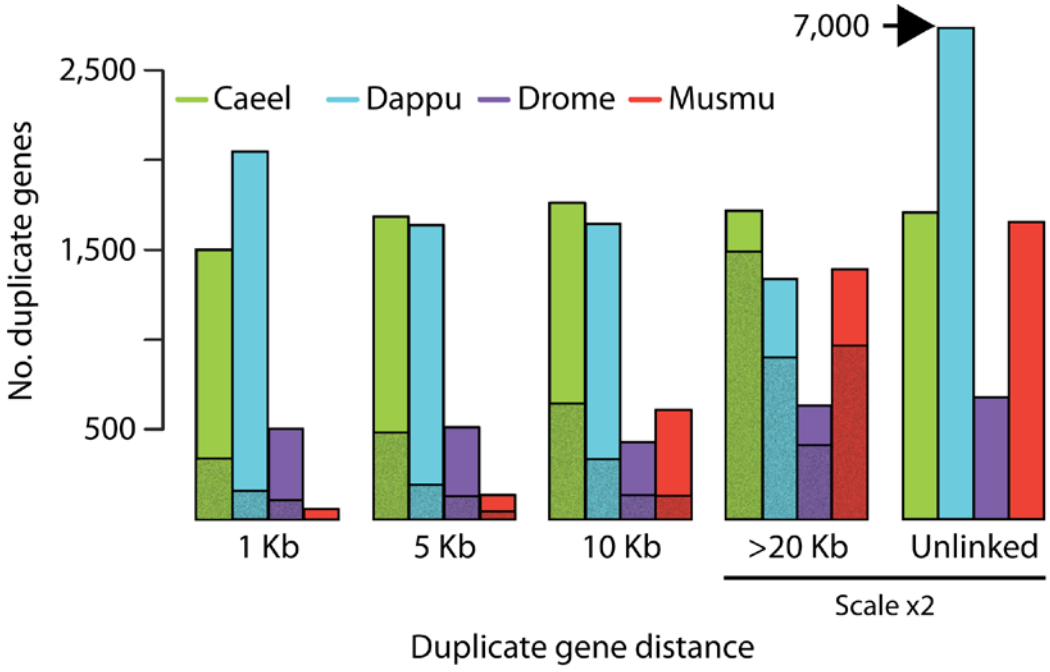


Figure S21. Phylogenetic relationships of 39 of the 46 *Daphnia pulex* opsin genes (listed in Table S32), labeled in red; Some clusters of the most recent gene duplications within the Crustacea Long wave-length opsins Lineages A and B are not shown because full-length protein sequences are not available for these tandemly duplicated genes that failed to assemble) and representative animal opsins (labeled in blue for crustacean sequences and in black for all others). Opsins are members of the GPCR-class family of proteins that mediates phototransduction cascades in eumetazoan animals. The phylogeny is constructed by first aligning amino acids using MUSCLE [S58] and assuming the a WAG+I+Γ model of amino acid evolution, as implemented in RaxML [S96]. The resulting phylogeny is rooted by the ciliary subfamily [S156]. At left, bootstrap support at the nodes is reported as concordance among 100 pseudo-replications, with nodes with <49% support collapsed. Several major opsin clades are labeled. Although low bootstrap support is obtained for the RGR/Go subfamily (49%), analysis of intron locations supports their monophyly [S157], as does more extensive sequence phylogenies (e.g. [S156]). At right is a phylogram showing branch lengths proportional to inferred number of amino acid changes. Gene names are the genus of the containing species, plus a number or accession number to identify uniquely multiple genes from the same species. We included all Branchiopoda opsin genes from a recent publication that studied three species [S94]. **Abbreviations for *D. pulex* genes:** LOP=Long-wave opsin; UVOP = Ultraviolet Opsin and BLOP = Blue Opsin are named based on similarity to functionally characterized opsins of other species, no functional analyses have yet been performed for these; 'Arthropsin' is used to describe a new clade of opsins, known so far only from *Daphnia*. We find these genes to be a sister-group to 'rhabdomeric' (Gq-coupled) opsins with strong support (100%). Based on multiple three-taxon, maximum likelihood relative rate tests implemented in HyPhy [S158], we found no evidence for rapid rates of evolution in arthropsin genes, and therefore no support for long branch topological artifacts [S159] caused by rapid arthropsin evolution.

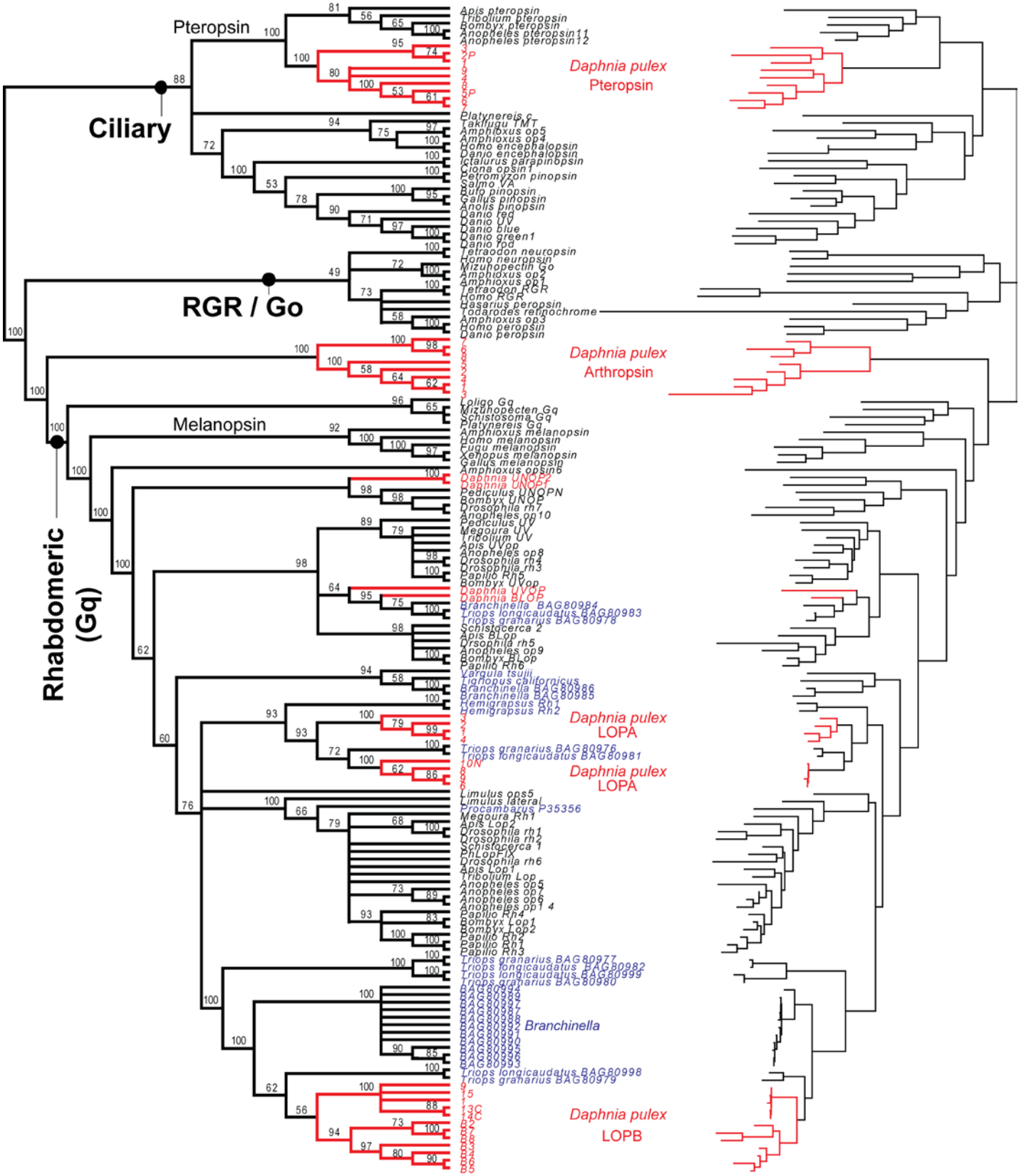
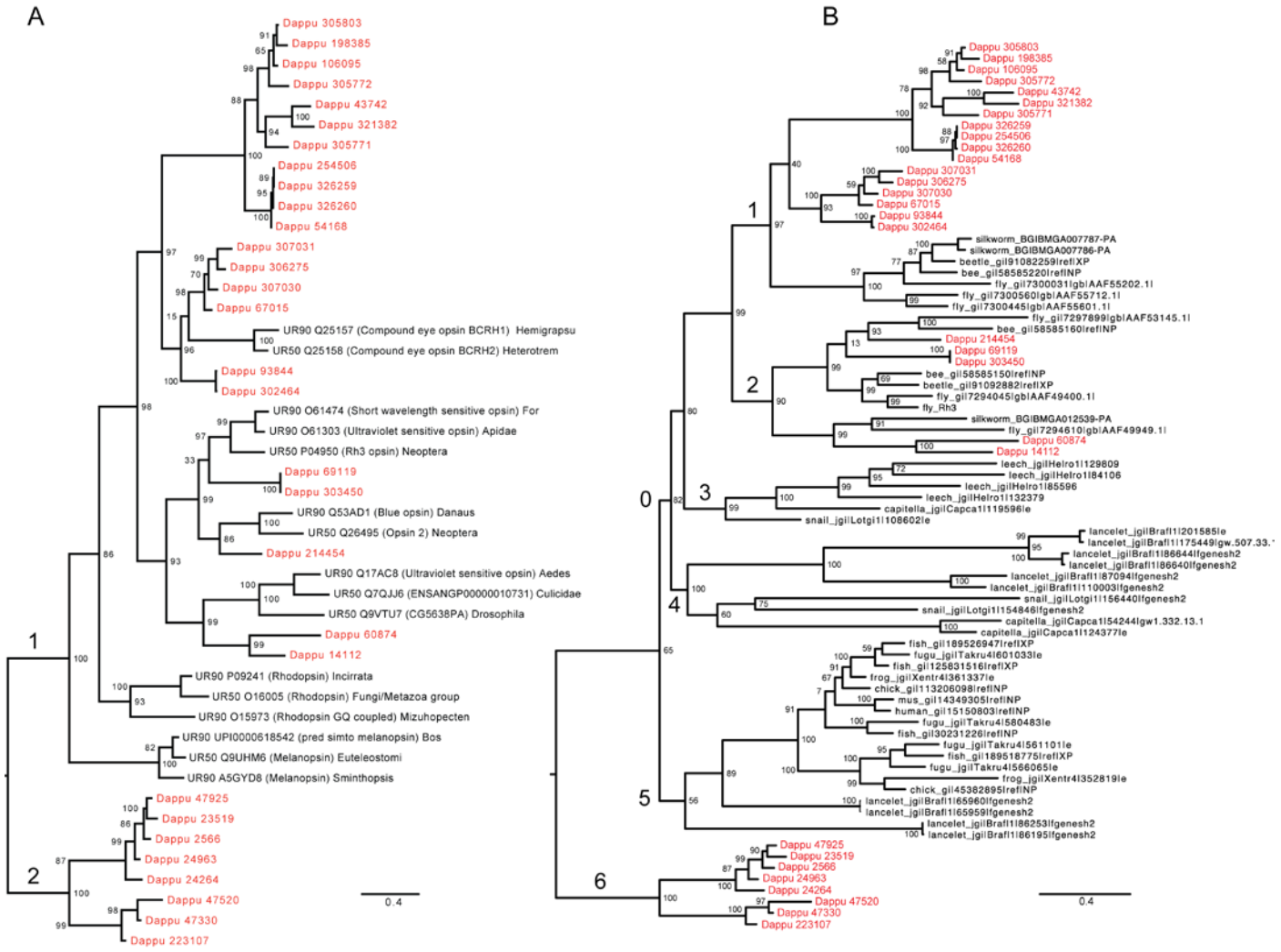

Pteropsin
Ciliary
RGR / Go
Melanopsin
Rhabdomeric (Gq)
Daphnia pulex Pteropsin
Daphnia pulex Arthropsin
Daphnia pulex LOPA
Daphnia pulex LOPA
Daphnia pulex LOPB
Branchinella
Apis pteropsin
Tribolium pteropsin
Bombyx pteropsin
Anopheles pteropsin11
Anopheles pteropsin12
Platynereis c
Takifugu TMT
Amphioxus op5
Amphioxus op4
Homo encephalopsin
Danio encephalopsin
Ictalurus parapinopsin
Ciona opsin1
Petromyzon pinopsin
Salmo VA
Bufo pinopsin
Gallus pinopsin
Anolis pinopsin
Danio red
Danio UV
Danio blue
Danio green1
Danio rod
Tetraodon neuropsin
Homo neuropsin
Mizuhopectin Go
Amphioxus op2
Amphioxus op1
Tetraodon RGR
Homo RGR
Hasarius peropsin
Todarodes retinochrome
Amphioxus op3
Homo peropsin
Danio peropsin
Loligo Gq
Mizuhopecten Gq
Schistosoma Gq
Platynereis Gq
Amphioxus melanopsin
Homo melanopsin
Fugu melanopsin
Xenopus melanopsin
Gallus melanopsin
Amphioxus opsin6
Daphnia UVOP2
Daphnia UVOP1
Pediculus UVOPN
Bombyx UVOP
Drosophila rh7
Anopheles op10
Pediculus UV
Megoura UV
Tribolium UV
Apis UVop
Anopheles op8
Drosophila rh4
Drosophila rh3
Papilio Rh5
Bombyx UVop
Daphnia UVOP
Daphnia BLOP
Branchinella BAG80984
Triops longicaudatus BAG80983
Triops granarius BAG80978
Schistocerca 2
Apis BLop
Drosophila rh5
Anopheles op9
Bombyx BLop
Papilio Rh6
Vargula tsujii
Tigriopus californicus
Branchinella BAG80986
Branchinella BAG80985
Hemigrapsus Rh1
Hemigrapsus Rh2
Triops granarius BAG80976
Triops longicaudatus BAG80981
Limulus ops5
Limulus lateral
Procambarus P35356
Megoura Rh1
Apis Lop2
Drosophila rh1
Drosophila rh2
Schistocerca 1
PhLopFIX
Drosophila rh6
Apis Lop1
Tribolium Lop
Anopheles op5
Anopheles op7
Anopheles op6
Anopheles op1 4
Papilio Rh4
Bombyx Lop1
Bombyx Lop2
Papilio Rh2
Papilio Rh1
Papilio Rh3
Triops granarius BAG80977
Triops longicaudatus BAG80982
Triops longicaudatus BAG80999
Triops granarius BAG80980
Triops longicaudatus BAG80998
Triops granarius BAG80979

Figure S22. Maximum-likelihood phylogenies of *Daphnia pulex* opsin genes for comparison to evolution of other gene families involved in vision and eye development [S66]. Bootstrap support at the nodes is reported as concordance among 100 pseudo-replications. We included two clades of opsins in this analysis: rhabdomeric opsins (panel **A**, lineage 1), and the newly described arthropsin clade (panel **A**, lineage 2). Consistent with previous analyses [S97], this analysis recovers rhabdomeric opsins only from the bilaterian animals. A reconciled tree analysis (inferring the timing of gene duplication and loss events by comparing a gene tree to a species tree [S66, S160]) identifies 43 well-supported gene duplication events in the evolutionary history of rhabdomeric and arthropsin opsins across all taxa examined – far more than any other phototransduction locus considered by [S66]. Twenty-five of these duplications occurred within the *D. pulex* lineage alone. One duplication of rhabdomeric opsin predated the bilaterians (panel **B**, lineage 0); two duplications occurred at least prior to the origin of the Pancrustacea (panel **B**, lineages 1 and 2), and two duplications preceded the evolution of the vertebrates (Panel **B**, lineages 4 and 5). This analysis also recorded 13 loss events for rhabdomeric opsins. Because the node joining arthropsins (panel **B**, lineage 6) to the larger rhabdomeric opsin radiation was weakly supported in this analysis (alrt = 65) (panel B) and because we assigned loss and gain using nodes supported by alrt = 0.9 or greater, we did not record any loss events for this clade. However, the finding that the *Daphnia*-specific opsin clade arthropsin is basal to rhabdomeric opsins in rooted analyses (Figure S21) suggests that a more complicated history of loss for arthropsins is likely. **Panel B lineages**: 1 = Rh6/Rh2; 2 = Rh3/Rh4/Rh5/Rh7; 3 = Loph Rh; 4 = Bilaterian Rh; 5 = Melanopsin; 6 = Arthropsin.



F. Consequence *Daphnia*'s Genome Structure

Figure S23. Rates of gene conversion (as percent of converted paralogs) and number of intervening genes between duplicates in *Daphnia pulex* (blue) and the average of five *Drosophila* species (red: *D. melanogaster*, *D. yakuba*, *D. pseudoobscura*, *D. virilis*, *D. grimshawi*). Values on the X-axis represent intervening genes between pairs of duplicates. Strictly tandem pairs have zero intervening genes. Bars above and below the mean values are maximum and minimum values among the *Drosophila* species.

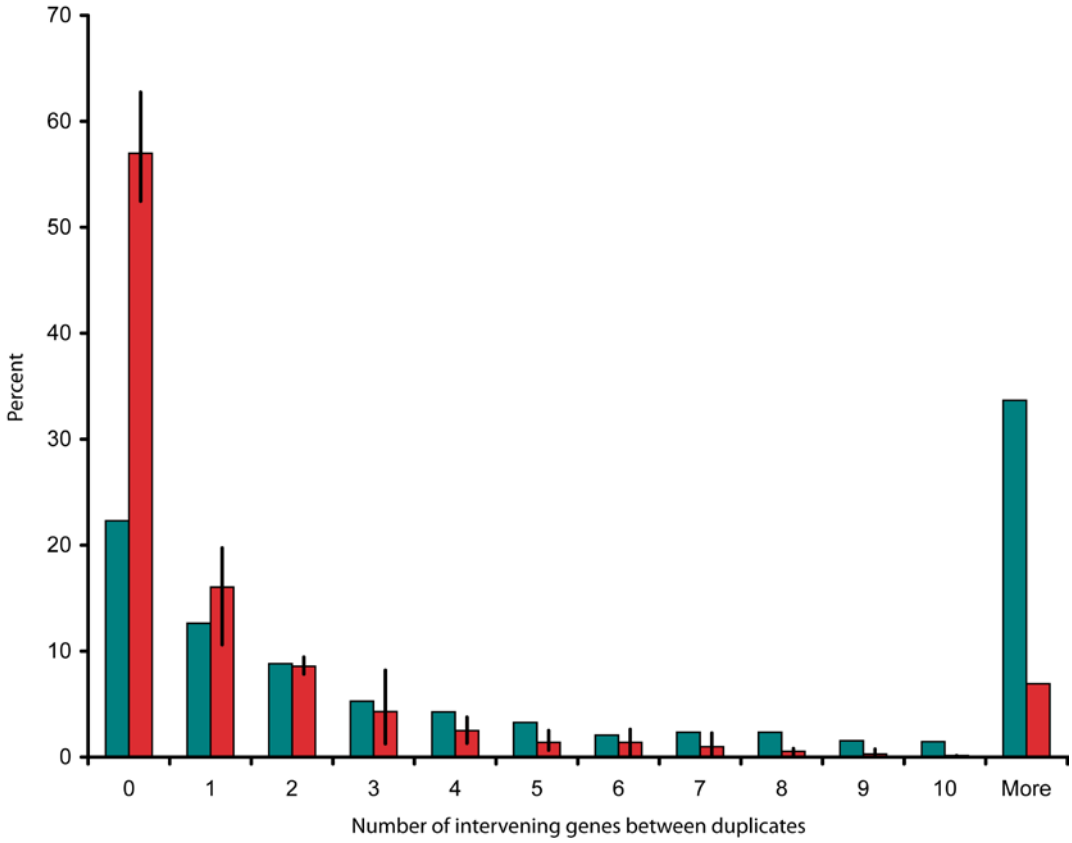


Figure S24. Rates of gene conversion (as percent of converted paralogs) and divergence between duplicates in *Daphnia pulex* (blue) and the average of five *Drosophila* species (red: *D. melanogaster*, *D. yakuba*, *D. pseudoobscura*, *D. virilis*, *D. grimshawi*). Values on the x-axis represent divergence estimates for synonymous nucleotide substitutions. Bars above and below the mean values are maximum and minimum values among the *Drosophila* species.

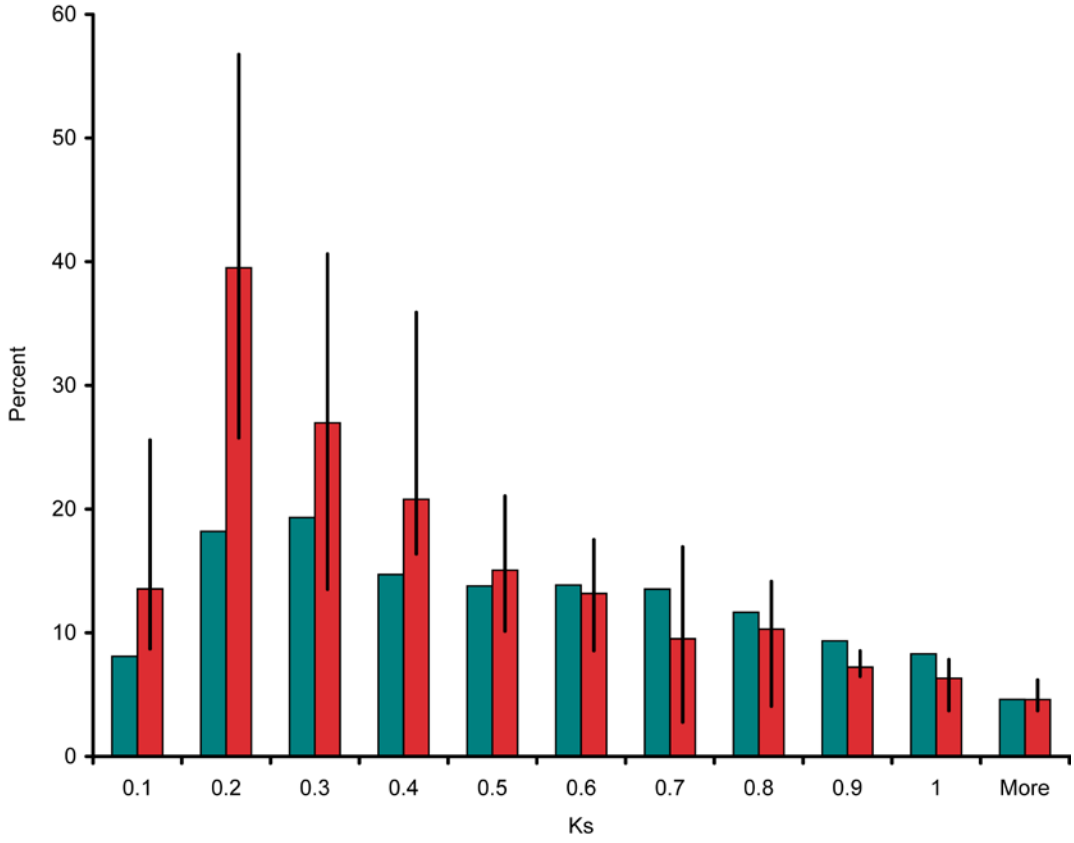


Figure S25. Amino acid sequence alignment of di-domain hemoglobins (Hb) of *Daphnia pulex* and *D. magna*. The amino acid sequences used in the alignment and their accession number of NCBI/EMBL/DDBJ databases are: *Moina macroccopa* Hb1 and Hb2 (AB055113, AB055114), *Barbatia lima* Hb1 and Hb2 (D63931, D58417), *Barbatia reeveana* Hb (M73328), *Ascaris suum* Hb (L03351), and *Pseudoterranova decipiens* Hb (M63298). A to H helices in the globin folding are indicated above the first amino acid of each helix. N-terminal extension and pre-A are also indicated. The most conserved residues in all Hb are shaded black. Other highly conserved residues are shaded gray. **Abbreviations:** Dpul, *Daphnia pulex*; Dmag, *Daphnia magna*; Mmac, *Moina macroccopa*; Bl, *Barbatia lima*; Br, *Barbatia reeveana*; As, *Ascaris sum*; Pd, *Pseudoterranova decipiens*.

Note. All Hb proteins from both *Daphnia* species and from outgroup species have conserved amino acids, such as a Trp residue at the twelfth position of helix A (A12), Pro (C2), Phe (CD1), His (F8), and Trp (H8), which are important for heme binding in the first and second domains. An exception is found at position F8 containing a substitution of Tyr for His in the first domain of Dpul-Hb9. Generally, positions B10, E7, and E11 residues are most important for oxygen affinity, and amino acid residues at B10 and E11 play a pivotal role in formation of the distal heme pocket [S161]. We find Leu (B10), His (E7), and Val (E11) conserved among vertebrate Hb and myoglobin, while Gln (E7) is common in the invertebrate Hb, including Hb in *Daphnia*. However, Leu is replaced by Phe at position B10 of the second Hb domains in *Daphnia*, except for Dpul-Hb10 and Dpul-Hb11, while Leu at position E11 is replaced by Ile in the first domains and Val in the second domains, respectively. A study of *Ascaris* (nematode) Hb suggests that substituting Leu for Phe increases the rate of oxygen association, resulting in an increase of oxygen affinity [S162], while a 10 fold benefit is observed in myoglobin [S163]. Presumably, a similar equilibrium between oxygen affinity and dissociation is reached by *Daphnia*'s second domains. Finally, *D. pulex* Hb have characteristic Thr rich sequence in their pre-A sequences, located upstream of the first domain of waterflea Hb, except for Dpul-Hb6 and Dpul-Hb9. The identities between the first and the second domain of the same Hb subunit in Cladocera (*Daphnia* plus *Moina*) are remarkably low in contrast to clam and nematode di-domain Hb (average amino acid identities in Cladocera, clam and nematode are 25.1%, 79.2% and 56.7%, respectively). This observation suggests that the duplication of an Hb gene encoding a single heme-binding domain preceded the fusion and formation of di-domain Hb genes in Cladocera, which occurred much earlier than in clam and nematode.

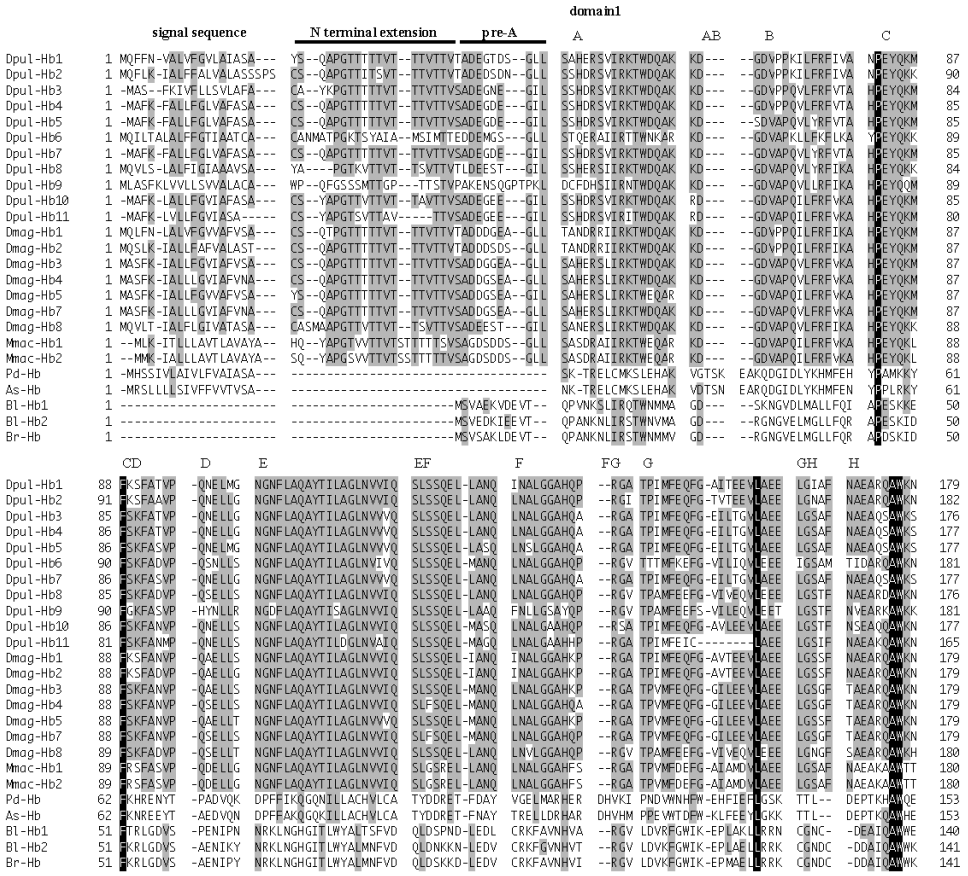

domain1
signal sequence
N terminal extension
pre-A
A
AB
B
C
CD
D
E
EF
F
FG
G
GH
H
domain2
pre-A
A
AB
B
C
CD
D
E
EF
F
FG
G
GH
H

Figure S26. Nucleotide sequence alignment of di-domain hemoglobins (Hb) in coding regions of genes. The most conserved residues in all Hb are shaded black. Other highly conserved residues are shaded gray. **Abbreviations:** dpul, *Daphnia pulex*; dmag, *Daphnia magna*; asuumc, *Ascaris sum*; pdecic, *Pseudoterranova decipiens*.

Note. The nucleotide alignment was analyzed using GENECONV [S99] by assigning a gap penalty of 1 and creating 10,000 permutations for detecting copied DNA with probability $p < 0.05$. Five copied DNA segments in the *D. magna* Hb gene cluster and the eight DNA segments in the *D. pulex* Hb gene cluster were identified. In the *D. magna* Hb gene cluster, gene conversion events occurred between Hb1/2, Hb2/3, Hb2/4, Hb2/5, Hb2/7 and Hb1/7. By contrast, gene conversion events occurred between Hb1/3, Hb2/3, Hb2/4, Hb2/7, Hb3/4, Hb3/7, Hb4/7 and Hb5/7 in the *D. pulex* Hb gene cluster.

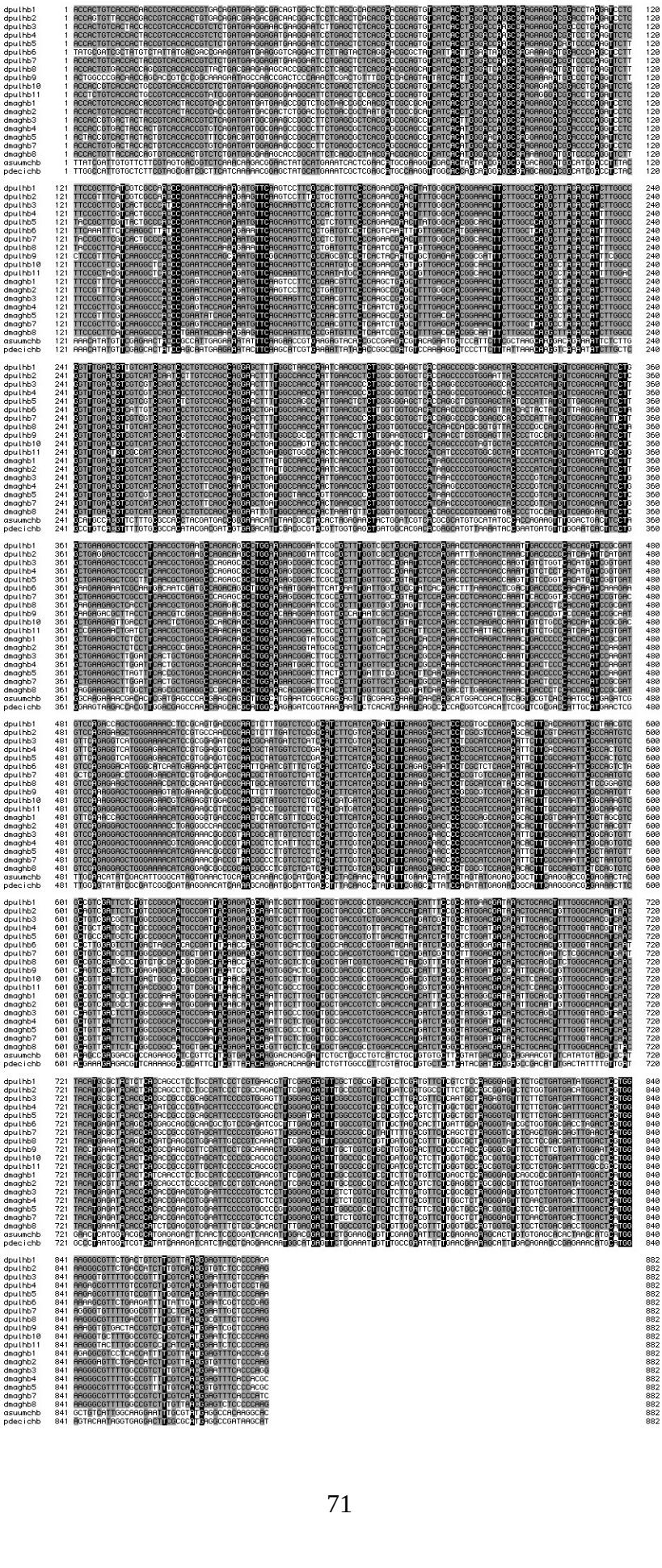
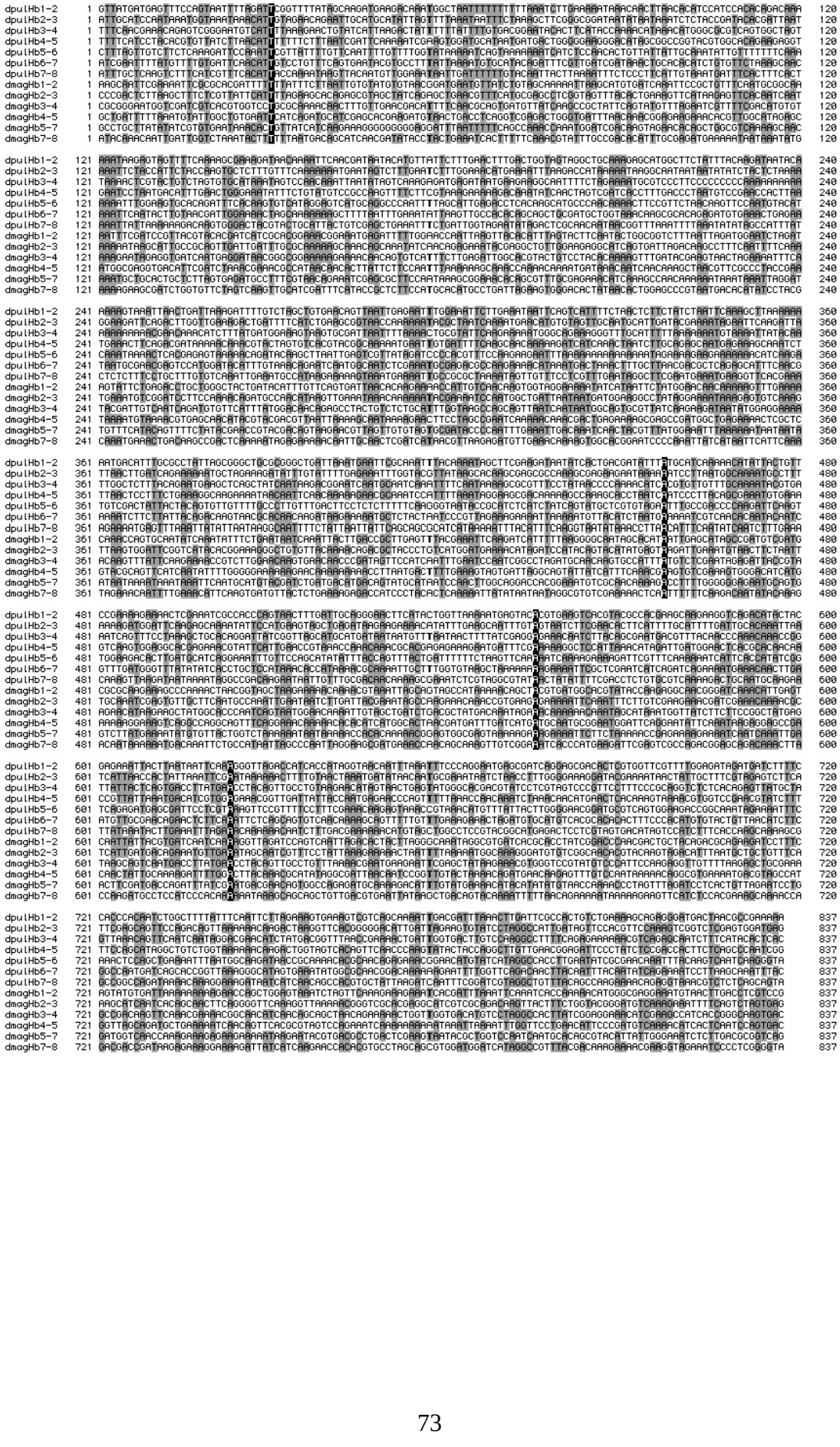


Figure S27. Nucleotide sequence alignment of intergenic regions between the stop codons of upstream genes and the TATA of the downstream genes of all *Daphnia pulex* (dpul) and *D. magna* (dmag) di-domain hemoglobins (Hb). The most conserved residues in all Hb are shaded black. Other highly conserved residues are shaded gray.

Note. Consensus core sequence of HRE (T/G/C ACGTG) in the Hb gene clusters were found by using the homology search tool of GENETYX v11 (www.sdc.co.jp/genetyx, , Tokyo, Japan). Presumptive hypoxia response elements (HREs) in all intergenic regions were identified (Figure 2B). Some elements were accompanied by conserved ancillary sequences (VTACGTG(N)7YCACGY) (Figure 2B, marked with asterisks). Alignment of the intergenic sequences showed that many of them are located exactly the same position relative to the translation start point of the downstream Hb genes in the two clusters (Figure 2B, marked with sharps).



G. Evolutionary Diversification of Duplicated Genes

Figure S28. Differential expression (DE) profiles of 37 of the 46 *Daphnia pulex* opsin genes from eight microarray experiments (A-H). **A.** Heat map showing results from the hierarchical clustering by un-centered expression correlation of genes from all of the major clades. Red designates up-regulation against the reference condition. Green designates down-regulation against the reference condition. Dark shades denote no change in gene expression. **B.** Differential gene expression (DE) pattern correlations among paralogs of the long-wavelength opsin genes, including lineage A (LOPA) and lineage B (LOPB), as a function of their pair-wise genetic divergence at silent sites (K_s). Symbols indicate whether the paralogs both stem from lineage A (triangles), both stem from lineage B (circles), or are each from separate lineages (squares). Observations that are encircled involve all comparisons involving LOPA3 (Dappu-67015). By eliminating this gene with the most divergent expression patterns, long-wavelength opsins are seen to gradually diverge with increasing age (best fit regression line is shown). Relative time since duplication is inferred from K_s.

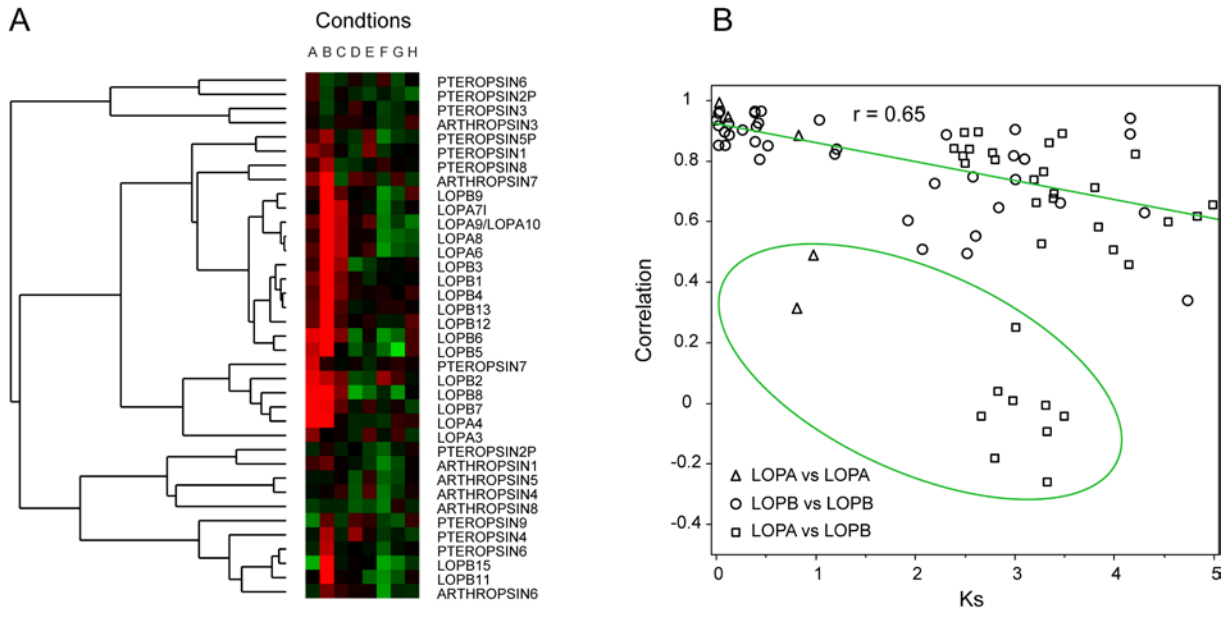
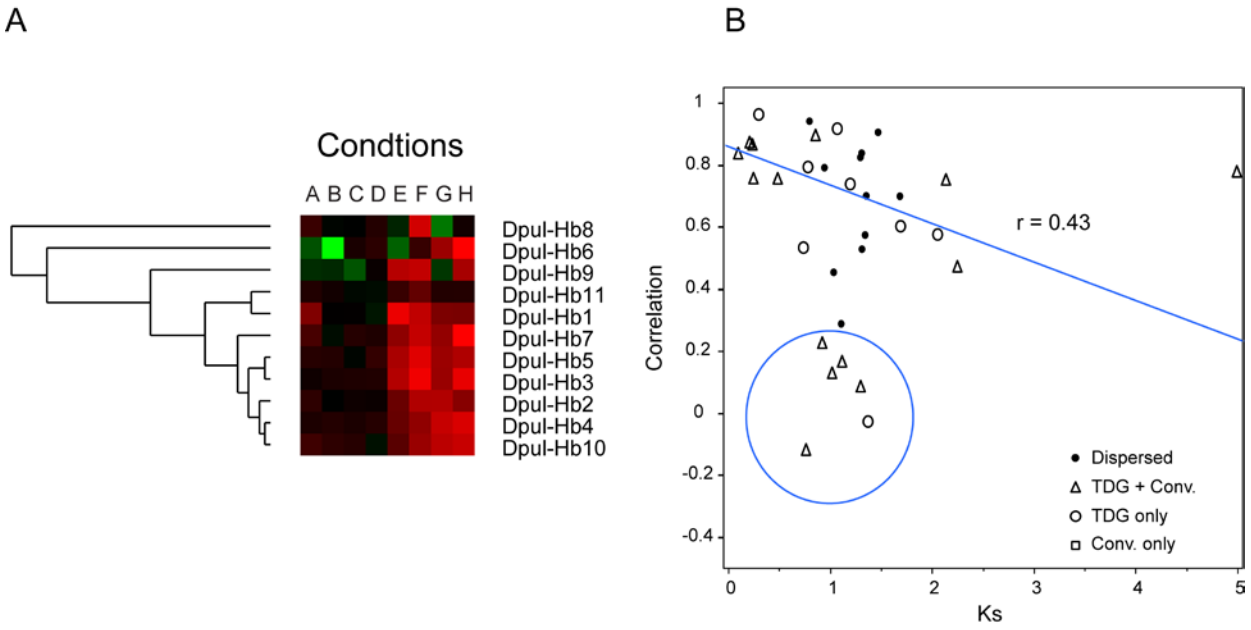


Figure S29. Differential expression (DE) profiles of 11 *Daphnia pulex* di-domain hemoglobin genes from eight microarray experiments (A-H). **A.** Heat map showing results from the hierarchical clustering by un-centered expression correlation of genes from all major clades. Red designates up-regulation against the reference condition. Green designates down-regulation against the reference condition. Dark shades denote no change in gene expression. **B.** Differential gene expression (DE) pattern correlations among paralogs of the di-domain hemoglobin genes, including duplicates that are within the tandem duplicated gene (TDG) cluster (TDG only), duplicates sharing gene conversion tracts (Conv. only), duplicates within TDG clusters that also show signatures of gene conversion (TDG + Conv.), and duplicated genes that are dispersed in the genome, as a function of their pair-wise genetic divergence at silent sites (K_s). Observations that are encircled involve all by one comparison involving Dpul-Hb8 (Dappu-230333). By eliminating these comparisons with the pair of genes with K_s = 5, hemoglobins are seen to diverge with increasing age (best fit regression line is shown). Relative time since duplication is inferred from K_s.



H. Functional Significance of Expanded Gene Families

Figure S30. Thirty-eight expanded and 54 contracted metabolic genes in arthropod genomes compared to vertebrates. All enzymes are supported by the Fisher exact test (15 dark green and 2 red bars represent genes supported by Bonferroni correction for multiple testing), based on the distribution of the number of genes encoding corresponding enzymes among the following species: *Homo sapiens*, *Mus musculus*, *Gallus gallus* and *Tetraodon nigroviridis* represent vertebrates, *Drosophila melanogaster*, *Apis mellifera*, *Anopheles gambiae* represent arthropods.

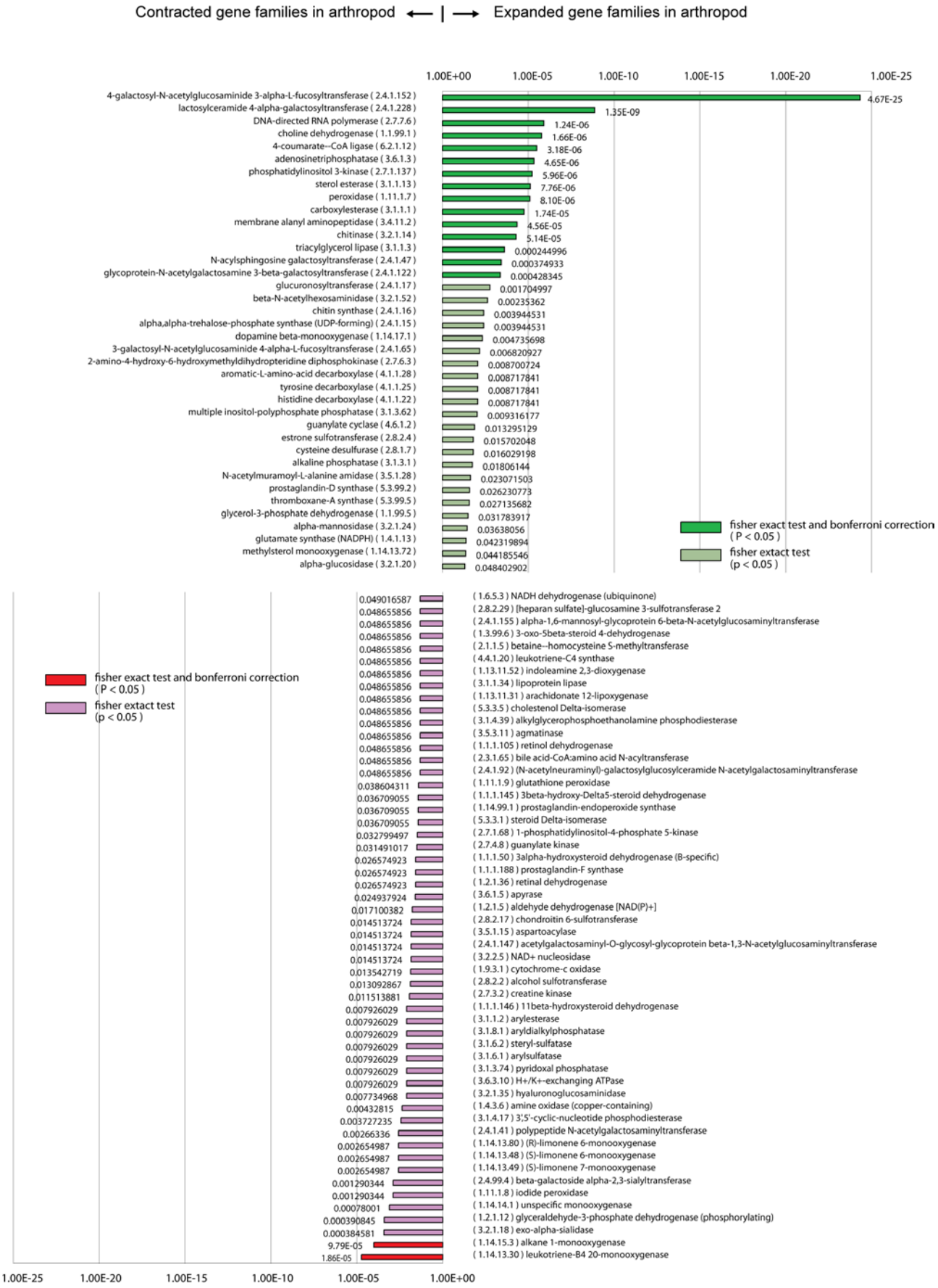

Contracted gene families in arthropod ⟵ | ⟶ Expanded gene families in arthropod
1.00E+00
1.00E-05
1.00E-10
1.00E-15
1.00E-20
1.00E-25
4-galactosyl-N-acetylglucosaminide 3-alpha-L-fucosyltransferase (2.4.1.152) 4.67E-25
lactosylceramide 4-alpha-galactosyltransferase (2.4.1.228) 1.35E-09
DNA-directed RNA polymerase (2.7.7.6) 1.24E-06
choline dehydrogenase (1.1.99.1) 1.66E-06
4-coumarate--CoA ligase (6.2.1.12) 3.18E-06
adenosinetriphosphatase (3.6.1.3) 4.65E-06
phosphatidylinositol 3-kinase (2.7.1.137) 5.96E-06
sterol esterase (3.1.1.13) 7.76E-06
peroxidase (1.11.1.7) 8.10E-06
carboxylesterase (3.1.1.1) 1.74E-05
membrane alanyl aminopeptidase (3.4.11.2) 4.56E-05
chitinase (3.2.1.14) 5.14E-05
triacylglycerol lipase (3.1.1.3) 0.000244996
N-acylsphingosine galactosyltransferase (2.4.1.47) 0.000374933
glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase (2.4.1.122) 0.000428345
glucuronosyltransferase (2.4.1.17) 0.001704997
beta-N-acetylhexosaminidase (3.2.1.52) 0.00235362
chitin synthase (2.4.1.16) 0.003944531
alpha,alpha-trehalose-phosphate synthase (UDP-forming) (2.4.1.15) 0.003944531
dopamine beta-monooxygenase (1.14.17.1) 0.004735698
3-galactosyl-N-acetylglucosaminide 4-alpha-L-fucosyltransferase (2.4.1.65) 0.006820927
2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase (2.7.6.3) 0.008700724
aromatic-L-amino-acid decarboxylase (4.1.1.28) 0.008717841
tyrosine decarboxylase (4.1.1.25) 0.008717841
histidine decarboxylase (4.1.1.22) 0.008717841
multiple inositol-polyphosphate phosphatase (3.1.3.62) 0.009316177
guanylate cyclase (4.6.1.2) 0.013295129
estrone sulfotransferase (2.8.2.4) 0.015702048
cysteine desulfurase (2.8.1.7) 0.016029198
alkaline phosphatase (3.1.3.1) 0.01806144
N-acetylmuramoyl-L-alanine amidase (3.5.1.28) 0.023071503
prostaglandin-D synthase (5.3.99.2) 0.026230773
thromboxane-A synthase (5.3.99.5) 0.027135682
glycerol-3-phosphate dehydrogenase (1.1.99.5) 0.031783917
alpha-mannosidase (3.2.1.24) 0.03638056
glutamate synthase (NADPH) (1.4.1.13) 0.042319894
methylsterol monooxygenase (1.14.13.72) 0.044185546
alpha-glucosidase (3.2.1.20) 0.048402902
fisher exact test and bonferroni correction (P < 0.05)
fisher extact test (p < 0.05)
fisher exact test and bonferroni correction (P < 0.05)
fisher extact test (p < 0.05)
0.049016587 (1.6.5.3) NADH dehydrogenase (ubiquinone)
0.048655856 (2.8.2.29) [heparan sulfate]-glucosamine 3-sulfotransferase 2
0.048655856 (2.4.1.155) alpha-1,6-mannosyl-glycoprotein 6-beta-N-acetylglucosaminyltransferase
0.048655856 (1.3.99.6) 3-oxo-5beta-steroid 4-dehydrogenase
0.048655856 (2.1.1.5) betaine--homocysteine S-methyltransferase
0.048655856 (4.4.1.20) leukotriene-C4 synthase
0.048655856 (1.13.11.52) indoleamine 2,3-dioxygenase
0.048655856 (3.1.1.34) lipoprotein lipase
0.048655856 (1.13.11.31) arachidonate 12-lipoxygenase
0.048655856 (5.3.3.5) cholestenol Delta-isomerase
0.048655856 (3.1.4.39) alkylglycerophosphoethanolamine phosphodiesterase
0.048655856 (3.5.3.11) agmatinase
0.048655856 (1.1.1.105) retinol dehydrogenase
0.048655856 (2.3.1.65) bile acid-CoA:amino acid N-acyltransferase
0.048655856 (2.4.1.92) (N-acetylneuraminyl)-galactosylglucosylceramide N-acetylgalactosaminyltransferase
0.038604311 (1.11.1.9) glutathione peroxidase
0.036709055 (1.1.1.145) 3beta-hydroxy-Delta5-steroid dehydrogenase
0.036709055 (1.14.99.1) prostaglandin-endoperoxide synthase
0.036709055 (5.3.3.1) steroid Delta-isomerase
0.032799497 (2.7.1.68) 1-phosphatidylinositol-4-phosphate 5-kinase
0.031491017 (2.7.4.8) guanylate kinase
0.026574923 (1.1.1.50) 3alpha-hydroxysteroid dehydrogenase (B-specific)
0.026574923 (1.1.1.188) prostaglandin-F synthase
0.026574923 (1.2.1.36) retinal dehydrogenase
0.024937924 (3.6.1.5) apyrase
0.017100382 (1.2.1.5) aldehyde dehydrogenase [NAD(P)+]
0.014513724 (2.8.2.17) chondroitin 6-sulfotransferase
0.014513724 (3.5.1.15) aspartoacylase
0.014513724 (2.4.1.147) acetylgalactosaminyl-O-glycosyl-glycoprotein beta-1,3-N-acetylglucosaminyltransferase
0.014513724 (3.2.2.5) NAD+ nucleosidase
0.013542719 (1.9.3.1) cytochrome-c oxidase
0.013092867 (2.8.2.2) alcohol sulfotransferase
0.011513881 (2.7.3.2) creatine kinase
0.007926029 (1.1.1.146) 11beta-hydroxysteroid dehydrogenase
0.007926029 (3.1.1.2) arylesterase
0.007926029 (3.1.8.1) aryldialkylphosphatase
0.007926029 (3.1.6.2) steryl-sulfatase
0.007926029 (3.1.6.1) arylsulfatase
0.007926029 (3.1.3.74) pyridoxal phosphatase
0.007926029 (3.6.3.10) H+/K+-exchanging ATPase
0.007734968 (3.2.1.35) hyaluronoglucosaminidase
0.00432815 (1.4.3.6) amine oxidase (copper-containing)
0.003727235 (3.1.4.17) 3',5'-cyclic-nucleotide phosphodiesterase
0.00266336 (2.4.1.41) polypeptide N-acetylgalactosaminyltransferase
0.002654987 (1.14.13.80) (R)-limonene 6-monooxygenase
0.002654987 (1.14.13.48) (S)-limonene 6-monooxygenase
0.002654987 (1.14.13.49) (S)-limonene 7-monooxygenase
0.001290344 (2.4.99.4) beta-galactoside alpha-2,3-sialyltransferase
0.001290344 (1.11.1.8) iodide peroxidase
0.00078001 (1.14.14.1) unspecific monooxygenase
0.000390845 (1.2.1.12) glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)
0.000384581 (3.2.1.18) exo-alpha-sialidase
9.79E-05 (1.14.15.3) alkane 1-monooxygenase
1.86E-05 (1.14.13.30) leukotriene-B4 20-monooxygenase
1.00E-25
1.00E-20
1.00E-15
1.00E-10
1.00E-05
1.00E+00

Figure S31. Expanded metabolic genes in the *Daphnia pulex* genome compared to other arthropods and vertebrates. Thirty-two enzymes are supported by the Fisher exact test (dark green bars represent 20 genes supported by Bonferroni correction for multiple testing), based on the distribution of the number of genes encoding corresponding enzymes among the following species: *Homo sapiens*, *Mus musculus*, *Gallus gallus* and *Tetraodon nigroviridis* represent vertebrates, *Drosophila melanogaster*, *Apis mellifera*, *Anopheles gambiae* represent arthropods.

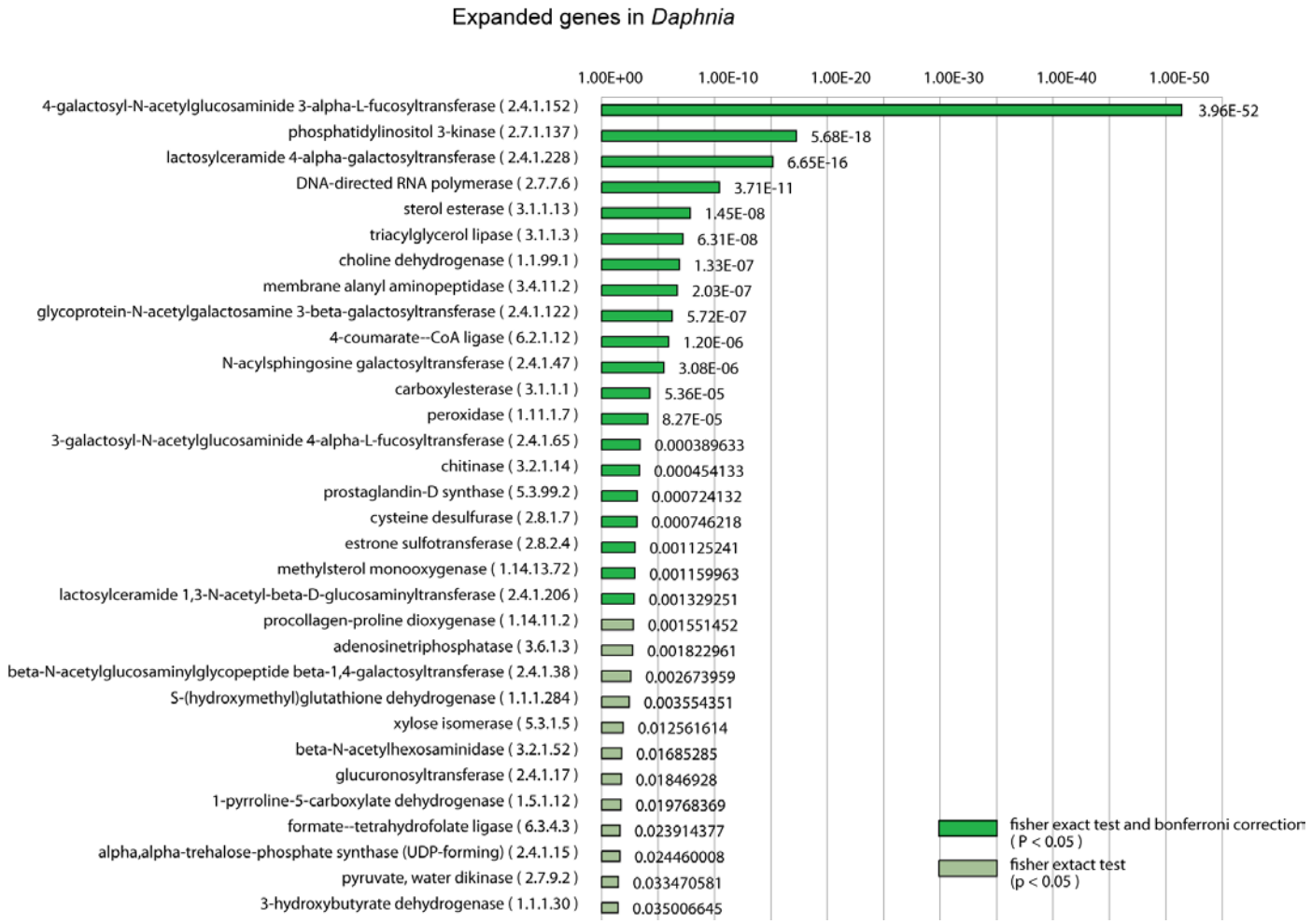


Figure S32. Distribution of the number of amplified genes with interactions, derived from 1,000 randomized metabolic networks. The horizontal axis represents the number of interacting genes with the vertical line at p = 0.03, and the vertical axis represents the frequency from sampling 1,000 randomized metabolic networks. Nineteen amplified genes are observed in the real network as having interactions, which is significantly higher than in randomized networks.

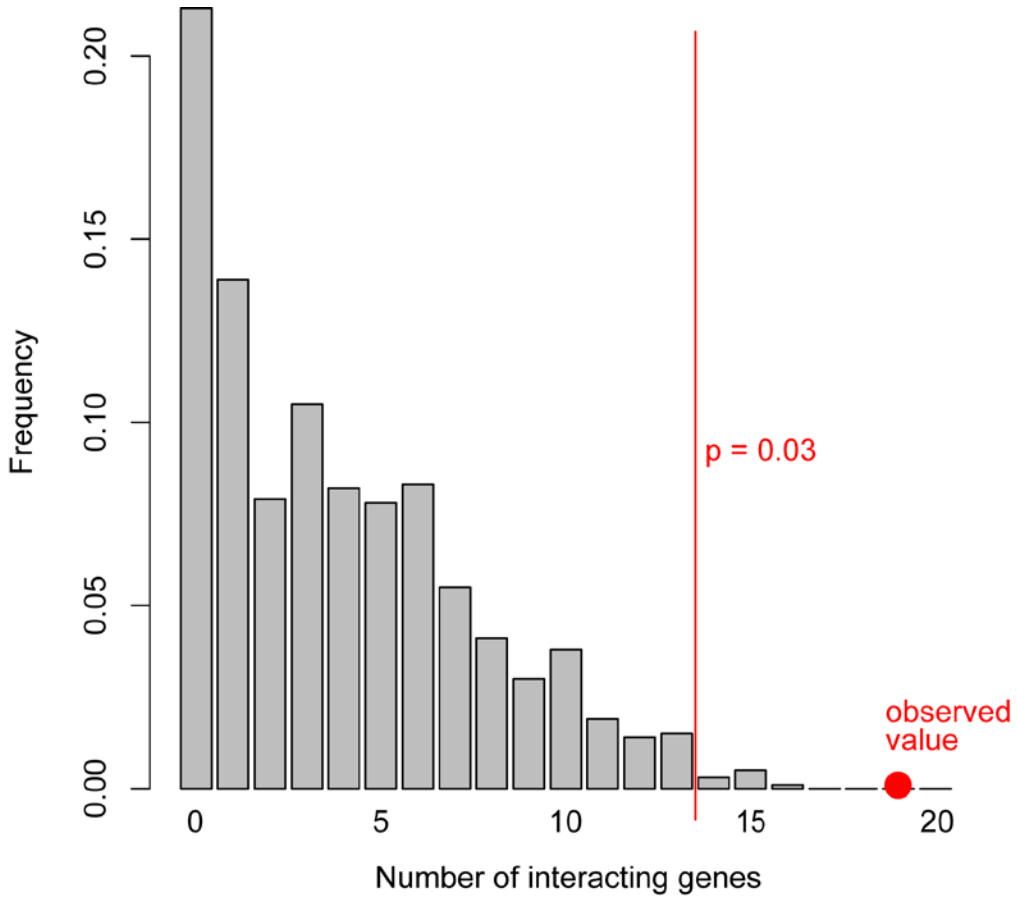


Figure S33. Phylogenetic relationships of members of the three expanded gene families of the *Daphnia pulex* glycosphingolipid biosynthesis neo-lactoseries pathway of metabolism (KEGG map000602). Phylogenetic trees are constructed by the maximum likelihood method using the Phylip ProML algorithm [S109] with corrected distances by the Jones-Taylor-Thornton model of molecular evolution [S110], using aligned amino acid sequences by MUSCLE [S58](Tables S45-48). Orthologs from the *Tribolium castaneum* (labeled blue) and *Ixodes scapularis* (labeled orange) genome sequences are included to bracket the *Daphnia pulex* paralogs.

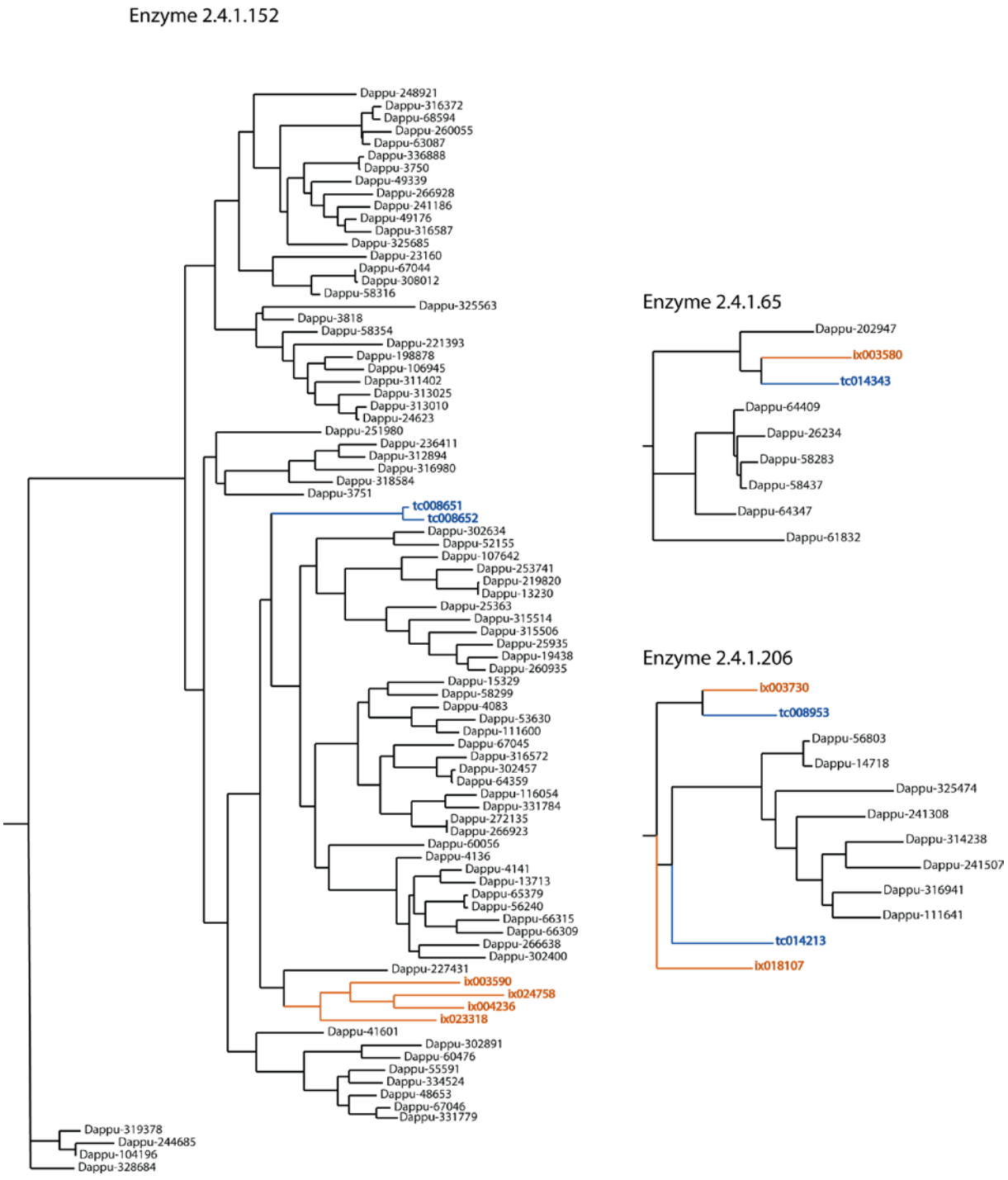


Figure S34. Differential expression (DE) pattern correlations among the *Daphnia pulex* gene-members of three lineage-specific gene family expansions from microarray experiments. The three enzymes are known to interact within the glycosphingolipid biosynthesis neo-lactoseries metabolic pathway of other model species. Correlations are plotted as a function of their pair-wise genetic divergence at silent sites (K_s). **Enzyme names:** 2.4.1.152, Alpha-1,3-fucosyltransferase C; 2.4.1.206, Beta-1,3-galactosyltransferase 5; 2.4.1.65, glycosyltransferase.

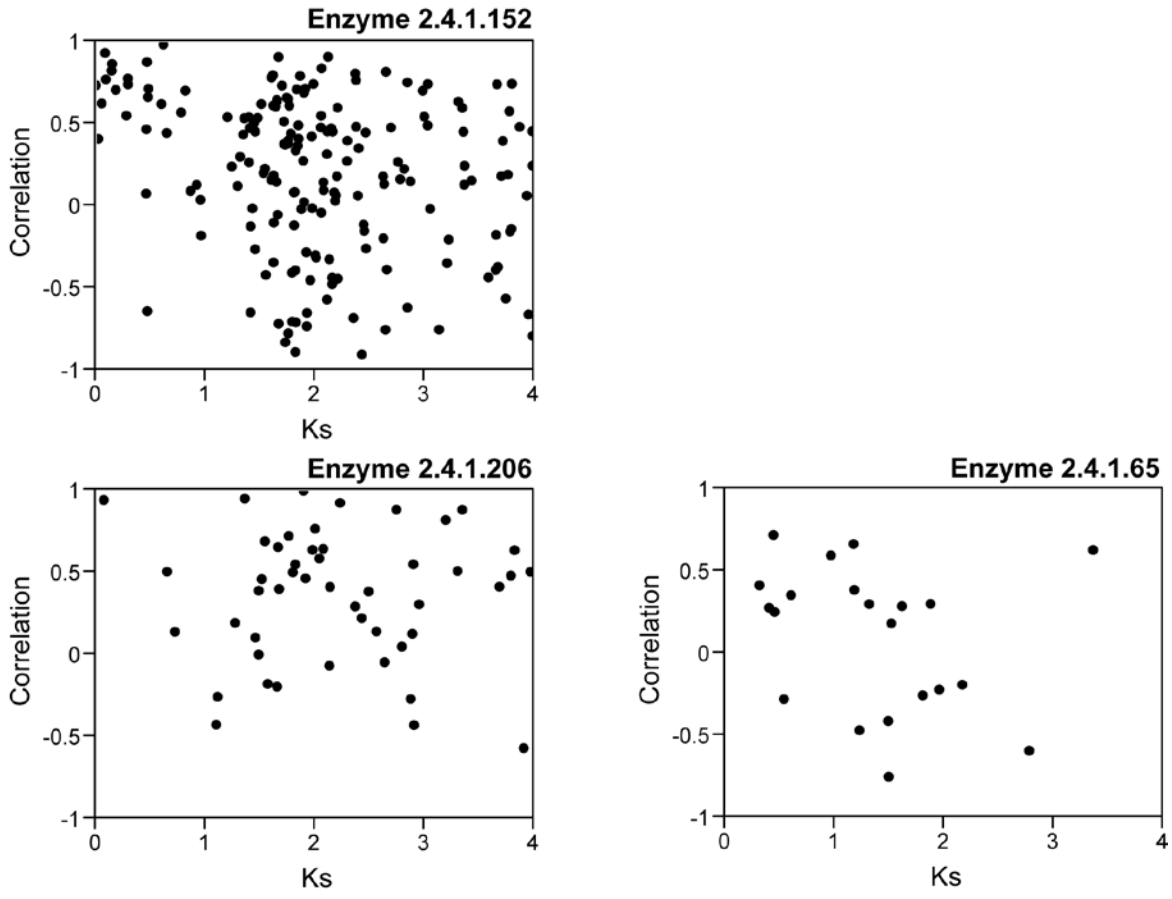


Figure S35. The phylogeny of duplicated fucosyltransferase genes (Enzyme 2.4.1.152) compared to their differential expression (DE) profiles across 8 experimental conditions (A-H) on microarrays. Gene phylogeny is identical to the panel in Figure S34. Internal nodes labeled blue are clades containing genes with average genetic distance between 0.4 and 0.5. Heat map on the left shows results from the hierarchical clustering by un-centered expression correlation of 79 genes from the expanded fucosyltransferase family plus 8 genes from the expanded glycosyl transferases family (enzyme 2.4.1.65 labeled by filled circles). Red designates up-regulation against the reference condition. Green designates down-regulation against the reference condition. Dark shades denote no change in gene expression. The two enzymes are required for biochemical reactions of glycosphingolipid biosynthesis. Subclusters labeled 1-7 contain at least one 2.4.1.65 gene. All 2.4.1.152 genes are grouped into one of these subclusters, except for Dappu-58299, Dappu-52155, Dappu-48653 and Dappu-248921. Lines are colored based on the membership of genes within clades stemming from marked nodes of the protein phylogeny.

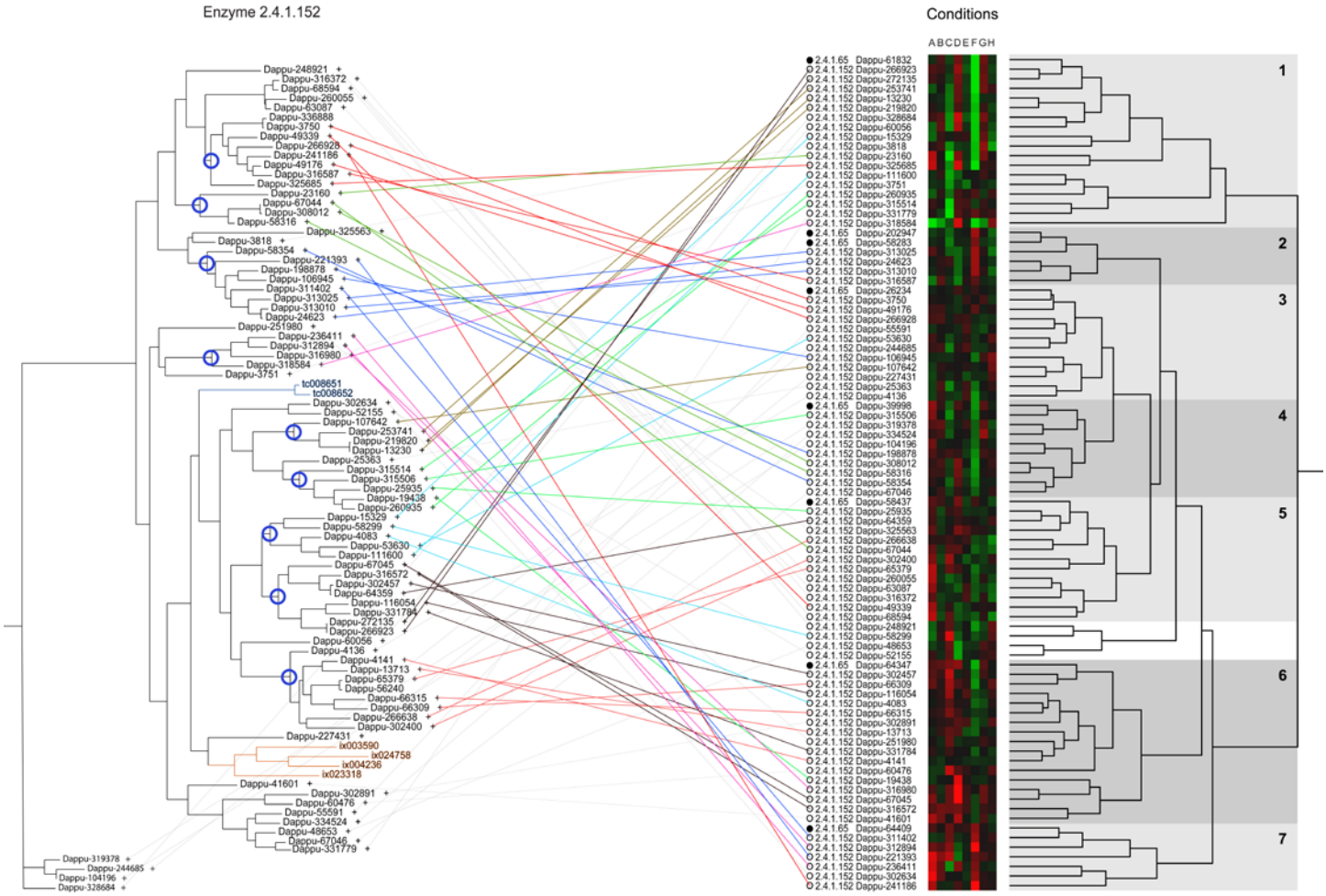
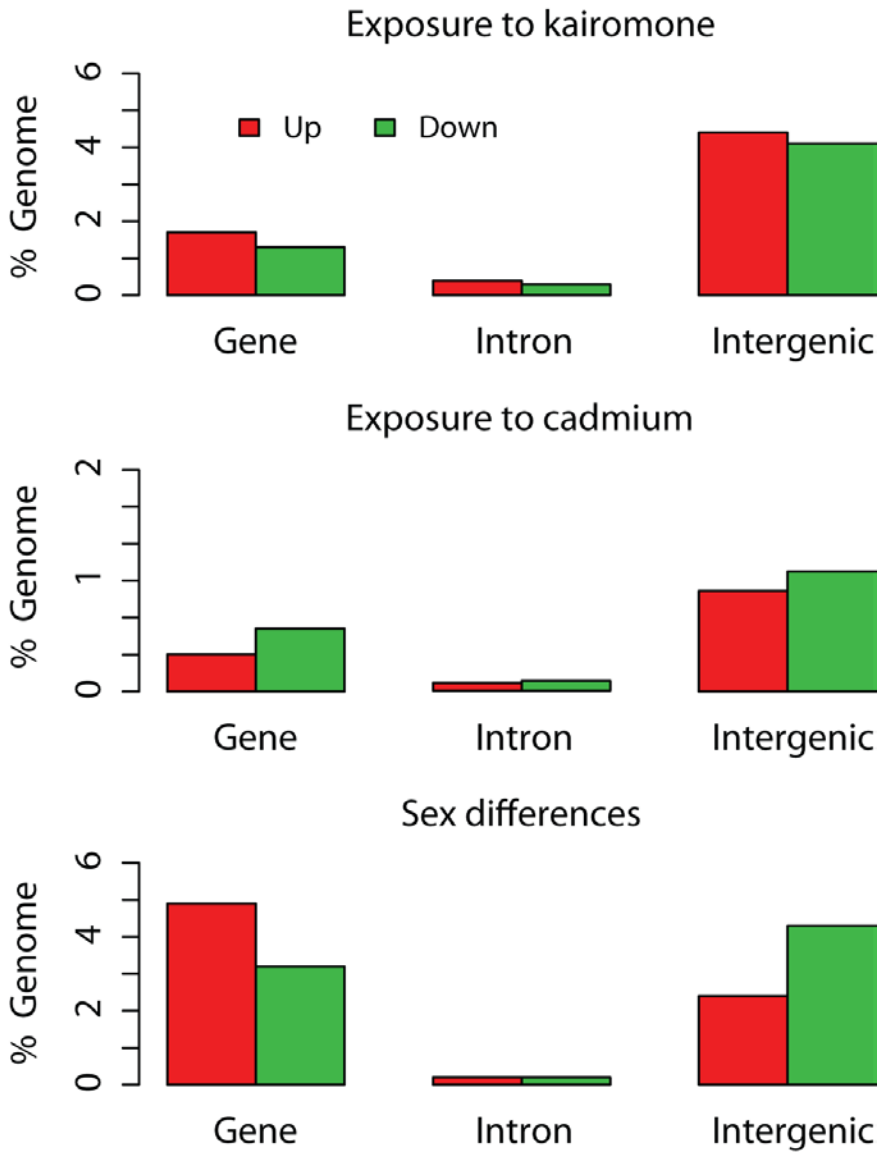


Figure S36. Differential transcription of the genome from *D. pulex* exposed to kairomone produced by the larval dipteran predator *Chaoborus* (biotic challenge), from *D. pulex* exposed to cadmium (abiotic challenge) and from male and females (standard conditions) measured by genome-wide tiling path microarray experiments. Differential transcription is twice as pronounced in genomic regions that are currently void of gene models (Intergenic) compared to regions with annotated genes when *D. pulex* are exposed to ecological conditions.



SUPPLEMENTARY TABLES

A. Introduction

Table S1. Open-source web-portals for *Daphnia pulex* genome data, analysis results and bioinformatic tools.

Daphnia informatics	URL address	Citation
wFleaBase	http://wFleaBase.org/	[S153]
JGI Genome Portal	http://www.jgi.doe.gov/Daphnia/	[S154]
PASA Database	http://wfleabase.org/genome/Daphnia_pulex/current/pasa/	[S21, S164]
ESTPiper	https://estpiper.cgb.indiana.edu/	[S26]
Superfamily	http://supfam.org/	[S83]
Cado	http://omics.informatics.indiana.edu/lab/CADO/precalculated/DpulInterPro/	[S165]
OrthoDB	http://cegg.unige.ch/orthodb	[S61]
miROrtho	http://cegg.unige.ch/mirortho/	[S48]
DGC Web Portal	http://daphnia.cgb.indiana.edu/	[S166]
Scaffold Dotplot	http://cancer.informatics.indiana.edu/cgi-bin/jeochoi/daphnia/tandemduplicategene/index.cgi	[S167]
MGEScan-LTR	http://darwin.informatics.indiana.edu/cgi-bin/evolution/daphnia_ltr.pl/	[S49]
DGC Wiki Portal	https://wiki.cgb.indiana.edu/display/DGC/Home	[S166]
NIH Model Organisms	http://www.nih.gov/science/models/	[S168]
NCBI UniGene	http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=6669/	[S169]
euGenes Arthropods	http://arthropods.eugenes.org/arthropods/	[S54]
Companion papers for the genome sequence	http://www.biomedcentral.com/series/Daphnia.	[S39]

B. Genome Sequence, Assembly and Chromosomes

Table S2. Summary of the *Daphnia pulex* genome assemblies using three assemblers. The official assembly for the current annotation is JAZZ, where numbers in parentheses are for the scaffolds and contigs from the nuclear genome. Other numbers refer to the full sequence data. The ARACHNE and PCAP assemblies are used to validate JAZZ.

	JAZZ	ARACHNE	PCAP
Number of reads	2,711,298	2,724,768	2,615,317
Number of reads placed	1,645,566 (1,554,564)	1,401,492	1,968,495
Length of reads placed (bp)	1,199,451,926	1,188,616,421	1,688,271,557
Number of scaffolds	26,848 (5,191)	23,643	61,858
Length of scaffolds (bp)	256,659,416 (197,261,574)	395,871,249	262,945,580
Length of largest scaffold (bp)	4,193,030 (4,193,030)	2,075,369	1,945,001
Avg. Length of scaffolds (bp)	9,660 (38,001)	16,743	4,250
Length of N50 scaffold (bp)	318,519 (642,089)	40,486	92,912
Number of N50 scaffold	142 (75)	1,734	376
Number of contigs	44,403 (19,008)	80,844	74,521
Length of contigs (bp)	186,524,647 (158,634,814)	209,098,385	239,506,399
Length of largest contig (bp)	528,830 (528,830)	144,860	302,603
Avg. Length of contigs (bp)	4,201 (8,346)	2,586	3,213
Length of N50 contig (bp)	1,170 (831)	14,037	14,037
Number of N50 contig	34,096 (49,250)	3,158	3,158
Number of gaps	17,555 (13,817)	57,201	12,663
Length of gaps (bp)	70,117,214 (38,612,943)	186,772,864	23,439,181

Table S3. Analysis of shotgun reads from TCO and TRO derived libraries. Two genomic read libraries from TRO (ANIT,ANIS) and three libraries from TCO (AZSN, AZWZ, AZSH) were aligned to the TCO assembly using the BLAST algorithm and a strict filter was used to identify potential scaffold bridging reads (see SOM). Approximate insert size for each library is shown in parenthesis. Each row of the table between "Starting Reads" and "Different Scaffolds" represents a criterion on which the alignment failed to pass through the filter, with the number of failed reads shown for each read library.

	TRO			TCO			
gDNA Library	ANIT (8kb)	ANIS (3kb)	Total	AZSN (35kb)	AZWZ (7kb)	AZSH (3kb)	Total
Starting Reads	151,381	137,603	288,984	201,262	1,202,060	1,139,438	2,542,760
No Pair in Library	3,322	9,427	12,749	15,869	177,097	55,295	248,261
No Blast Hit	18,286	21,303	39,589	--	--	--	--
No pair	5,601	4,299	9,900	17,489	43,401	56,395	117,285
e-value not met	38,654	36,206	74,860	11,368	45,522	50,054	106,944
Multiple hits	64,088	47,494	111,582	120,882	719,482	771,530	1,611,894
Potential inversion	596	380	976	380	3,206	7,660	11,246
Different scaffolds	**1,920**	**1,520**	**3,440**	**1,828**	**12,760**	**14,922**	**29,510**
Candidate Reads	18,914	16,974	35,888	33,446	200,592	183,582	417,620

TRO = reads from The Rejected One
TCO = reads from The Chosen One

Table S4. Putative super-scaffolds based on focused paired-end read analysis. Super-scaffolds are ordered by sequence length, excluding gaps. Scaffolds which clustered with a super-scaffold but could not be unambiguously placed are listed as Additional Scaffolds. Each bridged scaffold is listed in order of assembly with the number of bridging reads in the column to the right. Scaffolds anchored on the genetic map (Table S5) are indicated by an "lg" followed by the genetic map chromosome number. Scaffolds that must be reverse complemented in order linked in the proper orientation are marked with an "rc". In a few cases, no direct bridging reads were found between two scaffolds but flanking scaffolds were found to be linked. These scaffolds are indicated with an “na” in their Bridging Scaffold column followed by the number of reads that bridge the flanking scaffolds.

Please download Tables S4 and S11-14 from:

http://wfleabase.org/release1/current_release/supplement/

Table S5. Scaffolds genetically mapped to chromosomes. Markers and map IDs are described in [S8, S170]. Chromosomes are numbered starting from the largest map distance to the smallest.

Scaffold ID	Start position	End position	Marker ID	Map distance (cM)	Map ID
Chromosome 1					
scaffold 130	265403	264989	Dp840	.	d115
scaffold 42	674271	674164	Dp40	3.9	d039
scaffold 42	766448	766130	Dp564	.	d076
scaffold 130	317751	317116	Dp1354	8.4	d167
scaffold 42	132315	132609	Dp589	15.8	d098
scaffold 42	112093	112556	Dp1290	17.9	d170
scaffold 53	685669	685917	Dp199	22.8	d048
scaffold 53	515514	515330	Dp571	26.1	d096
scaffold 53	289336	289000	Dp884	30.4	d114
scaffold 53	192875	193436	Dp1073	32.5	d181
scaffold 207	91612	91361	Dp553	42.6	d134
scaffold 106	175949	175582	Dp802	57	d101
scaffold 211	98831	99327	Dp1495	72.3	d063
scaffold 3	3212918	3212697	Dp1189	77.5	d193
scaffold 3	3015577	3015192	Dp300	86.5	d053
scaffold 3	2484047	2483839	Dp729	98.2	d103
scaffold 3	2318013	2318270	Dp1266	102.6	d174
scaffold 3	2276376	2275958	Dp1368	104.3	d163
scaffold 3	2064769	2064530	Dp149	109.8	d001
scaffold 3	1923990	1923837	Dp754	.	d130
scaffold 3	1384162	1383978	Dp655	122.4	d091
scaffold 3	739228	739030	Dp74	148.6	d007
scaffold 61	198861	198594	Dp1155	168.6	d148
scaffold 61	211843	212169	Dp1195	.	d188
scaffold 80	268005	267759	Dp48	182	d009
scaffold 53	150886	150445	Dp957	.	d138
Chromosome 2					
scaffold 5	2127458	2127138	Dp1048	.	d197
scaffold 70	111417	111546	Dp725	2.4	d102
scaffold 86	425898	426113	Dp848	7.5	d112
scaffold 58	394484	394271	Dp967	22	d140
scaffold 27	318072	318300	Dp785	25.6	d095
scaffold 159	63421	63621	Dp1497	25.8	d123
scaffold 84	320005	319800	Dp339	25.8	d016
scaffold 84	323106	323340	Dp742	25.8	d104
scaffold 63	450291	450076	Dp1491	28.3	d050
scaffold 63	480969	480790	Dp389	29.9	d074
scaffold 134	191172	191603	Dp1494	70.5	d044
scaffold 80	364906	364846	Dp969	82.5	d195
scaffold 30	735909	736343	Dp1005	89.3	d196
scaffold 30	741780	741652	Dp637	89.3	d120
scaffold 30	883779	883988	Dp28	93.1	d004

scaffold 237	42137	42598	Dp821	99.2	d109
scaffold 112	95327	95763	Dp325	.	d070
scaffold 71	331251	330699	Dp1363	99.4	d175
scaffold 1	4020601	4020755	Dp321	101	d002
scaffold 1	3532645	3532515	Dp395	107.4	d047
scaffold 1	3510933	3510772	Dp147	.	d015
scaffold 1	3257753	3258007	Dp1056	116.8	d183
scaffold 1	2755698	2755478	Dp224	128.7	d069
scaffold 1	2459188	2459642	Dp557	137.8	d079
scaffold 1	2184113	2183854	Dp117	145.6	d124
scaffold 1	1891813	1892387	Dp1346	156.4	d162
Chromosome 3					
scaffold 219	85465	85231	Dp1498	.	d008
scaffold 16	499003	499212	Dp308	11.4	d067
scaffold 26	674530	674325	Dp71	.	d010
scaffold 87	19282	19769	Dp1490	.	d064
scaffold 26	185444	185604	Dp572	30.9	d097
scaffold 21	598039	598442	Dp1058	31.7	d169
scaffold 21	813362	813165	Dp581	.	d082
scaffold 21	835598	835104	Dp1276	35.6	d177
scaffold 21	1051430	1051139	Dp24	39.5	d003
scaffold 21	1248646	1248825	Dp50	41.5	d122
scaffold 62	124936	124649	Dp616	62.4	d078
scaffold 62	51713	51836	Dp115	66.4	d054
scaffold 97	406130	406310	Dp337	76.2	d019
scaffold 2	3187522	3187300	Dp137	.	d041
scaffold 2	3138447	3137984	Dp770	.	d094
scaffold 2	2984218	2983993	Dp144	93.3	d049
scaffold 2	2984384	2984631	Dp895	93.3	d132
scaffold 2	2184048	2184167	Dp111	.	d062
scaffold 128	62140	62654	Dp196	111.9	d059
scaffold 32	851822	852212	Dp1492	111.9	d066
scaffold 2	1834370	1834050	Dp998	116.8	d136
scaffold 2	1851473	1851638	Dp813	116.8	d108
scaffold 2	1247523	1247096	Dp530	127.1	d075
scaffold 2	465349	465115	Dp1041	147.2	d147
Chromosome 4					
scaffold 31	806857	806709	Dp675	.	d089
scaffold 31	224972	225256	Dp311	0.9	d071
scaffold 31	224972	225256	Dp311	0.9	d071
scaffold 31	647755	648063	Dp1311	0.9	d156
scaffold 28	18864	18481	Dp1372	5.9	d168
scaffold 2784	447	465	Dp430	.	d029
scaffold 110	104935	104741	Dp605	.	d081
scaffold 11	725719	726062	Dp687	8	d084
scaffold 11	343688	343786	Dp878	15.3	d116
scaffold 11	203644	203416	Dp924	19.6	d139

scaffold 86	299694	299675	Dp1376	26.8	d179
scaffold 8	2259957	2260245	Dp78	33	d057
scaffold 8	1927865	1927328	Dp1185	36.3	d155
scaffold 8	1410937	1411116	Dp779	36.6	d105
scaffold 47	238533	238708	Dp295	.	d021
scaffold 8	213952	214644	Dp830	80.5	d106
scaffold 43	380284	379974	Dp143	114.6	d018
scaffold 43	254164	254353	Dp1409	120	d180
scaffold 43	119984	120207	Dp1396	123.1	d172
scaffold 163	158324	158011	Dp1148	143.4	d143
Chromosome 5					
scaffold 89	595719	595282	Dp838	.	d126
scaffold 89	542785	542357	Dp1123	3.1	d160
scaffold 89	311220	310743	Dp1262	7.1	d164
scaffold 39	1016773	1016558	Dp91	29.6	d013
scaffold 39	589286	589170	Dp240	49.3	d031
scaffold 39	427971	428255	Dp231	54.1	d024
scaffold 39	379458	379327	Dp208	57.6	d042
scaffold 39	284948	284755	Dp319	60.9	d030
scaffold 39	5470	5195	Dp721	69.3	d093
scaffold 12	1025201	1024971	Dp21	95.5	d055
scaffold 12	368224	368109	Dp648	96	d087
scaffold 15	212509	212011	Dp632	96	d119
scaffold 38	309466	309811	Dp775	113.5	d100
Chromosome 6					
scaffold 43	874249	874269	Dp907	.	d111
scaffold 47	77609	77800	Dp1232	22.9	d144
scaffold 47	88108	87872	Dp170	22.9	d020
scaffold 47	76063	76223	Dp815	.	d125
scaffold 191	128146	127729	Dp642	38.4	d085
scaffold 32	244697	244205	Dp298	48.6	d025
scaffold 32	253637	253489	Dp126	49.8	d014
scaffold 32	295131	295288	Dp475	.	d035
scaffold 32	376888	377193	Dp1040	52.9	d142
scaffold 32	471100	471385	Dp985	55.3	d135
scaffold 183	105222	105353	Dp1399	60	d190
scaffold 32	749728	750056	Dp146	60.7	d012
scaffold 32	772534	772153	Dp361	60.7	d073
scaffold 32	1085664	1085897	Dp385	61.6	d028
scaffold 57	715932	715294	Dp1327	63.4	d152
scaffold 251	4363	4584	Dp1350	84.9	d153
scaffold 28	807177	807482	Dp1238	107.2	d151
Chromosome 7					
scaffold 46	900788	901047	Dp156	.	d027
scaffold 4	2237553	2237059	Dp112	21.2	d058
scaffold 184	39118	39342	Dp786	31.6	d133
scaffold 93	181025	180507	Dp1328	31.6	d157

scaffold 91	402113	401836	Dp1347	45	d189
scaffold 91	391760	391276	Dp1391	46.6	d191
scaffold 46	483793	484121	Dp1300	53.3	d166
scaffold 46	456561	456798	Dp867	54.5	d107
scaffold 22	102030	102641	Dp1489	80.4	d040
Chromosome 8					
scaffold 7	1979117	1979476	Dp53	.	d068
scaffold 7	2037319	2037409	Dp142	4.3	d065
scaffold 83	464580	465066	Dp559	23.9	d077
scaffold 136	226201	225971	Dp887	46.5	d117
scaffold 151	287252	287019	Dp1404	.	d165
scaffold 77	181266	181816	Dp1160	.	d150
scaffold 77	57562	57923	Dp1493	.	d045
scaffold 199	40471	40664	Dp883	50.5	d113
scaffold 32	551409	551789	Dp1351	75.2	d192
scaffold 2	11597	11824	Dp1485	76.1	d121
Chromosome 9					
scaffold 9	761358	761163	Dp1278	.	d178
scaffold 9	848851	848517	Dp1309	0.9	d171
scaffold 9	1082325	1082581	Dp1325	6.2	d145
scaffold 9	1369547	1369704	Dp621	20	d118
scaffold 9	2143217	2143047	Dp330	48.1	d043
scaffold 99	316482	316596	Dp660	49.8	d088
scaffold 13	1397943	1397633	Dp609	64.9	d099
scaffold 13	1333322	1333135	Dp1236	65.2	d149
scaffold 13	1074377	1074598	Dp123	72.4	d011
Chromosome 10					
scaffold 103	332371	332765	Dp696	.	d127
scaffold 49	90586	90309	Dp460	1.9	d023
scaffold 49	580950	580271	Dp304	2.3	d034
scaffold 49	611970	612305	Dp463	5.1	d005
scaffold 100	375967	375623	Dp1496	17.5	d072
scaffold 29	225287	225013	Dp1302	26.9	d161
scaffold 29	424770	425099	Dp1057	35.4	d186
scaffold 17	914491	914609	Dp641	61.3	d083
Chromosome 11					
scaffold 24	541082	540777	Dp808	.	d092
scaffold 24	272325	272150	Dp70	16.5	d006
scaffold 24	141703	142351	Dp1112	22	d173
scaffold 111	208013	207484	Dp693	55.4	d086
Chromosome 12					
scaffold 5	123699	123547	Dp726	.	d128
scaffold 5	134282	134058	Dp936	.	d137
scaffold 5	126817	126598	Dp1080	0.5	d187
scaffold 5	117840	118469	Dp1144	1	d182
scaffold 5	109480	109307	Dp1079	6.9	d184

Table S6. Pair-wise comparison of genome assemblies by using different assemblers. The JAZZ contigs were matched with the contigs generated by Arachne and the contigs generated by PCAP using genome sequence alignment program, MUMmer [S10]. C_n and C_l represent the total number and total length of all matched contigs in corresponding assembly, respectively; U_l represent the total lengths of uniquely matched contigs in both assemblies. We also applied the two additional criteria to filter the MUMmer matches (denoted as regular and stringent; see the texts for details) and show the comparison results below. In general, >95% JAZZ contigs are consistent with Arachne and PCAP contigs, indicating JAZZ assembly that we used for the analysis in this manuscript has satisfactory quality.

MUMmer Filtering	Assemblies	(C_n) Total no. of matched contigs	(C_l) Total length of matched contigs (Mb)	Fraction of (C_l) over total length of contigs	(U_l) Total length of uniquely matched contigs (Mb)	Fraction of (U_l) over total length of matched contigs (C_l)
All	Arachne vs. JAZZ	78,569	205.9	0.98	202.4	0.98
		33,734	175.6	0.94	170	0.97
	PCAP vs. JAZZ	64,973	228	0.95	221.2	0.97
		40,682	182.5	0.98	180	0.99
Regular	Arachne vs. JAZZ	52,176	172.6	0.83	164.3	0.95
		26,814	168.1	0.90	156	0.93
	PCAP vs. JAZZ	34,033	189.5	0.79	174.2	0.92
		34,960	175.9	0.94	170.2	0.97
Stringent	Arachne vs. JAZZ	35,958	151.8	0.73	145	0.96
		17,087	157.6	0.84	143.1	0.9
	PCAP vs. JAZZ	19,851	172.2	0.72	158.9	0.92
		19,792	160.9	0.86	155.4	0.97

Table S7. GAV (Genome Assembly Validator) is a machine learning approach that combines multiple evidences to detect putatively mis-assembled regions in genome assemblies [S11]. The features used in GAV include read and clone coverage, clone length statistics, and repeat content in the region. We used the regions in the assemblies that are supported by EST sequences as positive samples for the training (shown in **A**) The statistics of detected mis-assembled regions by GAV are shown in **B**.

A.

Criteria	Class	No. scaffolds	No. contigs	No. regions	No. bases
Regular	Correct	711	2,905	117,211	26,634,131
	Mis-assembly	500	1,841	13,029	369,702
	Total	940	4,746	130,240	27,003,833
Stringent	Correct	710	2,894	116,714	26,608,272
	Mis-assembly	474	1,642	10,232	334,507
	Total	920	4,536	126,946	26,942,779

B.

Criteria		Model	Correct assembly	Mis-assembly	Total
Regular	No. scaffolds	512	1,799	771	1,862
	No. contigs	2,502	7,653	2,823	7,887
	No. blocks	48,009	208,363	5,906	262,278
Stringent	No. scaffolds	512	1,817	653	1,862
	No. contigs	2,496	7,710	2,097	7,887
	No. blocks	47,870	210,742	3,666	262,278

Table S8. Chromosome size measurements. Chromosomes are numbered starting from the largest in length to the smallest, and are not necessarily congruent with the chromosome numbers for the genetic map. Heterochromatic regions are measured as the proportion of total chromosome length in DAPI and G banding stained regions.

Chromosome	Area (square µm)	Length (µm)
1	3.67 – 6.18	5.66 – 6.59
2	1.8 – 3.66	2.3 – 3.37
3	1.74 – 2.11	2.09 – 2.33
4	1.4 – 1.96	1.93 – 2.07
5	1.38 – 1.58	1.77 – 2.00
6	1.33 – 1.41	1.71 – 1.79
7	1.27 – 1.3	1.56 – 1.67
8	1.21 – 1.28	1.38 – 1.46
9	0.86 – 1.06	1.16 – 1.31
10	0.58 – 0.94	0.9 – 1.28
11	0.53 – 0.83	0.81 – 1.28
12	0.44 – 0.86	0.71 – 1.28
Total	16.21 – 23.17	21.98 – 26.43
Heterochromatic region	4.2 – 5.85	(25% of total area)

C. Largest Gene Inventory

Table S9. Results from the automated gene annotation procedures. Gnomon, Fgenesh++ and SNAP are *ab-initio* predictors, but also using additional EST and protein evidence. GeneWise maps known protein genes to the genome, and PASA maps ESTs into gene models. Many gene predictions were post-processed to extend models with EST evidence. Gene models were improved by manual annotation and by automated verification against EST assemblies using PASA. These improvements included UTR additions, internal rearrangements and refinements of intron-exon boundaries, and merging or splitting of gene models. The criteria for assigning gene models to the Chosen models (v1.1 frozen gene set) are described in the SOM.

Source of gene prediction	Chosen models	All models	Alternate transcripts modeled from EST data	Average protein length (AA)	Average exons per gene
Gnomon	7,717	37,329	137	323	4.7
PASA	4,059	11,845	1319	534	6.2
SNAP	7,364	41,310	na	306	3.9
Fgenesh++	5,863	34,193	na	384	3.7
GeneWise	3,328	29,488	na	na	4.8
EstExt	2,434	45,066	na	406	7.8
Manual	175	na	na	na	na
Total	30,940	--	0	325	4.6

Table S10. The *Daphnia pulex* cDNA libraries and EST sequencing effort. cDNA clones were sequenced from both ends. Clone diversity is calculated by dividing the # of EST clusters (assembled ESTs including clusters of 1) by the # of clones in clusters. This estimate is inflated, especially for non-normalized libraries, by ignoring clones containing organelle transcripts (6% to 45% of ESTs are mitochondrial, depending on library). By contrast, the normalized libraries typically contain between <1% and 10% organelle ESTs. Libraries were created from two isolates: TRO = The Rejected One; TCO = The Chosen One.

Library ID	Condition, Developmental Stage	# Sequenced Clones	# Nuclear ESTs	# EST Clusters†	# Clones in Clusters†
Non-normalized					
TRO-1	Hypoxia, adult	2,304	3,355	1,039	1,823
TRO-2	Hypoxia, juvenile	3,840	5,567	1,524	3,033
TRO-3	Low dose UV exposure, mixed	2,304	2,620	1,013	1,433
TRO-4§	High dose UV exposure, mixed	384	450	188	243
TRO-5	Unchallenged, juvenile	1,152	1,580	553	827
TRO-6	Low dose cadmium, mixed	2,688	4,048	1,209	2,170
TRO-7	Low dose arsenic, mixed	4,224	6,370	1,867	3,399
TRO-8	Low dose zinc, mixed	4,224	6,817	1,535	3,709
TRO-9	High dose mixed metals, mixed	4,608	7,185	1,863	3,770
TRO-10§	Unchallenged, mixed	384	390	159	232
TRO-11§	Unchallenged, mixed	384	405	167	238
TRO-12	Invertebrate (*Chaoborus*) predation, adult	4,608	6,542	2,034	3,511
TRO-13	Food starvation, juvenile	2,304	2,826	924	1,378
TRO-14	Food starvation, adult	2,304	2,684	860	1,291
TRO-15§	Microcystis fed, juvenile	384	307	150	175
TRO-16§	Microcystis fed, adult	384	368	164	208
TRO-17	Fish predation, juvenile	3,840	4,750	1,548	2,638
TRO-18§	Fish predation, adult	384	425	177	249
TRO-19§	Methyl Farnesoate hormone, juvenile	384	413	170	227
TRO-20	Methyl Farnesoate hormone, adult	3,840	4,833	1,323	2,604
Total		50,070	70,765		33,158
Library ID	Condition, Developmental Stage	# Sequenced Clones	# Nuclear ESTs	# Clusters†	# Clones in Clusters†
Normalized					
TRO-21	Unchallenged, mixed	5,376	8,962	3,413	4,762
TCO-1§	Females, juvenile	384	211	98	121
TCO-2	Females, adult	3,456	5,313	2,252	2,821
TCO-3	Males, adult	4,224	5,425	2,168	2,883
TCO-4	Low dose nickel, mixed	4,224	6,484	2,865	3,599
TCO-5	Low dose copper, mixed	4,224	6,852	2,963	3,685
TCO-6	Acid stress pH 6.0, mixed	3,840	6,626	2,870	3,514
TCO-7	High salinity, mixed	3,840	6,121	2,645	3,275
TCO-8	Fullerene nanoparticle, mixed	4,224	5,643	2,428	3,044

TCO-9	Bacterial infection, mixed	3,456	5,639	2,553	2,935
TCO-10	High dose mixed metals, mixed	3,840	4,398	2,030	2,452
TCO-11	Low dose mixed metals, mixed	3,456	5,407	2,447	2,967
TCO-12	Low dose monomethylarsenic III, mixed	4,224	6,274	2,768	3,387
TCO-13	Titanium dioxide nanoparticle, mixed	4,224	5,742	2,490	3,037
TCO-14	Microcystis fed, mixed	3,072	4,734	2,052	2,522
TCO-15	Calcium starvation, mixed	3,840	5,309	2,278	2,887
Total		59,904	89,140		47,891

§ Libraries failing stringent quality control checks and were therefore excluded from high throughput EST sequencing.
† These numbers are of clusters and clones of nuclear genes only.

Table S11. Observed homology and transcription evidence for v1.1 annotated gene set of the *Daphnia pulex* genome. Evidence columns include (1) homology found within the 8-fold coverage draft genome assembly for the distantly related *Daphnia magna* using BLAST searches with e-value cutoff set at 10^{-10}; (2) EST evidence when the degree of sequence identity is 90% and above; (3) homology bit scores from BLAST sequence similarity searches against the NCBI non-redundant (NR) protein database with e-value cutoff set at 10^{-5}; (4) evidence of transcription in tiling array experiments where transcriptionally active regions (TARs) and gene models overlap by 80% or more; (5) paralog IDs assigned by the OrthoMCL algorithm [S79, S80]. The gene location information includes strand (+/-), while the listed gene models are those summarized in Table S9. Alternative Gnomon model IDs are also provided. Summary of the results: 23,239 predicted genes only have evidence from homology to other proteomes; 18,451 genes only have evidence from EST and tiling array experiments; 27,090 have at least one line of evidence, including paralogs; 25,690 genes have at least one line of evidence, excluding paralogs. Only 4,040 genes have no comparative or empirical support.

Note: By requiring 80% overlap between detected TARs and gene models, 57,294 exons from 14,135 v1.1 genes are supported. In addition, we detected 10,125 TARs that overlap exons from 4,227 alternative gene models. Yet further, we count 68,033 remaining TARs that do not overlap with any predicted exons. Of these, 9,783 TARs are found inside genes and outside of predicted exons but within 500 bp of exons, and 9,620 intergenic TARs are directly neighboring predicted gene boundaries by 200 bp. These transcribed areas of the genome are most likely untranslated genic regions (UTRs) or model corrections. Finally, 48,630 TARs are unattached to existing gene models. By clustering unattached TARs in groups of 3 or more exon-like structures within 200 bases from each other, we detect 7,965 gene-like groupings. Most of these TAR-predicted loci (7,059) have an open reading frame greater than 40 amino-acids (see Table S12).

Please download Tables S4 and S11-14 from:

http://wfleabase.org/release1/current_release/supplement/

Table S12. Supporting evidence is found for 4,480 Transcriptional Active Regions (TAR)-predicted loci, despite their being overlooked by gene finding algorithms or their rejection from the v1.1 gene builds. Of the 7,965 gene-like TAR groupings, most (7,059) have open reading frames greater than 40 amino acids; 1,275 (16%) have EST support and 1,514 (19%) overlap with discarded Gnomon gene predictions, some containing erroneous stop-points in open reading frames. A search for protein homologs in the NCBI non-redundant database, at the 1×10^{-3} statistical cut-off value, uncovers matches for 171 TAR-predicted loci.

Please download Tables S4 and S11-14 from:

http://wfleabase.org/release1/current_release/supplement/

Table S13. List of identified proteins. Values in row “Protein ID probability” are calculated using Scaffold V. 02.01.00. **A)** Proteins identified in v1.1 gene catalog; **B)** Proteins identified among all predicted models, yet not included in the v1.1 set.

Please download Tables S4 and S11-14 from:

http://wfleabase.org/release1/current_release/supplement/

Table S14. List of identified peptides. Values in rows “Protein ID probability” and “Best peptide ID probability” are calculated using Scaffold V. 02.01.00. **A)** Peptides identified in v1.1 gene catalog; **B)** Peptides identified among all predicted models, yet not included in the v1.1 set.

Please download Tables S4 and S11-14 from:

http://wfleabase.org/release1/current_release/supplement/

Table S15. List of 716 genes conserved as single-copy othologs across eukaryotic genomes. The first 17 listed genes are missing from the v1.1 set of *Daphnia pulex* annotated gene list, yet all except two are either listed in this set or predicted by NCBI Gnomon gene models. Only two genes (KOG3086/CG8031 and KOG3499/CG18001) are missing from both sets. This analysis serves as a control for the assembly quality (2/716=0.3% missing). The *D. pulex* proteins were added to the clusters of orthologous genes of eukaryotes (KOGs), which were obtained by comparison of 7 genomes: *Homo sapiens*, the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, the dicot plant *Arabidopsis thaliana*, the ascomycete fungi *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, and the intracellular microsporidian parasite *Encephalitozoon cuniculi* [S56]. The *D. pulex* genes were assigned to the COGs using the "index ortholog" method [S171]. To compare with other genome assemblies, we measured the frequency of identifying orthologs of these same genes within the annotated genomes of 10 arthropods [S54]: *Aedes aegyptii; Anopheles gambiae; Culex pipiens*; *Drosophila pseudoobscura; Bombyx mori; Tribolium castaneum; Nasonia vitripennis; Pediculus humanus; Acyrthosiphon pisum; Ixodes scapularis*. The number of missing genes range from 1 to 9, placing the *D. pulex* genome on par among the better arthropod genome sequence assemblies for identifying these single-copy orthologs.

KOG ID	*Daphnia pulex* gene	*Daphnia pulex* gene	*Drosophila melanogaster* gene
KOG3086	NULL	NULL	CG8031
KOG3499	NULL	NULL	CG18001
KOG0333	NULL	NCBI_GNO_286924	CG10333
KOG0923	NULL	NCBI_GNO_140324	CG10689
KOG0924	NULL	NCBI_GNO_630594	CG32604
KOG0998	NULL	NCBI_GNO_280604	CG16932
KOG1119	NULL	NCBI_GNO_156434	CG13623
KOG1643	NULL	NCBI_GNO_9034	CG2171
KOG1748	NULL	NCBI_GNO_1452013	CG9160
KOG1758	NULL	NCBI_GNO_278513	CG2968
KOG1790	NULL	NCBI_GNO_85284	CG6090
KOG2145	NULL	NCBI_GNO_348024	CG9735
KOG2917	NULL	NCBI_GNO_680113	CG8549
KOG3045	NULL	NCBI_GNO_2332013	CG7137
KOG3152	NULL	NCBI_GNO_246373	CG32708
KOG3336	NULL	NCBI_GNO_884033	CG9131
KOG3974	NULL	NCBI_GNO_502283	CG10424
KOG1467	DAPPU-100447	NCBI_GNO_122154	CG10315
KOG1784	DAPPU-100799	NCBI_GNO_90163	CG2021
KOG2781	DAPPU-100904	NCBI_GNO_262163	CG11920
KOG1322	DAPPU-101964	NCBI_GNO_482193	CG1129
KOG0214	DAPPU-102782	NCBI_GNO_102234	CG3180
KOG1436	DAPPU-106454	NCBI_GNO_862373	CG9741
KOG2090	DAPPU-107701	NCBI_GNO_340424	CG7791
KOG2609	DAPPU-109004	NCBI_GNO_604473	CG12343
KOG0981	DAPPU-109061	NCBI_GNO_476474	CG6146
KOG0346	DAPPU-109408	NCBI_GNO_172493	CG1666
KOG2518	DAPPU-110118	NCBI_GNO_140524	CG10387

KOG2270	DAPPU-111070	NCBI_GNO_530563	CG11660
KOG3478	DAPPU-111574	NCBI_GNO_352593	CG7770
KOG3399	DAPPU-111718	NCBI_GNO_662594	CG15309
KOG3273	DAPPU-113251	NCBI_GNO_416663	CG11738
KOG3202	DAPPU-113324	NCBI_GNO_60673	CG7736
KOG1443	DAPPU-113613	NCBI_GNO_8693	CG14971
KOG0361	DAPPU-127024	NCBI_GNO_606044	CG8351
KOG0645	DAPPU-127130	NCBI_GNO_394054	CG12797
KOG2698	DAPPU-127379	NCBI_GNO_616074	CG9441
KOG1637	DAPPU-127463	NCBI_GNO_510084	CG5353
KOG1499	DAPPU-127740	NCBI_GNO_99163	CG6554
KOG0094	DAPPU-128059	NCBI_GNO_108183	CG6601
KOG0305	DAPPU-128430	NCBI_GNO_346264	CG3000
KOG3502	DAPPU-128589	NCBI_GNO_374313	CG2998
KOG1872	DAPPU-129179	NCBI_GNO_106473	CG5384
KOG0242	DAPPU-129226	NCBI_GNO_406474	CG10923
KOG0898	DAPPU-129273	NCBI_GNO_238483	CG8332
KOG0021	DAPPU-129499	NCBI_GNO_102554	CG32495
KOG2740	DAPPU-129909	NCBI_GNO_822643	CG3035
KOG2570	DAPPU-130021	NCBI_GNO_208693	CG4303
KOG1299	DAPPU-186897	NCBI_GNO_380013	CG8228
KOG2012	DAPPU-186898	NCBI_GNO_438014	CG1782
KOG2485	DAPPU-186925	NCBI_GNO_986013	CG17141
KOG2250	DAPPU-187316	NCBI_GNO_40053	CG5320
KOG2897	DAPPU-187339	NCBI_GNO_414054	CG4621
KOG1486	DAPPU-187388	NCBI_GNO_82063	CG6195
KOG3149	DAPPU-187412	NCBI_GNO_744063	CG9207
KOG0370	DAPPU-187511	NCBI_GNO_370084	CG18572
KOG0425	DAPPU-187657	NCBI_GNO_1092103	CG7656
KOG2330	DAPPU-187692	NCBI_GNO_424113	CG3605
KOG0042	DAPPU-187868	NCBI_GNO_322154	CG8256
KOG2509	DAPPU-187913	NCBI_GNO_98173	CG17259
KOG2382	DAPPU-188037	NCBI_GNO_176203	CG2059
KOG1211	DAPPU-188353	NCBI_GNO_130324	CG6007
KOG2882	DAPPU-188403	NCBI_GNO_538344	CG5567
KOG1415	DAPPU-188571	NCBI_GNO_26403	CG4265
KOG3073	DAPPU-188726	NCBI_GNO_130463	CG3527
KOG1586	DAPPU-188735	NCBI_GNO_388463	CG6625
KOG0557	DAPPU-188759	NCBI_GNO_156473	CG5261
KOG2310	DAPPU-188962	NCBI_GNO_164544	CG16928
KOG1780	DAPPU-189031	NCBI_GNO_196563	CG13277
KOG1602	DAPPU-189103	NCBI_GNO_84593	CG5300
KOG0433	DAPPU-189542	NCBI_GNO_206814	CG5414
KOG3103	DAPPU-189944	NCBI_GNO_656013	CG12404
KOG2654	DAPPU-190512	NCBI_GNO_418023	CG13625
KOG2035	DAPPU-190647	NCBI_GNO_1034023	CG6258

KOG3381	DAPPU-190910	NCBI_GNO_2108023	CG7949
KOG1986	DAPPU-190968	NCBI_GNO_344034	CG1250
KOG0396	DAPPU-191245	NCBI_GNO_1198033	CG31357
KOG0188	DAPPU-191289	NCBI_GNO_866034	CG13391
KOG0529	DAPPU-191339	NCBI_GNO_940034	CG12007
KOG3400	DAPPU-191687	NCBI_GNO_758043	CG11246
KOG2298	DAPPU-192225	NULL	CG6778
KOG2413	DAPPU-193046	NCBI_GNO_46083	CG6291
KOG1921	DAPPU-193197	NCBI_GNO_392084	CG9272
KOG1239	DAPPU-193442	NCBI_GNO_84093	CG6404
KOG0057	DAPPU-193653	NCBI_GNO_870093	CG7955
KOG1514	DAPPU-194189	NCBI_GNO_418114	CG10667
KOG0126	DAPPU-194198	NCBI_GNO_834113	CG10466
KOG2613	DAPPU-194355	NCBI_GNO_336123	CG3460
KOG2674	DAPPU-194862	NCBI_GNO_302143	CG6194
KOG2671	DAPPU-195302	NCBI_GNO_146173	CG1074
KOG1490	DAPPU-195334	NCBI_GNO_250173	CG8801
KOG0811	DAPPU-195372	NCBI_GNO_400173	CG5081
KOG1774	DAPPU-195506	NCBI_GNO_806173	CG18591
KOG0284	DAPPU-195849	NCBI_GNO_276194	CG1109
KOG3871	DAPPU-195940	NCBI_GNO_18203	CG1430
KOG2600	DAPPU-196099	NCBI_GNO_408204	CG13097
KOG0567	DAPPU-196422	NCBI_GNO_180233	CG2245
KOG1760	DAPPU-196440	NCBI_GNO_268233	CG10635
KOG1793	DAPPU-196656	NCBI_GNO_320243	CG9915
KOG0376	DAPPU-196725	NCBI_GNO_372244	CG8402
KOG4086	DAPPU-196840	NCBI_GNO_202253	CG1057
KOG1335	DAPPU-197043	NCBI_GNO_76263	CG7430
KOG0450	DAPPU-197428	NCBI_GNO_352294	CG11661
KOG0240	DAPPU-197668	NCBI_GNO_266314	CG7765
KOG2015	DAPPU-198149	NCBI_GNO_282353	CG13343
KOG0925	DAPPU-198525	NCBI_GNO_660373	CG11107
KOG1498	DAPPU-198904	NCBI_GNO_566393	CG1100
KOG2713	DAPPU-198965	NCBI_GNO_124403	CG7441
KOG0768	DAPPU-198969	NCBI_GNO_204403	CG4743
KOG1639	DAPPU-199063	NCBI_GNO_616403	CG10849
KOG0357	DAPPU-200182	NCBI_GNO_66483	CG8439
KOG2840	DAPPU-200348	NCBI_GNO_26493	CG8078
KOG1278	DAPPU-200371	NCBI_GNO_104494	CG9318
KOG3065	DAPPU-200537	NCBI_GNO_132513	CG9474
KOG2159	DAPPU-200873	NCBI_GNO_128544	CG2100
KOG1915	DAPPU-201152	NCBI_GNO_24563	CG3193
KOG0526	DAPPU-201206	NCBI_GNO_252563	CG4817
KOG0509	DAPPU-201447	NCBI_GNO_144584	CG6017
KOG0153	DAPPU-202081	NCBI_GNO_366623	CG14641
KOG3203	DAPPU-202590	NCBI_GNO_250663	CG10603

KOG2805	DAPPU-202999	NCBI_GNO_240703	CG3021
KOG1076	DAPPU-203047	NCBI_GNO_146714	CG4954
KOG1489	DAPPU-203237	NCBI_GNO_110744	CG13390
KOG3262	DAPPU-203390	NCBI_GNO_124764	CG4038
KOG0810	DAPPU-203416	NCBI_GNO_52774	CG31136
KOG1961	DAPPU-204587	NCBI_GNO_112974	CG7371
KOG0622	DAPPU-204767	NCBI_GNO_111144	CG8721
KOG1111	DAPPU-205016	NCBI_GNO_300013	CG6401
KOG2750	DAPPU-205037	NCBI_GNO_436014	CG8414
KOG0820	DAPPU-205158	NCBI_GNO_852013	CG11837
KOG2909	DAPPU-205240	NCBI_GNO_776014	CG8048
KOG1144	DAPPU-205517	NCBI_GNO_2380013	CG10840
KOG1122	DAPPU-205675	NCBI_GNO_440023	CG8545
KOG0090	DAPPU-205698	NCBI_GNO_536023	CG33162
KOG3800	DAPPU-205778	NCBI_GNO_954023	CG7614
KOG2851	DAPPU-206450	NCBI_GNO_1514033	CG7108
KOG0691	DAPPU-206490	NCBI_GNO_1692033	CG17187
KOG2920	DAPPU-207481	NCBI_GNO_1202053	CG17219
KOG1439	DAPPU-207615	NCBI_GNO_120063	CG4422
KOG2187	DAPPU-208793	NCBI_GNO_770094	CG3808
KOG3431	DAPPU-208840	NCBI_GNO_198103	CG13072
KOG2785	DAPPU-209129	NCBI_GNO_96113	CG6769
KOG1985	DAPPU-209148	NCBI_GNO_178114	CG1472
KOG0677	DAPPU-209234	NCBI_GNO_520113	CG9901
KOG0530	DAPPU-209281	NCBI_GNO_744113	CG2976
KOG0739	DAPPU-209675	NCBI_GNO_154133	CG6842
KOG0372	DAPPU-209708	NCBI_GNO_266133	CG32505
KOG3448	DAPPU-209723	NCBI_GNO_296133	CG10418
KOG2967	DAPPU-210151	NCBI_GNO_238154	CG14618
KOG0003	DAPPU-210344	NCBI_GNO_470113	CG2960
KOG2014	DAPPU-210457	NCBI_GNO_326173	CG12276
KOG2103	DAPPU-211018	NCBI_GNO_348194	CG2943
KOG0018	DAPPU-211085	NCBI_GNO_266254	CG6057
KOG1243	DAPPU-212197	NCBI_GNO_56274	CG1973
KOG1077	DAPPU-212220	NCBI_GNO_250273	CG4260
KOG0358	DAPPU-212503	NCBI_GNO_86293	CG5525
KOG3141	DAPPU-212601	NCBI_GNO_298303	CG8288
KOG2151	DAPPU-213023	NULL	CG2161
KOG2680	DAPPU-213265	NCBI_GNO_50363	CG9750
KOG4655	DAPPU-213273	NCBI_GNO_60363	CG4866
KOG0985	DAPPU-213575	NCBI_GNO_704373	CG9012
KOG1521	DAPPU-213644	NCBI_GNO_32383	CG3756
KOG1396	DAPPU-213964	NCBI_GNO_586394	CG31678
KOG2472	DAPPU-214647	NCBI_GNO_453	CG5706
KOG2261	DAPPU-214661	NCBI_GNO_94454	CG7776
KOG2952	DAPPU-215020	NCBI_GNO_400473	CG9947

KOG3230	DAPPU-215551	NULL	CG14542
KOG2523	DAPPU-215796	NCBI_GNO_258533	CG5941
KOG3151	DAPPU-215895	NCBI_GNO_222543	CG4157
KOG0952	DAPPU-216243	NCBI_GNO_310564	CG5205
KOG1295	DAPPU-216440	NCBI_GNO_236583	CG11184
KOG3239	DAPPU-216824	NCBI_GNO_402603	CG9099
KOG2936	DAPPU-216884	NCBI_GNO_88613	CG1416
KOG0432	DAPPU-217307	NCBI_GNO_106644	CG4062
KOG0209	DAPPU-217456	NCBI_GNO_480644	CG6230
KOG3293	DAPPU-218645	NCBI_GNO_120793	CG17768
KOG0342	DAPPU-219114	NCBI_GNO_46873	CG6375
KOG1936	DAPPU-219129	NCBI_GNO_130873	CG6335
KOG0329	DAPPU-219465	NCBI_GNO_148953	CG7269
KOG0761	DAPPU-219561	NCBI_GNO_34983	CG2616
KOG1705	DAPPU-219863	NCBI_GNO_53423	CG2984
KOG1607	DAPPU-220106	NCBI_GNO_1286013	CG3576
KOG0366	DAPPU-220648	NCBI_GNO_2132023	CG18627
KOG2693	DAPPU-220710	NCBI_GNO_160033	CG10449
KOG1189	DAPPU-220868	NCBI_GNO_1066033	CG1828
KOG1070	DAPPU-220970	NCBI_GNO_930034	CG5728
KOG1506	DAPPU-220996	NCBI_GNO_1734033	CG2674
KOG1541	DAPPU-221497	NCBI_GNO_392054	CG10903
KOG1332	DAPPU-221812	NCBI_GNO_1014063	CG6773
KOG1062	DAPPU-221821	NCBI_GNO_1088063	CG9113
KOG0368	DAPPU-222043	NCBI_GNO_624074	CG11198
KOG1612	DAPPU-222073	NCBI_GNO_102083	CG8395
KOG3068	DAPPU-222167	NCBI_GNO_746083	CG9667
KOG0937	DAPPU-222170	NCBI_GNO_782083	CG9388
KOG0796	DAPPU-222176	NCBI_GNO_804083	CG7564
KOG2767	DAPPU-222844	NCBI_GNO_1010113	CG9177
KOG0414	DAPPU-222916	NCBI_GNO_270124	CG1911
KOG0261	DAPPU-222950	NCBI_GNO_554123	CG17209
KOG0343	DAPPU-223010	NCBI_GNO_628124	CG5800
KOG0668	DAPPU-223674	NCBI_GNO_440173	CG17520
KOG2061	DAPPU-224172	NCBI_GNO_594203	CG3260
KOG0010	DAPPU-224315	NCBI_GNO_524214	CG14224
KOG0803	DAPPU-224848	NCBI_GNO_234274	CG32210
KOG0979	DAPPU-226104	NCBI_GNO_174404	CG32438
KOG3460	DAPPU-226183	NCBI_GNO_150413	CG31184
KOG0330	DAPPU-226459	NCBI_GNO_214443	CG9253
KOG0362	DAPPU-226650	NCBI_GNO_556463	CG8258
KOG0481	DAPPU-227386	NCBI_GNO_212543	CG4082
KOG2227	DAPPU-227580	NCBI_GNO_224604	CG5971
KOG0457	DAPPU-227604	NCBI_GNO_164564	CG9638
KOG3080	DAPPU-227674	NCBI_GNO_402564	CG1542
KOG1147	DAPPU-227683	NCBI_GNO_44574	CG5394

KOG2924	DAPPU-228087	NCBI_GNO_340734	CG8005
KOG1390	DAPPU-228205	NCBI_GNO_246634	CG10932
KOG2529	DAPPU-228342	NCBI_GNO_286643	CG3333
KOG2166	DAPPU-228373	NCBI_GNO_260644	CG11861
KOG2141	DAPPU-228663	NCBI_GNO_184694	CG9004
KOG2403	DAPPU-228809	NCBI_GNO_180714	CG17246
KOG0911	DAPPU-229030	NCBI_GNO_22773	CG6523
KOG0103	DAPPU-229110	NCBI_GNO_370774	CG6603
KOG3406	DAPPU-230117	NCBI_GNO_1718013	CG11271
KOG3452	DAPPU-230277	NCBI_GNO_1796033	CG7622
KOG0402	DAPPU-230286	NCBI_GNO_2028033	CG5827
KOG0122	DAPPU-230293	NCBI_GNO_56043	CG10881
KOG3301	DAPPU-230323	NCBI_GNO_1242043	CG3395
KOG1664	DAPPU-230423	NCBI_GNO_354063	CG1088
KOG0407	DAPPU-230521	NCBI_GNO_576093	CG1524
KOG0077	DAPPU-230523	NCBI_GNO_650093	CG7073
KOG3079	DAPPU-230557	NCBI_GNO_1032103	CG6092
KOG1749	DAPPU-230600	NCBI_GNO_168123	CG8415
KOG0767	DAPPU-230626	NCBI_GNO_292133	CG4994
KOG0179	DAPPU-230628	NCBI_GNO_298133	CG4097
KOG3424	DAPPU-230652	NCBI_GNO_358143	CG3751
KOG1775	DAPPU-230679	NCBI_GNO_122153	CG6610
KOG3090	DAPPU-230683	NCBI_GNO_268153	CG15081
KOG1728	DAPPU-230714	NCBI_GNO_190163	CG8857
KOG1647	DAPPU-230767	NCBI_GNO_826173	CG8186
KOG2299	DAPPU-230884	NCBI_GNO_468233	CG13690
KOG0181	DAPPU-230922	NCBI_GNO_128253	CG5266
KOG3318	DAPPU-230984	NCBI_GNO_202283	CG11137
KOG3271	DAPPU-231006	NCBI_GNO_614283	CG3186
KOG3163	DAPPU-231033	NCBI_GNO_116304	CG5277
KOG2691	DAPPU-231046	NCBI_GNO_242303	CG3284
KOG3049	DAPPU-231048	NCBI_GNO_258303	CG3283
KOG0466	DAPPU-231062	NCBI_GNO_286323	CG6476
KOG1816	DAPPU-231070	NCBI_GNO_104333	CG6233
KOG2708	DAPPU-231071	NCBI_GNO_150333	CG4933
KOG3418	DAPPU-231081	NCBI_GNO_110343	CG4759
KOG1678	DAPPU-231144	NCBI_GNO_758373	CG17420
KOG1742	DAPPU-231165	NCBI_GNO_62383	CG15442
KOG3387	DAPPU-231249	NCBI_GNO_212443	CG3949
KOG3343	DAPPU-231254	NCBI_GNO_50453	CG3948
KOG1688	DAPPU-231343	NCBI_GNO_546493	CG11857
KOG0723	DAPPU-231471	NCBI_GNO_290574	CG7394
KOG0182	DAPPU-231549	NCBI_GNO_230613	CG30382
KOG3421	DAPPU-231574	NCBI_GNO_228623	CG6253
KOG0279	DAPPU-231635	NCBI_GNO_552643	CG7111
KOG2239	DAPPU-231649	NCBI_GNO_122663	CG8759

KOG3408	DAPPU-231670	NCBI_GNO_326663	CG3224
KOG0707	DAPPU-231676	NCBI_GNO_276673	CG11811
KOG0176	DAPPU-231717	NCBI_GNO_300713	CG10938
KOG0178	DAPPU-231772	NCBI_GNO_138773	CG9327
KOG0180	DAPPU-231791	NCBI_GNO_34813	CG11981
KOG0121	DAPPU-231808	NCBI_GNO_40843	CG15923
KOG3188	DAPPU-231825	NCBI_GNO_48863	CG6750
KOG3411	DAPPU-231884	NCBI_GNO_148963	CG5338
KOG0127	DAPPU-232196	NCBI_GNO_554014	CG4806
KOG2686	DAPPU-233115	NCBI_GNO_650024	CG2201
KOG0190	DAPPU-234212	NCBI_GNO_1936033	CG6988
KOG0846	DAPPU-235934	NCBI_GNO_886063	CG5219
KOG2653	DAPPU-240002	NCBI_GNO_240154	CG3724
KOG0326	DAPPU-244975	NCBI_GNO_266303	CG4916
KOG0291	DAPPU-257213	NCBI_GNO_432643	CG12325
KOG0734	DAPPU-258323	NCBI_GNO_134684	CG3499
KOG0468	DAPPU-260737	NCBI_GNO_86773	CG4849
KOG1057	DAPPU-260966	NCBI_GNO_372774	CG14616
KOG2963	DAPPU-263794	NCBI_GNO_70883	CG5786
KOG2573	DAPPU-299559	NCBI_GNO_494593	CG13849
KOG2795	DAPPU-299643	NCBI_GNO_410034	CG7753
KOG0548	DAPPU-299674	NCBI_GNO_270553	CG2720
KOG1433	DAPPU-299714	NCBI_GNO_316173	CG7948
KOG0671	DAPPU-299765	NCBI_GNO_336054	CG33553
KOG1367	DAPPU-299795	NCBI_GNO_196013	CG3127
KOG1526	DAPPU-299809	NCBI_GNO_404013	CG7176
KOG2725	DAPPU-299821	NCBI_GNO_576013	CG3803
KOG2726	DAPPU-299896	NCBI_GNO_1376013	CG5705
KOG0011	DAPPU-299929	NCBI_GNO_1804013	CG1836
KOG0108	DAPPU-299933	NCBI_GNO_1822013	CG7697
KOG0419	DAPPU-299959	NCBI_GNO_2240013	CG2013
KOG3323	DAPPU-299963	NCBI_GNO_2288013	CG18643
KOG1507	DAPPU-299973	NCBI_GNO_2384013	CG5330
KOG0534	DAPPU-299980	NCBI_GNO_2502013	CG5946
KOG2732	DAPPU-299981	NCBI_GNO_2504013	CG12018
KOG1980	DAPPU-299992	NCBI_GNO_358014	CG7338
KOG2784	DAPPU-300019	NCBI_GNO_446014	CG2263
KOG1234	DAPPU-300027	NCBI_GNO_486014	CG32649
KOG2013	DAPPU-300104	NCBI_GNO_816014	CG7528
KOG1092	DAPPU-300107	NCBI_GNO_1378013	CG5745
KOG2957	DAPPU-300127	NCBI_GNO_930014	CG2934
KOG0328	DAPPU-300157	NCBI_GNO_1066014	CG7483
KOG0211	DAPPU-300165	NCBI_GNO_2230013	CG17291
KOG2815	DAPPU-300200	NCBI_GNO_30103	CG4207
KOG1448	DAPPU-300219	NCBI_GNO_238103	CG6767
KOG1446	DAPPU-300247	NCBI_GNO_376103	CG17293

KOG0858	DAPPU-300294	NCBI_GNO_832103	CG14899
KOG2248	DAPPU-300412	NCBI_GNO_720104	CG12877
KOG1158	DAPPU-300447	NCBI_GNO_938104	CG11567
KOG1523	DAPPU-300535	NCBI_GNO_708643	CG8978
KOG0900	DAPPU-300540	NCBI_GNO_762643	CG15693
KOG1757	DAPPU-300571	NCBI_GNO_76653	CG5499
KOG3228	DAPPU-300592	NCBI_GNO_258653	CG12135
KOG3185	DAPPU-300595	NCBI_GNO_300653	CG17611
KOG0901	DAPPU-300630	NCBI_GNO_140663	CG3661
KOG0171	DAPPU-300640	NCBI_GNO_180663	CG9240
KOG3437	DAPPU-300659	NCBI_GNO_380663	CG11419
KOG2808	DAPPU-300689	NCBI_GNO_162673	CG6011
KOG3311	DAPPU-300691	NCBI_GNO_208673	CG8900
KOG2005	DAPPU-300699	NCBI_GNO_198674	CG7762
KOG2572	DAPPU-300779	NCBI_GNO_492114	CG10206
KOG1098	DAPPU-300798	NCBI_GNO_412114	CG8939
KOG1123	DAPPU-300803	NCBI_GNO_434114	CG8019
KOG0177	DAPPU-300859	NCBI_GNO_78683	CG17331
KOG2484	DAPPU-300860	NCBI_GNO_96683	CG3983
KOG3361	DAPPU-300869	NCBI_GNO_222683	CG9836
KOG0347	DAPPU-300875	NCBI_GNO_36684	CG9143
KOG3036	DAPPU-300941	NCBI_GNO_204693	CG14213
KOG0271	DAPPU-300951	NCBI_GNO_282693	CG2863
KOG0482	DAPPU-301022	NCBI_GNO_108704	CG4978
KOG1014	DAPPU-301034	NCBI_GNO_220703	CG1444
KOG2016	DAPPU-301131	NCBI_GNO_1076123	CG7828
KOG0478	DAPPU-301153	NCBI_GNO_48714	CG1616
KOG2463	DAPPU-301213	NCBI_GNO_252714	CG2972
KOG1394	DAPPU-301461	NCBI_GNO_352744	CG12170
KOG3060	DAPPU-301475	NCBI_GNO_136753	CG17556
KOG1374	DAPPU-301492	NCBI_GNO_238753	CG3157
KOG3291	DAPPU-301516	NCBI_GNO_168763	CG8922
KOG0964	DAPPU-301524	NCBI_GNO_28763	CG9802
KOG0533	DAPPU-301602	NCBI_GNO_708143	CG1101
KOG2121	DAPPU-301666	NCBI_GNO_672144	CG3298
KOG0989	DAPPU-301682	NCBI_GNO_268763	CG8142
KOG0187	DAPPU-301703	NCBI_GNO_118773	CG3922
KOG0885	DAPPU-301731	NCBI_GNO_204773	CG10907
KOG2830	DAPPU-301783	NCBI_GNO_34783	CG31852
KOG3064	DAPPU-301797	NCBI_GNO_96783	CG10648
KOG2738	DAPPU-301830	NCBI_GNO_110153	CG13630
KOG2670	DAPPU-301844	NCBI_GNO_250153	CG17654
KOG0212	DAPPU-301925	NCBI_GNO_262154	CG5608
KOG2675	DAPPU-301953	NCBI_GNO_404154	CG33979
KOG3198	DAPPU-302042	NCBI_GNO_96813	CG4457
KOG2747	DAPPU-302096	NCBI_GNO_132304	CG6121

KOG0416	DAPPU-302154	NCBI_GNO_474163	CG2257
KOG0397	DAPPU-302274	NCBI_GNO_140173	CG7726
KOG2446	DAPPU-302344	NCBI_GNO_354174	CG8251
KOG1770	DAPPU-302358	NCBI_GNO_410173	CG17737
KOG1889	DAPPU-302360	NCBI_GNO_442173	CG9128
KOG3428	DAPPU-302421	NCBI_GNO_222833	CG10753
KOG3031	DAPPU-302434	NCBI_GNO_38843	CG7993
KOG3364	DAPPU-302435	NCBI_GNO_44843	CG17510
KOG0063	DAPPU-302438	NCBI_GNO_72844	CG5651
KOG2552	DAPPU-302528	NCBI_GNO_176183	CG13089
KOG1309	DAPPU-302577	NCBI_GNO_440183	CG9617
KOG0922	DAPPU-302589	NCBI_GNO_360403	CG8241
KOG1112	DAPPU-302612	NCBI_GNO_226183	CG5371
KOG1722	DAPPU-302765	NCBI_GNO_322863	CG9282
KOG0960	DAPPU-302792	NCBI_GNO_238193	CG3731
KOG1813	DAPPU-302909	NCBI_GNO_162873	CG4973
KOG1379	DAPPU-302940	NCBI_GNO_64883	CG15035
KOG3235	DAPPU-303043	NCBI_GNO_1744023	CG11989
KOG2229	DAPPU-303148	NCBI_GNO_752023	CG8070
KOG2007	DAPPU-303176	NCBI_GNO_700024	CG8431
KOG1898	DAPPU-303178	NCBI_GNO_1212023	CG13900
KOG1301	DAPPU-303233	NCBI_GNO_1024024	CG3539
KOG1539	DAPPU-303234	NCBI_GNO_1026024	CG9799
KOG1914	DAPPU-303263	NCBI_GNO_1148024	CG17170
KOG0175	DAPPU-303318	NCBI_GNO_500203	CG12323
KOG0373	DAPPU-303396	NCBI_GNO_10213	CG12217
KOG2907	DAPPU-303407	NCBI_GNO_176213	CG13418
KOG0173	DAPPU-303409	NCBI_GNO_148213	CG3329
KOG3342	DAPPU-303416	NCBI_GNO_286213	CG2358
KOG3283	DAPPU-303431	NCBI_GNO_504213	CG7808
KOG2067	DAPPU-303461	NCBI_GNO_296214	CG8728
KOG1999	DAPPU-303545	NCBI_GNO_174904	CG7626
KOG0285	DAPPU-303553	NCBI_GNO_226904	CG1796
KOG0058	DAPPU-303561	NCBI_GNO_258904	CG3156
KOG3234	DAPPU-303577	NCBI_GNO_102913	CG14222
KOG1772	DAPPU-303591	NCBI_GNO_28223	CG6213
KOG2314	DAPPU-303630	NCBI_GNO_288914	CG4878
KOG1302	DAPPU-303678	NCBI_GNO_270234	CG12230
KOG0289	DAPPU-303758	NCBI_GNO_408233	CG5519
KOG2635	DAPPU-303794	NCBI_GNO_208924	CG14813
KOG1662	DAPPU-303877	NCBI_GNO_680243	CG4307
KOG0217	DAPPU-303932	NCBI_GNO_370244	CG7003
KOG1550	DAPPU-303934	NCBI_GNO_374244	CG10221
KOG2386	DAPPU-303937	NCBI_GNO_576243	CG1810
KOG0119	DAPPU-303941	NCBI_GNO_426244	CG5836
KOG1692	DAPPU-304000	NCBI_GNO_188253	CG3564

KOG0864	DAPPU-304059	NCBI_GNO_94254	CG11856
KOG4409	DAPPU-304102	NCBI_GNO_336254	CG1882
KOG1943	DAPPU-304145	NCBI_GNO_522254	CG7261
KOG1465	DAPPU-304186	NCBI_GNO_202263	CG2677
KOG1270	DAPPU-304225	NCBI_GNO_208954	CG9249
KOG0360	DAPPU-304295	NCBI_GNO_320283	CG5374
KOG3174	DAPPU-304315	NCBI_GNO_592283	CG17158
KOG0948	DAPPU-304340	NCBI_GNO_182284	CG4152
KOG0947	DAPPU-304359	NCBI_GNO_248284	CG10210
KOG0225	DAPPU-304437	NCBI_GNO_90293	CG7010
KOG2144	DAPPU-304445	NCBI_GNO_362293	CG4561
KOG2799	DAPPU-304522	NCBI_GNO_52983	CG11963
KOG2861	DAPPU-304531	NCBI_GNO_148983	CG11679
KOG3087	DAPPU-304577	NCBI_GNO_330033	CG10673
KOG0727	DAPPU-304599	NCBI_GNO_700033	CG16916
KOG1671	DAPPU-304712	NCBI_GNO_934033	CG7361
KOG1173	DAPPU-304732	NCBI_GNO_756034	CG6759
KOG0100	DAPPU-304735	NCBI_GNO_762034	CG4147
KOG0967	DAPPU-304798	NCBI_GNO_1092034	CG5602
KOG0467	DAPPU-304801	NCBI_GNO_1108034	CG33158
KOG2988	DAPPU-304893	NCBI_GNO_50993	CG10652
KOG1984	DAPPU-304899	NCBI_GNO_58994	CG10882
KOG2885	DAPPU-304910	NCBI_GNO_288313	CG4510
KOG1562	DAPPU-304917	NCBI_GNO_404313	CG8327
KOG2978	DAPPU-304921	NCBI_GNO_434313	CG10166
KOG3487	DAPPU-304934	NCBI_GNO_120313	CG5161
KOG2672	DAPPU-304936	NCBI_GNO_122313	CG5231
KOG1555	DAPPU-304982	NCBI_GNO_48323	CG18174
KOG0479	DAPPU-305021	NCBI_GNO_54323	CG4206
KOG2030	DAPPU-305191	NCBI_GNO_324343	CG11847
KOG2580	DAPPU-305300	NCBI_GNO_414353	CG11779
KOG0306	DAPPU-305316	NCBI_GNO_352353	CG8064
KOG0712	DAPPU-305330	NCBI_GNO_536353	CG8863
KOG2825	DAPPU-305381	NCBI_GNO_76364	CG1598
KOG2420	DAPPU-305406	NCBI_GNO_226364	CG5991
KOG2335	DAPPU-305459	NCBI_GNO_486043	CG3645
KOG2874	DAPPU-305506	NCBI_GNO_1114043	CG4258
KOG3003	DAPPU-305539	NCBI_GNO_1602043	CG6155
KOG3013	DAPPU-305543	NCBI_GNO_1672043	CG3931
KOG3106	DAPPU-305544	NCBI_GNO_1712043	CG5183
KOG1294	DAPPU-305561	NCBI_GNO_1186044	CG3178
KOG1867	DAPPU-305599	NCBI_GNO_324043	CG4166
KOG2340	DAPPU-305636	NCBI_GNO_562044	CG3735
KOG0959	DAPPU-305640	NCBI_GNO_590044	CG5517
KOG0282	DAPPU-305671	NCBI_GNO_734044	CG6015
KOG3240	DAPPU-305782	NCBI_GNO_506363	CG9245

KOG3092	DAPPU-305830	NCBI_GNO_156373	CG15224
KOG3187	DAPPU-305850	NCBI_GNO_250373	CG6746
KOG0687	DAPPU-305955	NCBI_GNO_120383	CG5378
KOG0418	DAPPU-305966	NCBI_GNO_232383	CG8284
KOG0096	DAPPU-305970	NCBI_GNO_274383	CG1404
KOG1241	DAPPU-306007	NCBI_GNO_132384	CG2637
KOG2004	DAPPU-306039	NCBI_GNO_264384	CG8798
KOG0477	DAPPU-306072	NCBI_GNO_556384	CG7538
KOG1255	DAPPU-306118	NCBI_GNO_390393	CG1065
KOG2848	DAPPU-306184	NCBI_GNO_634393	CG3812
KOG0264	DAPPU-306287	NCBI_GNO_728403	CG4236
KOG1698	DAPPU-306295	NCBI_GNO_784403	CG8039
KOG1018	DAPPU-306296	NCBI_GNO_786403	CG5292
KOG2217	DAPPU-306323	NCBI_GNO_202413	CG6686
KOG2915	DAPPU-306325	NCBI_GNO_254413	CG14544
KOG2387	DAPPU-306334	NCBI_GNO_132413	CG6854
KOG0729	DAPPU-306359	NCBI_GNO_126053	CG1341
KOG0359	DAPPU-306375	NCBI_GNO_354053	CG8231
KOG2981	DAPPU-306423	NCBI_GNO_1026053	CG6877
KOG0184	DAPPU-306433	NCBI_GNO_1188053	CG1519
KOG1351	DAPPU-306451	NCBI_GNO_1334053	CG17369
KOG2916	DAPPU-306457	NCBI_GNO_1426053	CG9946
KOG2700	DAPPU-306459	NCBI_GNO_1438053	CG3590
KOG1567	DAPPU-306462	NCBI_GNO_1472053	CG8975
KOG2481	DAPPU-306492	NCBI_GNO_322053	CG4364
KOG2759	DAPPU-306505	NCBI_GNO_448053	CG17332
KOG0110	DAPPU-306569	NCBI_GNO_660054	CG3335
KOG3181	DAPPU-306633	NCBI_GNO_192424	CG6779
KOG1781	DAPPU-306642	NCBI_GNO_330423	CG13277
KOG2696	DAPPU-306646	NCBI_GNO_4423	CG2051
KOG3274	DAPPU-306670	NCBI_GNO_198423	CG5902
KOG0559	DAPPU-306760	NCBI_GNO_154434	CG5214
KOG2253	DAPPU-306802	NCBI_GNO_326433	CG4119
KOG0364	DAPPU-306806	NCBI_GNO_404433	CG8977
KOG3368	DAPPU-306826	NCBI_GNO_640433	CG1359
KOG2537	DAPPU-306844	NCBI_GNO_620433	CG10627
KOG0260	DAPPU-306846	NCBI_GNO_512434	CG1554
KOG1906	DAPPU-306880	NCBI_GNO_106444	CG11265
KOG1300	DAPPU-306883	NCBI_GNO_110444	CG15811
KOG1331	DAPPU-306906	NCBI_GNO_278444	CG17807
KOG0337	DAPPU-306922	NCBI_GNO_340444	CG32344
KOG1342	DAPPU-306940	NCBI_GNO_16453	CG7471
KOG1145	DAPPU-306984	NCBI_GNO_384454	CG12413
KOG4020	DAPPU-307032	NCBI_GNO_10063	CG4180
KOG0817	DAPPU-307101	NCBI_GNO_195163	CG8627
KOG1148	DAPPU-307141	NCBI_GNO_802064	CG10506

KOG0263	DAPPU-307153	NCBI_GNO_1096063	CG7704
KOG1493	DAPPU-307195	NULL	CG34441
KOG0480	DAPPU-307300	NCBI_GNO_302473	CG4039
KOG1357	DAPPU-307317	NCBI_GNO_494473	CG4162
KOG0816	DAPPU-307342	NCBI_GNO_786473	CG1381
KOG1652	DAPPU-307357	NCBI_GNO_918473	CG15257
KOG2540	DAPPU-307441	NCBI_GNO_336483	CG31915
KOG3347	DAPPU-307499	NCBI_GNO_494483	CG8816
KOG3355	DAPPU-307542	NCBI_GNO_248493	CG12534
KOG3282	DAPPU-307543	NCBI_GNO_260493	CG1307
KOG1672	DAPPU-307585	NCBI_GNO_494493	CG4511
KOG2421	DAPPU-307611	NCBI_GNO_412494	CG2818
KOG1800	DAPPU-307619	NCBI_GNO_470494	CG12390
KOG1487	DAPPU-307651	NCBI_GNO_268503	CG8340
KOG1992	DAPPU-307683	NCBI_GNO_31143	CG13281
KOG2023	DAPPU-307700	NCBI_GNO_458504	CG7398
KOG2792	DAPPU-307706	NCBI_GNO_114513	CG8885
KOG3372	DAPPU-307738	NCBI_GNO_64073	CG5677
KOG1655	DAPPU-307757	NCBI_GNO_382073	CG6259
KOG1753	DAPPU-307788	NCBI_GNO_918073	CG4046
KOG1590	DAPPU-307796	NCBI_GNO_1068073	CG14290
KOG0554	DAPPU-307813	NCBI_GNO_142074	CG6796
KOG2292	DAPPU-307851	NCBI_GNO_302074	CG7748
KOG0299	DAPPU-307893	NCBI_GNO_528074	CG33505
KOG2058	DAPPU-307907	NCBI_GNO_1220073	CG5916
KOG3432	DAPPU-308028	NCBI_GNO_386523	CG8210
KOG0605	DAPPU-308031	NCBI_GNO_468524	CG8637
KOG2770	DAPPU-308047	NCBI_GNO_126533	CG6415
KOG0325	DAPPU-308051	NCBI_GNO_66534	CG9804
KOG0651	DAPPU-308117	NCBI_GNO_198543	CG3455
KOG2608	DAPPU-308173	NCBI_GNO_424543	CG1333
KOG1137	DAPPU-308207	NCBI_GNO_472543	CG7698
KOG3349	DAPPU-308239	NCBI_GNO_126553	CG14512
KOG4098	DAPPU-308250	NCBI_GNO_332553	CG6302
KOG0580	DAPPU-308251	NCBI_GNO_286574	CG3068
KOG0688	DAPPU-308328	NCBI_GNO_96563	CG5605
KOG2768	DAPPU-308414	NCBI_GNO_706083	CG4153
KOG1979	DAPPU-308451	NCBI_GNO_318084	CG11482
KOG1149	DAPPU-308493	NCBI_GNO_876083	CG4573
KOG3298	DAPPU-308561	NCBI_GNO_290564	CG31344
KOG0585	DAPPU-308627	NCBI_GNO_288574	CG17698
KOG2769	DAPPU-308637	NCBI_GNO_74583	CG7757
KOG3430	DAPPU-308639	NCBI_GNO_54583	CG6998
KOG2171	DAPPU-308670	NCBI_GNO_583	CG1059
KOG3237	DAPPU-308779	NCBI_GNO_302593	CG1789
KOG3490	DAPPU-308814	NCBI_GNO_434593	CG12372

KOG0400	DAPPU-308825	NCBI_GNO_518593	CG13389
KOG1373	DAPPU-308832	NCBI_GNO_578593	CG9539
KOG0036	DAPPU-308837	NCBI_GNO_454594	CG32103
KOG0523	DAPPU-308853	NCBI_GNO_604594	CG8036
KOG0822	DAPPU-308857	NCBI_GNO_544593	CG3730
KOG3111	DAPPU-308929	NCBI_GNO_434603	CG30499
KOG2863	DAPPU-308938	NCBI_GNO_360603	CG7942
KOG3129	DAPPU-309024	NCBI_GNO_920093	CG9588
KOG1099	DAPPU-309098	NCBI_GNO_670093	CG5220
KOG2688	DAPPU-309103	NCBI_GNO_706093	CG7351
KOG3486	DAPPU-309158	NCBI_GNO_258613	CG2986
KOG3315	DAPPU-309327	NCBI_GNO_194633	CG10153
KOG3059	DAPPU-309328	NCBI_GNO_254634	CG12077
KOG1636	DAPPU-309373	NCBI_GNO_224633	CG9193
KOG3063	DAPPU-309382	NCBI_GNO_296634	CG14804
KOG1534	DAPPU-309440	NCBI_GNO_188643	CG2656
KOG1060	DAPPU-309510	NCBI_GNO_376014	CG11427
KOG3470	DAPPU-309511	NCBI_GNO_162013	CG1890
KOG1424	DAPPU-309652	NCBI_GNO_580014	CG14788
KOG1043	DAPPU-309740	NCBI_GNO_734014	CG4589
KOG0933	DAPPU-310261	NCBI_GNO_728023	CG10212
KOG2102	DAPPU-310536	NCBI_GNO_992024	CG6413
KOG1533	DAPPU-310542	NCBI_GNO_998024	CG10222
KOG2749	DAPPU-310603	NCBI_GNO_1098024	CG5970
KOG1210	DAPPU-310703	NCBI_GNO_360034	CG10425
KOG1651	DAPPU-310801	NCBI_GNO_516034	CG12013
KOG0050	DAPPU-311294	NCBI_GNO_176043	CG6905
KOG0302	DAPPU-311335	NCBI_GNO_362044	CG12792
KOG1315	DAPPU-311398	NCBI_GNO_482044	CG1407
KOG2568	DAPPU-311546	NCBI_GNO_720044	CG17660
KOG1416	DAPPU-312712	NCBI_GNO_1024064	CG9596
KOG2051	DAPPU-312843	NCBI_GNO_288074	CG2253
KOG1058	DAPPU-313075	NCBI_GNO_212084	CG6223
KOG1919	DAPPU-313092	NCBI_GNO_234084	CG6187
KOG1035	DAPPU-313217	NCBI_GNO_416084	CG1609
KOG0939	DAPPU-313219	NCBI_GNO_418084	CG8184
KOG2180	DAPPU-313250	NCBI_GNO_460084	CG3338
KOG0711	DAPPU-313317	NCBI_GNO_1204083	CG12389
KOG0102	DAPPU-313359	NCBI_GNO_618084	CG8542
KOG2811	DAPPU-313492	NCBI_GNO_296094	CG18048
KOG3075	DAPPU-313516	NCBI_GNO_230093	CG30410
KOG1220	DAPPU-314497	NCBI_GNO_852113	CG8073
KOG1440	DAPPU-315618	NCBI_GNO_654144	CG7962
KOG1107	DAPPU-315698	NCBI_GNO_170153	CG5625
KOG3153	DAPPU-315739	NCBI_GNO_292153	CG14721
KOG3457	DAPPU-315783	NCBI_GNO_476153	CG10130

KOG0434	DAPPU-316089	NCBI_GNO_320163	CG11471
KOG0733	DAPPU-316737	NCBI_GNO_344184	CG8571
KOG2360	DAPPU-317133	NCBI_GNO_520193	CG5558
KOG1620	DAPPU-317135	NCBI_GNO_73154	CG10082
KOG2807	DAPPU-317251	NCBI_GNO_118203	CG11115
KOG0969	DAPPU-317527	NCBI_GNO_264214	CG5949
KOG2245	DAPPU-318160	NCBI_GNO_398234	CG9854
KOG1968	DAPPU-318220	NCBI_GNO_166244	CG1119
KOG2971	DAPPU-318276	NCBI_GNO_306243	CG11583
KOG0250	DAPPU-318587	NCBI_GNO_372254	CG5524
KOG0436	DAPPU-319387	NCBI_GNO_210294	CG31322
KOG2754	DAPPU-319543	NCBI_GNO_76304	CG9022
KOG1073	DAPPU-319965	NCBI_GNO_216324	CG10686
KOG2992	DAPPU-319994	NCBI_GNO_42334	CG7421
KOG0616	DAPPU-319998	NCBI_GNO_58333	CG4379
KOG3222	DAPPU-320031	NCBI_GNO_158333	CG8891
KOG0344	DAPPU-321368	NCBI_GNO_548364	CG5589
KOG1625	DAPPU-321802	NCBI_GNO_108384	CG5923
KOG2308	DAPPU-321889	NCBI_GNO_244384	CG8552
KOG0962	DAPPU-322547	NCBI_GNO_316404	CG6339
KOG0740	DAPPU-322816	NCBI_GNO_258414	CG3326
KOG2554	DAPPU-323292	NCBI_GNO_132433	CG3045
KOG2068	DAPPU-323297	NCBI_GNO_98434	CG31716
KOG0313	DAPPU-323838	NCBI_GNO_92454	CG6724
KOG1956	DAPPU-324341	NCBI_GNO_168474	CG10123
KOG1131	DAPPU-324413	NCBI_GNO_280474	CG9433
KOG1831	DAPPU-324499	NCBI_GNO_470474	CG34407
KOG2111	DAPPU-324944	NCBI_GNO_304494	CG11975
KOG0780	DAPPU-325390	NCBI_GNO_244513	CG4659
KOG0219	DAPPU-325677	NCBI_GNO_282524	CG4215
KOG1597	DAPPU-325996	NCBI_GNO_132544	CG5193
KOG1349	DAPPU-326467	NCBI_GNO_284563	CG4406
KOG1540	DAPPU-326731	NCBI_GNO_86584	CG2453
KOG2201	DAPPU-327249	NCBI_GNO_128604	CG5725
KOG0524	DAPPU-327265	NCBI_GNO_270603	CG11876
KOG2020	DAPPU-327633	NCBI_GNO_62623	CG12234
KOG2585	DAPPU-327994	NCBI_GNO_478633	CG2974
KOG3022	DAPPU-328248	NCBI_GNO_814643	CG17904
KOG1989	DAPPU-328331	NCBI_GNO_126644	CG10637
KOG2438	DAPPU-328680	NCBI_GNO_230664	CG5463
KOG0213	DAPPU-328912	NCBI_GNO_70684	CG2807
KOG0152	DAPPU-329118	NCBI_GNO_158693	CG3542
KOG2198	DAPPU-329423	NCBI_GNO_26713	CG6133
KOG1268	DAPPU-329474	NCBI_GNO_152714	CG1345
KOG0355	DAPPU-329527	NCBI_GNO_226714	CG10223
KOG3327	DAPPU-331571	NCBI_GNO_48834	CG5757

KOG0442	DAPPU-331812	NCBI_GNO_184844	CG3697
KOG2423	DAPPU-332475	NCBI_GNO_246874	CG6501
KOG1596	DAPPU-333298	NCBI_GNO_202924	CG9888
KOG0216	DAPPU-333503	NCBI_GNO_310934	CG4033
KOG1491	DAPPU-333762	NCBI_GNO_156953	CG1354
KOG0435	DAPPU-334969	NCBI_GNO_79074	CG7479
KOG2574	DAPPU-337041	NCBI_GNO_47243	CG6876
KOG3295	DAPPU-34614	NCBI_GNO_1140113	CG4651
KOG0537	DAPPU-36528	NCBI_GNO_136383	CG2140
KOG4032	DAPPU-36719	NULL	CG14543
KOG0131	DAPPU-40110	NCBI_GNO_2314013	CG3780
KOG3034	DAPPU-40241	NCBI_GNO_1580023	CG9286
KOG3438	DAPPU-40279	NCBI_GNO_1170023	CG10685
KOG3454	DAPPU-40364	NCBI_GNO_1724023	CG5454
KOG0351	DAPPU-40427	NCBI_GNO_888024	CG6920
KOG1614	DAPPU-40604	NCBI_GNO_222023	CG9606
KOG2930	DAPPU-40730	NCBI_GNO_250023	CG16982
KOG2868	DAPPU-40876	NCBI_GNO_1560023	CG11183
KOG0812	DAPPU-41202	NCBI_GNO_1980033	CG4214
KOG1402	DAPPU-41314	NCBI_GNO_312033	CG8782
KOG1361	DAPPU-41973	NCBI_GNO_404044	CG10018
KOG3346	DAPPU-42659	NCBI_GNO_944044	CG10298
KOG2086	DAPPU-42769	NCBI_GNO_544053	CG11139
KOG3316	DAPPU-42928	NCBI_GNO_1158053	CG6196
KOG2050	DAPPU-43721	NCBI_GNO_110063	CG1685
KOG0420	DAPPU-44414	NCBI_GNO_440073	CG7375
KOG2884	DAPPU-44914	NCBI_GNO_832083	CG7619
KOG0717	DAPPU-44917	NCBI_GNO_1438083	CG2790
KOG2728	DAPPU-45011	NCBI_GNO_348083	CG5629
KOG1975	DAPPU-45556	NCBI_GNO_618093	CG3688
KOG1240	DAPPU-45609	NCBI_GNO_742094	CG9746
KOG2311	DAPPU-45739	NCBI_GNO_1056104	CG4610
KOG0620	DAPPU-46150	NCBI_GNO_316104	CG31137
KOG1549	DAPPU-46336	NCBI_GNO_180113	CG12264
KOG1336	DAPPU-46525	NCBI_GNO_196114	CG4199
KOG1600	DAPPU-46667	NCBI_GNO_229454	CG5887
KOG0804	DAPPU-47474	NCBI_GNO_322134	CG5555
KOG2519	DAPPU-47502	NCBI_GNO_70133	CG8648
KOG0201	DAPPU-47789	NCBI_GNO_390144	CG5169
KOG0563	DAPPU-48979	NCBI_GNO_220173	CG12529
KOG0462	DAPPU-49405	NCBI_GNO_624193	CG1410
KOG1038	DAPPU-49577	NCBI_GNO_340194	CG4644
KOG1556	DAPPU-49603	NCBI_GNO_250194	CG3416
KOG2520	DAPPU-49689	NCBI_GNO_162194	CG10890
KOG2711	DAPPU-51324	NCBI_GNO_564253	CG9042
KOG0728	DAPPU-52572	NCBI_GNO_348303	CG2241

KOG2467	DAPPU-52799	NCBI_GNO_322313	CG3011
KOG2268	DAPPU-53510	NCBI_GNO_388343	CG11859
KOG2036	DAPPU-54191	NCBI_GNO_560363	CG1994
KOG1253	DAPPU-54762	NCBI_GNO_88384	CG6388
KOG0243	DAPPU-55076	NCBI_GNO_374394	CG9191
KOG3233	DAPPU-56706	NCBI_GNO_396454	CG5380
KOG4018	DAPPU-5678	NCBI_GNO_170763	CG5515
KOG1380	DAPPU-57149	NCBI_GNO_298473	CG5037
KOG1135	DAPPU-58164	NCBI_GNO_268513	CG1957
KOG1463	DAPPU-58294	NCBI_GNO_378523	CG10149
KOG1783	DAPPU-59069	NCBI_GNO_48554	CG9344
KOG1381	DAPPU-59083	NCBI_GNO_110554	CG9613
KOG2989	DAPPU-59672	NCBI_GNO_150584	CG8435
KOG2877	DAPPU-60250	NCBI_GNO_38603	CG6016
KOG1725	DAPPU-61104	NCBI_GNO_488633	CG8331
KOG1272	DAPPU-61337	NCBI_GNO_182643	CG2260
KOG2841	DAPPU-61546	NCBI_GNO_412644	CG10215
KOG3000	DAPPU-62015	NCBI_GNO_86683	CG3265
KOG2241	DAPPU-62407	NCBI_GNO_460693	CG15100
KOG1656	DAPPU-62579	NCBI_GNO_60713	CG8055
KOG1461	DAPPU-62668	NCBI_GNO_162714	CG3806
KOG2322	DAPPU-63064	NCBI_GNO_296743	CG3798
KOG3172	DAPPU-63450	NCBI_GNO_142773	CG8427
KOG2280	DAPPU-63503	NCBI_GNO_104774	CG8454
KOG3405	DAPPU-64901	NCBI_GNO_40873	CG1163
KOG0202	DAPPU-65262	NCBI_GNO_164903	CG3725
KOG1274	DAPPU-65744	NCBI_GNO_130944	CG13350
KOG0996	DAPPU-67196	NCBI_GNO_418843	CG11397
KOG0876	DAPPU-67591	NCBI_GNO_297184	CG8905
KOG0371	DAPPU-68048	NCBI_GNO_135244	CG7109
KOG0009	DAPPU-93183	NCBI_GNO_462513	CG15697
KOG1626	DAPPU-93571	NCBI_GNO_110713	CG4634
KOG2659	DAPPU-93654	NCBI_GNO_170783	CG6617
KOG3386	DAPPU-93662	NCBI_GNO_310813	CG3977
KOG3325	DAPPU-93995	NCBI_GNO_376013	CG4764
KOG0270	DAPPU-96073	NCBI_GNO_510044	CG6751
KOG0813	DAPPU-96363	NCBI_GNO_1562043	CG4365
KOG2779	DAPPU-96817	NCBI_GNO_1180053	CG7436
KOG2804	DAPPU-97573	NCBI_GNO_250074	CG18330
KOG1670	DAPPU-98425	NCBI_GNO_582093	CG32859
KOG0556	DAPPU-99304	NCBI_GNO_830113	CG3821
KOG3020	DAPPU-99659	NCBI_GNO_916123	CG3358
KOG0185	DAPPU-99667	NCBI_GNO_942123	CG12000
KOG3257	DAPPU-99708	NCBI_GNO_1000123	CG3351

Table S16. Fifty predicted *Daphnia pulex* miRNA

miRNA name	Scaffold	Pre-miRNA Start position	Pre-miRNA End position	Strand	Mature miRNA Start position	Mature miRNA End position
dpul-bantam	scaffold_115	370155	370238	1	55	77
dpul-let-7	scaffold_71	446440	446534	-1	14	35
dpul-mir-1	scaffold_1	1720872	1720960	-1	57	78
dpul-mir-10	scaffold_7	304805	304905	-1	22	42
dpul-mir-100	scaffold_71	446641	446740	-1	21	43
dpul-mir-1175	scaffold_113	97584	97667	1	53	76
dpul-mir-12	scaffold_1	1847835	1847917	-1	13	35
dpul-mir-124	scaffold_120	76886	76970	1	55	77
dpul-mir-125	scaffold_71	445340	445450	-1	24	45
dpul-mir-125b-as	scaffold_71	445352	445433	1	55	76
dpul-mir-133	scaffold_1	1708481	1708584	-1	67	88
dpul-mir-137	scaffold_92	410926	411003	1	49	70
dpul-mir-13b	scaffold_80	240721	240800	1	51	73
dpul-mir-153	scaffold_3	3560633	3560719	-1	53	72
dpul-mir-193	scaffold_167	85443	85550	-1	73	94
dpul-mir-2-1	scaffold_80	240857	240946	1	55	77
dpul-mir-2-2	scaffold_80	241036	241112	1	48	70
dpul-mir-210	scaffold_51	480329	480413	-1	51	71
dpul-mir-219	scaffold_253	93588	93666	-1	11	33
dpul-mir-252a	scaffold_285	66051	66144	1	16	37
dpul-mir-252b	scaffold_8	127361	127465	1	20	42
dpul-mir-263b	scaffold_87	475808	475882	1	11	30
dpul-mir-275	scaffold_4	1790732	1790817	1	51	73
dpul-mir-276	scaffold_15	755622	755692	1	46	67
dpul-mir-277	scaffold_4	1242957	1243058	-1	61	85
dpul-mir-279	scaffold_43	177495	177579	1	52	70
dpul-mir-281	scaffold_11	1065349	1065415	-1	1	21
dpul-mir-283	scaffold_1	1848733	1848832	-1	21	40
dpul-mir-29	scaffold_1	332494	332591	1	62	83
dpul-mir-309	scaffold_24	361460	361528	-1	44	65
dpul-mir-315	scaffold_58	431897	431975	1	12	33
dpul-mir-317	scaffold_4	1243950	1244040	-1	56	80
dpul-mir-33	scaffold_90	265090	265171	-1	7	28
dpul-mir-34	scaffold_4	1242031	1242127	-1	14	35
dpul-mir-36	scaffold_32	68509	68591	-1	51	70
dpul-mir-7-1	scaffold_11571	1020	1108	-1	15	37
dpul-mir-7-2	scaffold_191	112539	112627	-1	15	37
dpul-mir-71	scaffold_80	240421	240502	1	10	31
dpul-mir-8	scaffold_131	139395	139479	1	52	74
dpul-mir-87-1	scaffold_1	2190890	2190989	1	70	89
dpul-mir-87-2	scaffold_1	2191051	2191151	1	71	90

dpul-mir-92b	scaffold_38	876312	876410	1	60	81
dpul-mir-92c	scaffold_38	876134	876234	1	61	82
dpul-mir-965	scaffold_32	27762	27867	-1	65	86
dpul-mir-981	scaffold_2	1450976	1451073	-1	62	83
dpul-mir-993	scaffold_7	282304	282393	1	57	79
dpul-mir-9a	scaffold_2	1526199	1526285	1	15	37
dpul-mir-9b	scaffold_32	69569	69641	-1	9	31
dpul-mir-iab-4	scaffold_7	515533	515617	1	15	36
dpul-mir-iab-4as	scaffold_7	515541	515609	-1	6	28

Table S17. Comparative analysis of transposable elements (TEs) in *Daphnia pulex.* Among arthropods, *D. pulex* is similar in terms of repeat content, with most families being present in low copy number. *Daphnia pulex* does, however, contain a large number of novel TE families [S172] and many, diverse families for which there is evidence of possible recent activity [S173].

		Daphnia	*Drosophila*	*Aedes*	*Anopheles*	*Apis*	*Mus*
Proportion of genome (euchromatin)	DNA transposons	0.70%	0.31%[1]	20%	n/a	~1%	0.88%
	Retrotransposons	8.66%	3.47%[1]	26.5%	n/a	almost none	37.29%
	Total	9.4%	5.3%[2]	47%	16%	1%	38.55%
Highest copy number family		*gypsy*	*roo*[1]	*Felai-B*	*Sine200*	*Mariner*	LINE1
References			[1][S174] [2][S175]	[S176]	[S177]	[S178]	[S179]

Table S18. Classification and distribution of transposable elements in *Daphnia pulex*. The *D. pulex* genome contains representatives of 10 of the known superfamilies of DNA transposons (see also [S173]), including *Helitrons* which are found in tandem arrays. Also, *D. pulex* has the highest number of families of *Copia* elements of which we are aware described to date (44) compared with other arthropod genomes (see also [S172]). In addition, 15 families of DIRS elements were found in this study, a group previously annotated mainly in fish genomes which have not been found in other arthropod genomes (except *Tribolium castaneum*). Copy number estimates are based on RepeatMasker [http://www.repeatmasker.org] output (masked regions >50 bp in length, >70% similarity, and >20% of the length of the query).

Class	Subclass	Superfamily	# of families	Copy number	Proportion of genome (%)
DNA Transposons	TIRs	CACTA	10	109	0.0536
		hAT	6	33	0.0180
		Merlin	1	26	0.0160
		Mutator	10	195	0.0657
		P element	9	70	0.0411
		PIF	2	15	0.0061
		TTAA	3	685	0.2321
		Tc1/mariner	7	217	0.0676
		SUBTOTAL	*48*	*1,350*	*0.5003*
	Helitrons	*Helitron*	4	573	0.2005
	Maverick	*Maverick*	4	5	0.0038
SUBTOTAL			*56*	*1,928*	*0.7046*
Retrotransposons	LTR retrotransposons	BEL	26	793	1.8249
		Copia	44	600	1.1596
		DIRS	15	218	0.2715
		Gypsy	56	2,163	4.7192
		SUBTOTAL	*141*	*3,774*	*7.9752*
	Non-LTR retrotransposons	I	19	633	0.2163
		LOA	16	244	0.0872
		L1	3	138	0.0787
		L2	27	593	0.2246
		NeSL	8	104	0.0270
		SINEs	5	404	0.0520
		SUBTOTAL	*78*	*2,116*	*0.6858*
SUBTOTAL			*219*	*5,890*	*8.6610*
TOTAL			*275*	*7,821*	*9.3656*

D. Attributes of a Compact Genome

Table 19. Gene richness within a comparatively small genome. Various features of the *Daphnia pulex* genome compared to those of *Drosophila melanogaster* (relatively small arthropod genome), *Apis mellifera* (somewhat larger arthropod genome), *Caenorhabditis elegans* (small, gene-rich genome) and *Mus musculus* (large, gene-rich genome). *Daphnia pulex* values for the number of genes, gene span, intron size and intergenic size are outside the 95% confidence intervals when randomly sampling six other arthropod genomes.

	Daphnia	*Apis*	*Drosophila*	*Caenorhabditis*	*Mus*
Genome size in Mbp[1]	200 (150)	235 (150)	180 (120)	100 (100)	3,450 (2,600)
Number of genes	31,000+	17,000	13,700	20,100	27,600
Avg. span of a coding gene in bp	2,300	9,900	4,000	3,000	32,000
Avg. number of exons/gene	6.6	7.1	4.0	6.0	8.0
Avg. number of introns/100 aa[2]	1.24	1.10	0.55	1.23	1.49
Avg. exon size in bp[3]	210	240	410	200	280
Avg. UTR size in bp[4]	370	340	800	260	NA
Avg. intron size in bp[5]	170	770	660	290	2,800
Proportion of long introns[6]	10%	36%	27%	33%	85%
Avg. intergenic size in bp	4,000	21,600	5,400	2,400	78,000
Total fraction TE[7]	8.8%	1%	5.3%	NA	38.5%
Number of STRs[8]	65,211	188,101	58,808	13,617	1,562,965
Avg. STR length in bp	19.2	23.7	30.2	26.4	35.7

[1] Numbers in parentheses indicate euchromatic genome size.

[2] "aa" abbreviates amino acid. Calculated from NCBI's genomes mapview data sets

[3] Distribution for *D. melanogaster* is strongly bimodal.

[4] UTR size is biased by counting cases where length = 0 bp.

[5] Intron size is non-normally distributed. The distributions in all species except *D. pulex* are bimodal.

[6] Proportion of the number of introns that is larger than average exon size. See Figure S13.

[7] "TE" abbreviates transposable elements. References for TE statistics are listed in Table S17.

[8] Short Tandem Repeat (microsatellite) loci [S180].

Table S20. Species used in the study of introns. Abbreviations are used in Figure S15.

Species	Abbreviation	Source
Daphnia pulex	Dappu	http://genome.jgi-psf.org/Dappu1/Dappu1.home.html
Aedes aegypti	Aedae	http://www.vectorbase.org/
Anopheles gambiae	Anoga	http://www.ncbi.nlm.nih.gov/
Apis mellifera	Apime	http://www.ncbi.nlm.nih.gov/
Drosophila melanogaster	Drome	http://www.ncbi.nlm.nih.gov/
Drosophila pseudoobscura	Drops	ftp://ftp.flybase.net/genomes/
Nematostella vectensis	Nemve	http://genome.jgi-psf.org/Nemve1/Nemve1.home.html
Danio rerio	Danre	http://www.ncbi.nlm.nih.gov/
Homo sapiens	Homsa	http://www.ncbi.nlm.nih.gov/

Table S21. Number and density (per 100 amino acids) of introns for nine species are calculated by dividing the number of introns present by the number of total amino acids (residues) in the proteins for all proteins in orthologous sets. *Daphnia pulex* has the greatest intron density among the arthropods, followed by *Apis mellifera*, for which genomic data are currently available, but a significantly lower intron density than that in vertebrates and, especially, in the only available cnidarian.

Species	Residue	Introns	Density	Rank
Daphnia pulex	1,409,089	18,485	1.311	1
Apis mellifera	1,681,706	18,827	1.119	2
Anopheles gambiae	1,465,363	10,590	0.722	3
Aedes aegypti	1,619,969	10,482	0.647	4
Drosophila pseudoobscura	1,801,498	11,084	0.615	5
Drosophila melanogaster	1,846,871	10,594	0.573	6
Homo sapiens	1,770,781	32,535	1.837	-
Danio rerio	1,638,418	30,674	1.872	-
Nematostella vectensis	1,358,638	26,604	1.958	-

Table S22. Conservation of *Daphnia pulex* introns. A conserved intron is one whose position is shared by orthologous genes from at least two of the animal species listed in the table.

Species	Conserved introns	Variable introns	% conserved	Rank
Drosophila melanogaster	1,300	4,652	21.84	4
Drosophila pseudoobscura	1,277	4,675	21.45	5
Anopheles gambiae	1,418	4,534	23.82	3
Aedes aegypti	1,440	4,512	24.19	2
Apis mellifera	2,882	3,070	48.42	1
Homo sapiens	3,411	2,541	57.31	-
Danio rerio	3,392	2,560	56.99	-
Nematostella vectensis	3,213	2,739	53.98	-

Table S23. Conservation of intron positions between *Daphnia pulex* and other animals. The table shows the percentage and the raw numbers (in parentheses) of shared intron positions in a set of 9 animal genomes including *D. pulex* for all pairs of annotated orthologous protein-coding genes (above the diagonal) and for pairs of orthologous genes confirmed with ESTs (at least one *D. pulex* EST per gene; below the diagonal). Abbreviations are given in Table S20.

	Dappu	Drome	Drops	Anoga	Aedae	Apime	Homsa	Danre	Nemve
Dappu	-	31.93 (2600)	31.48 (2554)	34.47 (2836)	34.63 (2880)	55.54 (5764)	47.85 (6822)	47.26 (6784)	43.25 (6426)
Drome	32.61 (2156)	-	94.99 (4134)	58.28 (2604)	58.94 (2686)	38.10 (2522)	25.00 (2624)	24.79 (2626)	22.01 (2442)
Drops	32.10 (2116)	95.15 (3338)	-	58.25 (2584)	58.70 (2656)	37.62 (2478)	24.77 (2592)	24.58 (2596)	21.78 (2410)
Anoga	35.50 (2378)	58.59 (2118)	58.41 (2100)	-	88.77 (4120)	40.49 (2714)	26.75 (2830)	26.69 (2850)	23.88 (2670)
Aedae	35.52 (2402)	59.43 (2186)	59.10 (2162)	88.82 (3344)	-	40.58 (2756)	27.18 (2900)	26.99 (2906)	23.89 (2692)
Apime	56.59 (4768)	38.87 (2076)	38.34 (2040)	41.27 (2240)	41.23 (2264)	-	45.28 (5764)	44.76 (5742)	40.26 (5368)
Homsa	48.65 (5628)	25.41 (2156)	25.09 (2124)	27.32 (2342)	27.68 (2390)	45.62 (4698)	-	94.88 (15850)	73.35 (12622)
Danre	48.26 (5628)	25.25 (2166)	24.96 (2136)	27.31 (2366)	27.50 (2400)	45.29 (4706)	95.32 (12900)	-	72.54 (12554)
Nemve	44.42 (5380)	22.51 (2032)	22.25 (2004)	24.51 (2234)	24.56 (2254)	41.09 (4454)	73.55 (10284)	73.10 (10290)	

Table S24. Maximum Likelihood reconstruction of intron gain and loss events in arthropods and three other metazoans.

Node	No. introns	No. losses	No. gains	Gain/loss ratio
Metazoa	N/A	N/A	N/A	N/A
Coelomata	8,162	N/A	N/A	N/A
Arthropoda	5,163	3,666	667	0.18
Insecta	4,396	767	0	0
Vertebrata	8,367	586	791	1.35
Diptera	2,918	2,033	555	0.27
Drosophilidae	2,207	997	286	0.29
Culicidae	2,408	714	204	0.29
Daphnia pulex	5,952	1,047	1,836	1.75
Drosophila melanogaster	2,192	59	44	0.75
Drosophila pseudoobscura	2,160	83	36	0.43
Anopheles gambiae	2,276	208	76	0.37
Aedes aegypti	2,365	153	111	0.73
Apis mellifera	4,427	874	905	1.04
Homo sapiens	8,304	192	129	0.67
Danio rerio	8,402	257	292	1.14
Nematostella vectensis	8,905	N/A	N/A	N/A

E. Origin and Preservation of *Daphnia pulex* Genes

Table S25. Similarity of *Daphnia pulex* genes and 12 other genome-sequenced arthropods to human and other model eukaryote reference proteins. Reference proteins are all UniProt-SwissProt curated entries of 6 model species, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae*, accessed on 2010 January from www.uniprot.org. Arthropod proteome sets are current as of December 2009 [S54]. BlastP of SwissProt reference proteins to arthropod proteins is used with cutoff at e-value = $1e^{-5}$. Results are summarized to indicate closest arthropod matches to reference proteins in 4 ways: **A.** Counts of closest matching proteins. **B.** Alignment to reference proteins (average and sum of aminos) **C.** Percent of reference proteins found (of any found genes) **D.** Summary from other orthology assessments. *Daphnia pulex* has best matches and longest alignments to all non-arthropod gene sets, and *Tribolium castaneum* has the longest of the insects. *Daphnia pulex* has significantly greater best matches to proteins than *Tribolium castaneum* (**A**, p-value < $1e^{-15}$ using Chi-square). *Daphnia pulex* has statistically longer alignments than *Tribolium castaneum* to each reference species, whether non-matched genes are included or the subset where both species have reference gene matches. The Wilcoxon rank order test for paired ortholog genes measures this, with p-value < $1e^{-3}$ for human genes, and p-value < $1e^{-5}$ for the other non-arthropod models (**B**). Similarly 1% more human and 1% to 5% more non-arthropod genes are found in *Daphnia pulex* than *Tribolium castaneum* or others (**C**, p-value < $1e^{-15}$). Related studies have compared arthropod genes to reference proteins with similar results (**D**). Using phylogenetic orthology methods with alignment and tree construction, Phylomedb [S181] and PHiGs [S45] both find *Daphnia* > *Tribolium* > other insects for alignment to human genes. *Ixodes scapularis* genes have a high proportion of best matches (**A**), but are poorer overall matches (**B, C**). *Ixodes scapularis* proteins are shorter than expected and missing many expected orthologs, possibly an artifact of a fragmented genome assembly.

A. Counts of best match to reference proteins

Arthropod	*Arabidopsis*	*Caenorhabditis*	*Drosophila*	*Homo*	*Mus*	*Saccharomyces*
Daphnia pulex	1,004*	573*	0	3,286*	2,849*	714*
Ixodes scapularis	447	279	0	2,465	2,180	322
Tribolium castaneum	524	283	8	1,969	1,707	403
Apis mellifera	482	235	2	1,724	1,486	318
Nasonia vitripennis	506	249	7	1,606	1,412	374
Pediculus humanus	410	204	0	1,593	1,352	282
Acyrthosiphon pisum	496	166	2	1,286	763	266
Aedes aegypti	291	122	1	696	560	202
Anopheles gambiae	282	135	0	622	550	195
Drosophila melanogaster	330	134	2925	563	463	220
D. mojavensis	350	137	30	514	469	193
Culex quinquefasciatus	253	103	3	410	368	156
D. pseudoobscura	243	104	54	383	314	171
Reference	*Arabidopsis*	*Caenorhabditis*	*Drosophila*	*Homo*	*Mus*	*Saccharomyces*
Ref_found	5,029	2,492	3,035	1,5345	1,3004	3,575
Ref_input	8,823	3,278	3,052	2,0276	1,6214	6,912

* p-value < $1e^{-15}$ for *Daphnia pulex* vs *Tribolium castaneum*

B. Alignment to reference proteins, average aligned amino acids / protein.

Arthropod	*Arabidopsis*	*Caenorhabditis*	*Drosophila*	*Homo*	*Mus*	*Saccharomyces*	Mean
Daphnia pulex	130**	186**	216	188*	191**	149**	169**
Tribolium castaneum	126	181	250	187	188	147	166
Nasonia vitripennis	125	179	239	183	185	145	163
Apis mellifera	123	178	239	184	186	141	162
Pediculus humanus	123	177	231	184	185	139	162
Drosophila melanogaster	126	181	586	178	180	141	161
Drosophila mojavensis	125	177	427	175	177	142	159
Anopheles gambiae	123	178	278	177	179	139	159
Aedes aegypti	123	176	271	174	176	140	158
Drosophila pseudoobscura	124	175	454	173	175	140	157
Acyrthosiphon pisum	124	172	216	172	172	139	156
Culex quinquefasciatus	119	168	254	165	167	133	150
Ixodes scapularis	109	154	165	157	161	117	140
Mean	123	176	294	177	179	139	159

** p-value < $1e^{-5}$; * p-value < $1e^{-3}$ for *Daphnia pulex* vs *Tribolium castaneum*. Mean column excludes *Drosophila melanogaster*

C. Percent of reference proteins found (blastp cut-off $1e^{-5}$)

Arthropod	*Arabidopsis*	*Caenorhabditis*	*Drosophila*	*Homo*	*Mus*	*Saccharomyces*	Mean
Daphnia pulex	88.1*	94.9*	83.9	90.4*	91.5*	90.4*	90.3*
Tribolium castaneum	85.7	93.3	90.0	89.7	90.7	85.9	89.5
Nasonia vitripennis	85.7	93.0	88.7	88.9	89.8	86.0	88.8
Apis mellifera	83.5	92.7	88.8	88.9	90.1	86.0	88.6
Drosophila melanogaster	86.9	93.1	99.2	87.6	88.7	84.7	88.7
Anopheles gambiae	85.1	92.2	90.7	87.7	88.8	83.9	87.9
Drosophila pseudoobscura	85.8	92.2	97.5	86.8	87.9	84.5	87.9
Aedes aegypti	84.0	92.2	90.8	87.6	88.7	84.3	87.7
Pediculus humanus	81.7	92.3	86.4	89.0	89.8	83.7	87.9
Drosophila mojavensis	85.0	92.1	96.6	86.8	87.9	84.6	87.8
Acyrthosiphon pisum	85.3	91.2	84.9	86.2	87.2	83.5	86.4
Culex quinquefasciatus	83.5	91.3	90.4	86.9	88.0	82.1	86.9
Ixodes scapularis	80.1	90.5	79.1	87.7	88.8	77.8	85.8
Mean	84.6	92.4	89.8	88.0	89.1	84.4	87.8
Ref_found	5029	2492	3035	15345	13004	3575	

* p-value < $1e^{-15}$ for *Daphnia* pulex vs *Tribolium castaneum*. Mean column excludes *Drosophila melanogaster*

D. Other orthology assessments, best match to human genes count

Phylomedb results (*Acyrthosiphon pisum* analysis) are for human gene trees with all of 6 arthropod species, n = 6,281. This set is produced only for gene families including *Acyrthosiphon pisum*, so only groups having all 6 arthropods are counted here. PhIGs results (s50.3, 2007 data) for human gene trees with at least 1 of 4 arthropod species, n = 14,818, using an early *Tribolium castaneum* gene subset (~ 1/2 current).

	Phylomedb			**PHiGs**	
Arthropod	Human	% Best	Arthropod	Human	% Best
Daphnia	2,888	46	*Daphnia*	9,156	61
Tribolium	1,324	21	*Tribolium*	2,623	18
Pediculus	1,117	18	*Drosophila*	1,262	9
Acyrthosiphon	441	7	*Anopheles*	2,649	18
Drosophila	191	3			
Anopheles	320	5			

Table S26. Gene families in *Daphnia pulex* with and without recognizable InterPro protein domains that have expanded relative to gene families in insects. Statistically significant differences are marked in bold for *D. pulex* counts > insect counts with $p < 0.05$ based on 2,000 random permutations of exact probability. Others are groups with 2+ *Daphnia pulex* genes for 1-1 Insect genes. **iAve**, **iMax** are average, maximum other (insect) gene counts for the group. **G** is log-likelihood G-score (chi-square like) of abundance differences for all species. Results indicated that 483 orthologous gene families are overly-represented in *Daphnia* ($p < 0.05$). Based on iMax scores = 0, we count 379 (or 78%) expanded gene families that are unique to the *Daphnia* lineage. To test whether *Daphnia* duplicated genes are significantly biased towards genes without homologs, we compared the number of duplicates in 13 other arthropod genomes. The average frequency of unique duplicates is 0.104. The expected number of unique *Daphnia* duplicates is 1,503, thus giving the predicted total of 14,486 duplicate genes for the *Daphnia* genome. The observed number of lineage-specific duplicated genes in the *Daphnia* genome (2,326) is significantly greater than expected (χ^2 (df = 1) = 450.55, $p < 0.0001$).

Gene families that are found to have expanded independently among insects with an aquatic larval stage (mosquitoes) are indicated (¥). Gene sets were compared from within the genomes of 11 insects (*Acyrthosiphon pisum*, *Pediculus humanus*, *Aedes aegypti*, *Anopheles gambiae*, *Culex quinquefasciatus*, *Apis mellifera*, *Nasonia vitripennis*, *Tribolium castaneum*, *Drosophila melanogaster*, D. *pseudoobscura* and *D. mojavensis*), *Ixodes scapularis* and *D. pulex*. To find co-expanded gene families in the *Daphnia* and mosquito lineages, *D. pulex*, *A. aegypti*, *A. gambiae* and *C. quinquefasciatus* (plus *I. scapularis*) were removed from the calculation of the terrestrial insect species average, and then over-abundant gene groups were tabulated for these four taxa relative to terrestrial insects.

ARP2 gene group ID	No. of species	No. of genes	*Daphnia pulex* gene count	iAve	iMax	G	Description
G19	2	133	**132**	0	0	612	neurexin IV; src=ixodes_ISCW023368-PA
G24	1	123	**123**	0	0	570	hypothetical protein
G53	1	91	**91**	0	0	409	
G37	12	107	**89**	1	2	347	Alpha-1,3-fucosyltransferase; alpha1,3-fucosyltransferase b homologue; glycoprotein A
G49	10	92	**82**	1	2	333	hypothetical protein; cuticle protein; cpr50cb
G64	2	81	**80**	0	1	351	hypothetical protein
G67	2	80	**79**	0	1	346	
G78	1	75	**75**	0	0	329	hypothetical protein; jmjc domain-containing histone demethylation protein; kdm4a
G81	1	74	**74**	0	0	324	
G69	2	77	**73**	0	4	312	hypothetical protein; btb/poz domain-containing protein; mgc154338 protein
G83	1	73	**73**	0	0	319	
G105	2	64	**62**	0	2	261	
G79	14	74	**60**	1	2	248	denn domain-containing protein; tubulin-specific chaperone D
G110	3	62	**59**	0	2	244	
G113	1	59	**59**	0	0	250	hypothetical protein

G149	1	52	**52**	0	0	216	hypothetical protein
G121	2	58	**47**	1	11	195	
G180	1	47	**47**	0	0	192	hypothetical protein
G199	1	45	**45**	0	0	182	hypothetical protein
G200	1	45	**45**	0	0	182	
G232	1	40	**40**	0	0	159	
G233	1	40	**40**	0	0	159	hypothetical protein; spz; spaetzle-like cytokine
G94	12	69	**39**	2	8	119	pupal cuticle protein; hypothetical protein; edg78e
G254	1	38	**38**	0	0	149	hypothetical protein
G268	1	37	**37**	0	0	145	cathepsin l-like
G276	1	36	**36**	0	0	140	
G277	1	36	**36**	0	0	140	
G296	1	35	**35**	0	0	135	
G309	1	34	**34**	0	0	131	hypothetical protein; malate dehydrogenase
G310	1	34	**34**	0	0	131	hypothetical protein;
G328	1	33	**33**	0	0	126	
G329	1	33	**33**	0	0	126	
G330	1	33	**33**	0	0	126	lactosylceramide; alpha-lactosylceramide
G379	1	31	**31**	0	0	117	
G380	1	31	**31**	0	0	117	hypothetical protein; btb/poz domain-containing protein; mgc154338 protein
G406	1	30	**30**	0	0	112	
G97	5	67	**29**	3	22	137	hypothetical protein
G425	1	29	**29**	0	0	108	hypothetical protein
G426	1	29	**29**	0	0	108	hypothetical protein
G159	13	50	**27**	1	2	86	cral/trio domain-containing protein
G349	3	32	**27**	0	4	96	
G375	5	31	**27**	0	1	93	brain chitinase and chia; vegfr-a splice form a; tyrosine-protein kinase
G481	1	27	**27**	0	0	98	cytochrome p450
G482	1	27	**27**	0	0	98	hypothetical protein; malate dehydrogenase
G483	1	27	**27**	0	0	98	hypothetical protein
G484	1	27	**27**	0	0	98	hypothetical protein
G526	1	26	**26**	0	0	94	hypothetical protein
G527	1	26	**26**	0	0	94	
G27	13	119	**25**	8	18	57	glucosyl/glucuronosyl transferases; gustatory receptor; class b scavenger receptor cd36 domain
G404	6	30	**25**	0	1	83	F-box only protein 21; src=daphnia_NCBI_GNO_116234
G579	1	25	**25**	0	0	90	membrane glycoprotein lig-1
G580	1	25	**25**	0	0	90	hypothetical protein
G42	13	102	**23**	7	23	80	histone h3 type
G578	2	25	**23**	0	0	79	proclotting enzyme precursor; src=ixodes_ISCW000320-PA
G687	1	23	**23**	0	0	81	hypothetical protein
G689	1	23	**23**	0	0	81	trypsin alpha precursor
G690	1	23	**23**	0	0	81	hypothetical protein
G691	1	23	**23**	0	0	81	hypothetical protein
G686	2	23	**22**	0	1	75	ankyrin repeat protein
G762	1	22	**22**	0	0	76	hypothetical protein

G763	1	22	**22**	0	0	76	
G765	1	22	**22**	0	0	76	hypothetical protein
G766	1	22	**22**	0	0	76	
G139¥	13	54	**21**	3	6	47	class a rhodopsin-like g-protein coupled receptor gprop1
G842	1	21	**21**	0	0	72	hypothetical protein
G843	1	21	**21**	0	0	72	
G844	1	21	**21**	0	0	72	hypothetical protein
G845	1	21	**21**	0	0	72	
G846	1	21	**21**	0	0	72	bestrophin; bestrophin-2
G148	14	51	**20**	2	4	51	carbonic anhydrase; wd and tetratricopeptide repeats protein; cytoplasmic carbonic anhydrase
G227¥	14	40	**20**	2	3	58	conserved hypothetical protein; src=ixodes_ISCW009102-PA
G345	11	32	**20**	1	2	54	secreted protein; hypothetical protein
G760	3	22	**20**	0	1	65	transcriptional regulator ycf27
G988	1	20	**20**	0	0	68	hypothetical protein
G989	1	20	**20**	0	0	68	
G990	1	20	**20**	0	0	68	
G991	1	20	**20**	0	0	68	heat shock protein; inositol receptor
G992	1	20	**20**	0	0	68	clip-domain serine protease; lumbrokinase-31 precursor; clip-domain serine protease subfamily D
G269	14	36	**19**	1	2	55	chorion peroxidase precursor; peroxidase precursor; chorion peroxidase precursor ec contains chorion peroxidase light chain
G420	4	29	**19**	1	8	68	hypothetical protein; transposase; centromere protein B
G761	4	22	**19**	0	1	60	hypothetical protein; discoidin domain receptor; discoidin domain-containing receptor 2 precursor
G1166	1	19	**19**	0	0	63	hypothetical protein
G1167	1	19	**19**	0	0	63	
G1168	1	19	**19**	0	0	63	hypothetical protein
G1170	1	19	**19**	0	0	63	hypothetical protein
G1172	1	19	**19**	0	0	63	hypothetical protein
G1173	1	19	**19**	0	0	63	hypothetical protein; jmjc domain-containing histone demethylation protein; kdm4a
G1163	2	19	**18**	0	1	58	hypothetical protein; solute carrier family member a3; protein star
G1169	2	19	**18**	0	0	58	conserved hypothetical protein; src=ixodes_ISCW020111-PA
G1441	1	18	**18**	0	0	59	
G1442	1	18	**18**	0	0	59	hypothetical protein
G1443	1	18	**18**	0	0	59	hypothetical protein
G1444	1	18	**18**	0	0	59	hypothetical protein
G1445	1	18	**18**	0	0	59	hypothetical protein
G107	14	61	**17**	4	13	52	protease m1 zinc metalloprotease; alanyl aminopeptidase; aminopeptidase n precursor
G178	8	47	**17**	2	17	77	transposase; src=ixodes_ISCW007041-PA
G427	4	29	**17**	1	10	66	Gly d 3; src=daphnia_NCBI_GNO_158563
G1845	1	17	**17**	0	0	55	
G1846	1	17	**17**	0	0	55	hypothetical protein
G1849	1	17	**17**	0	0	55	
G1850	1	17	**17**	0	0	55	hypothetical protein

G1851	1	17	**17**	0	0	55	hypothetical protein
G1852	1	17	**17**	0	0	55	lactosylceramide
G1853	1	17	**17**	0	0	55	
G1854	1	17	**17**	0	0	55	
G1855	1	17	**17**	0	0	55	hypothetical protein; solute carrier family member a3; protein star
G1856	1	17	**17**	0	0	55	brain chitinase and chia; vegfr-a splice form a; tyrosine-protein kinase
G1857	1	17	**17**	0	0	55	hypothetical protein
G73	14	73	**16**	5	11	27	glucose dehydrogenase precursor
G164¥	14	47	**16**	2	5	36	high choriolytic enzyme; zinc metalloproteinase nas-15 precursor; meprin a subunit beta
G759	4	22	**16**	0	3	49	hypothetical protein; jmjc domain-containing histone demethylation protein; kdm4a
G1161	3	19	**16**	0	2	49	hypothetical protein
G1831	2	17	**16**	0	1	50	hypothetical protein
G1838	2	17	**16**	0	1	50	lactosylceramide
G1847	2	17	**16**	0	0	50	conserved hypothetical protein; src=ixodes_ISCW020342-PA
G2462	1	16	**16**	0	0	51	hypothetical protein
G2463	1	16	**16**	0	0	51	hypothetical protein
G2464	1	16	**16**	0	0	51	
G2465	1	16	**16**	0	0	51	hypothetical protein
G2466	1	16	**16**	0	0	51	di-domain hemoglobin precursor
G74	12	74	**15**	5	13	45	serine-type enodpeptidase; src=aedes_AAEL003060-PA
G193	14	43	**15**	2	4	30	abc transporter; atp-binding cassette sub-family a member; nod factor export atp-binding protein I
G207	14	43	**15**	2	4	30	dna-directed rna polymerase II largest subunit
G306	13	34	**15**	1	4	33	gastric triacylglycerol lipase precursor; lipase 1 precursor; lysosomal acid lipase
G510	11	26	**15**	1	2	36	hypothetical protein
G2461	2	16	**15**	0	0	46	hypothetical protein
G3483	1	15	**15**	0	0	46	hypothetical protein
G3484	1	15	**15**	0	0	46	hypothetical protein
G3485	1	15	**15**	0	0	46	hypothetical protein
G3486	1	15	**15**	0	0	46	hypothetical protein
G3487	1	15	**15**	0	0	46	
G3488	1	15	**15**	0	0	46	hypothetical protein; transposase; centromere protein B
G3489	1	15	**15**	0	0	46	
G3490	1	15	**15**	0	0	46	hypothetical protein
G3491	1	15	**15**	0	0	46	clip-domain serine protease; lumbrokinase-31 precursor; clip-domain serine protease subfamily D
G3493	1	15	**15**	0	0	46	glucosyl/glucuronosyl transferases; gustatory receptor; class b scavenger receptor cd36 domain
G688	3	23	**14**	0	3	48	conserved hypothetical protein; src=ixodes_ISCW004589-PA
G1836	4	17	**14**	0	1	40	conserved hypothetical protein; src=culex_CPIJ016633
G3469	2	15	**14**	0	1	42	
G3476	2	15	**14**	0	1	42	hypothetical protein
G3492	2	15	**14**	0	1	42	Hypothetical protein
G4919	1	14	**14**	0	0	42	hypothetical protein

G4920	1	14	**14**	0	0	42	
G4921	1	14	**14**	0	0	42	r2d2; tar rna binding protein
G4922	1	14	**14**	0	0	42	hypothetical protein
G4923	1	14	**14**	0	0	42	
G4924	1	14	**14**	0	0	42	tudor domain-containing protein
G4925	1	14	**14**	0	0	42	hypothetical protein
G4926	1	14	**14**	0	0	42	hypothetical protein
G4927	1	14	**14**	0	0	42	hypothetical protein
G187	14	45	**13**	3	4	22	bumetanide-sensitive na-k-cl cotransport protein
G219	13	41	**13**	2	5	25	hypothetical protein; cytochrome p450 cyp15a1; cyp304a1
G229¥	14	32	**13**	2	4	27	tribolium castaneum heat shock protein
G324	13	33	**13**	2	3	25	lactosylceramide; alpha-lactosylceramide
G343	12	32	**13**	2	3	28	pancreatic triacylglycerol lipase; ves g 1 allergen precursor; pancreatic lipase related protein 1
G441	13	27	**13**	1	2	27	class b secretin-like g-protein coupled receptor gprmth4; class b secretin-like g-protein coupled receptor gprmth1; class b secretin-like g-protein coupled receptor gprmth3
G2457	3	16	**13**	0	1	38	abc transporter; atp-binding cassette sub-family a member; nod factor export atp-binding protein I
G5963	1	13	**13**	0	0	38	hypothetical protein
G5964	1	13	**13**	0	0	38	hypothetical protein
G5965	1	13	**13**	0	0	38	hypothetical protein
G5966	1	13	**13**	0	0	38	hypothetical protein
G5967	1	13	**13**	0	0	38	
G5968	1	13	**13**	0	0	38	hypothetical protein
G5969	1	13	**13**	0	0	38	hypothetical protein
G5970	1	13	**13**	0	0	38	
G5971	1	13	**13**	0	0	38	hypothetical protein
G5972	1	13	**13**	0	0	38	hypothetical protein
G5973	1	13	**13**	0	0	38	hypothetical protein
G5974	1	13	**13**	0	0	38	hypothetical protein
G5975	1	13	**13**	0	0	38	hypothetical protein
G5976	1	13	**13**	0	0	38	hypothetical protein
G5977	1	13	**13**	0	0	38	hypothetical protein
G212	13	40	**12**	2	4	22	cral/trio domain-containing protein
G299	13	33	**12**	2	3	22	amp dependent coa ligase; acyl-coa synthetase
G376	5	31	**12**	2	15	62	hypothetical protein; mariner transposase; set domain and marinertransposase fusion
G3465	3	15	**12**	0	2	34	polyprotein; hypothetical protein; hypothetical protein k02a2.6
G3470	2	15	**12**	0	3	36	hypothetical protein
G5959	2	13	**12**	0	1	34	hypothetical protein
G5962	2	13	**12**	0	1	34	c-type lectin ctl - mannose binding.; serine protease; c-type lectin ctl - mannose binding. transcript A
G6719	1	12	**12**	0	0	34	hypothetical protein;
G6720	1	12	**12**	0	0	34	denn domain-containing protein; tubulin-specific chaperone D
G6721	1	12	**12**	0	0	34	ubiquitin-protein e3 ligase; hypothetical protein
G6722	1	12	**12**	0	0	34	hypothetical protein

G6723	1	12	**12**	0	0	34	hypothetical protein
G6724	1	12	**12**	0	0	34	hypothetical protein; mariner transposase; set domain and marinertransposase fusion
G6725	1	12	**12**	0	0	34	hypothetical protein
G6727	1	12	**12**	0	0	34	
G6728	1	12	**12**	0	0	34	hypothetical protein
G6729	1	12	**12**	0	0	34	denn domain-containing protein; tubulin-specific chaperone D
G6730	1	12	**12**	0	0	34	
G6731	1	12	**12**	0	0	34	
G6732	1	12	**12**	0	0	34	hypothetical protein
G6734	1	12	**12**	0	0	34	
G6735	1	12	**12**	0	0	34	hypothetical protein
G6736	1	12	**12**	0	0	34	
G6737	1	12	**12**	0	0	34	
G6738	1	12	**12**	0	0	34	hypothetical protein
G6739	1	12	**12**	0	0	34	
G6740	1	12	**12**	0	0	34	hypothetical protein
G6741	1	12	**12**	0	0	34	rna-binding protein precursor; hypothetical protein; rna-binding protein
G192	13	45	**11**	3	5	19	zinc carboxypeptidase; zinc carboxypeptidase a; zinc carboxypeptidase a 1 precursor
G246	14	36	**11**	2	3	19	atp-binding cassette sub-family g member; abc transporter
G671	13	23	**11**	1	1	21	queuine tRNA-ribosyltransferase; src=culex_CPIJ003941
G1848	2	17	**11**	0	0	38	sulfotransferase sult; bile salt sulfotransferase; hypothetical protein
G4910	3	14	**11**	0	2	30	hypothetical protein
G6718	2	12	**11**	0	1	30	hypothetical protein
G7291	1	11	**11**	0	0	31	
G7292	1	11	**11**	0	0	31	
G7293	1	11	**11**	0	0	31	
G7294	1	11	**11**	0	0	31	
G7295	1	11	**11**	0	0	31	hypothetical protein
G7296	1	11	**11**	0	0	31	hypothetical protein
G7297	1	11	**11**	0	0	31	hypothetical protein
G7298	1	11	**11**	0	0	31	
G7299	1	11	**11**	0	0	31	hypothetical protein
G7300	1	11	**11**	0	0	31	hypothetical protein
G7302	1	11	**11**	0	0	31	hypothetical protein
G7303	1	11	**11**	0	0	31	
G7304	1	11	**11**	0	0	31	hypothetical protein
G7305	1	11	**11**	0	0	31	hypothetical protein
G7306	1	11	**11**	0	0	31	hypothetical protein
G301¥	12	28	**10**	1	3	19	receptor-type tyrosine-protein phosphatase alpha precursor; hypothetical protein; roundabout
G764	2	22	**10**	1	12	51	hypothetical protein
G1223	7	18	**10**	1	2	24	hypothetical protein
G6712	3	12	**10**	0	1	26	serine/threonine-protein kinase mph1

G7267	2	11	**10**	0	1	27	hypothetical protein
G7752	1	10	**10**	0	0	27	hypothetical protein
G7753	1	10	**10**	0	0	27	hypothetical protein
G7754	1	10	**10**	0	0	27	
G7755	1	10	**10**	0	0	27	hypothetical protein
G7756	1	10	**10**	0	0	27	hypothetical protein
G7757	1	10	**10**	0	0	27	sulfate transporter
G7758	1	10	**10**	0	0	27	hypothetical protein
G7759	1	10	**10**	0	0	27	hypothetical protein
G7760	1	10	**10**	0	0	27	hypothetical protein
G7761	1	10	**10**	0	0	27	hypothetical protein
G7762	1	10	**10**	0	0	27	hypothetical protein
G7763	1	10	**10**	0	0	27	hypothetical protein
G7764	1	10	**10**	0	0	27	hypothetical protein
G7765	1	10	**10**	0	0	27	hypothetical protein
G7766	1	10	**10**	0	0	27	hypothetical protein
G7767	1	10	**10**	0	0	27	hypothetical protein
G7768	1	10	**10**	0	0	27	
G7769	1	10	**10**	0	0	27	hypothetical protein
G7771	1	10	**10**	0	0	27	hypothetical protein
G7772	1	10	**10**	0	0	27	
G7774	1	10	**10**	0	0	27	brain chitinase and chia; vegfr-a splice form a; tyrosine-protein kinase
G7775	1	10	**10**	0	0	27	hypothetical protein
G7776	1	10	**10**	0	0	27	hypothetical protein
G7777	1	10	**10**	0	0	27	hypothetical protein
G271	13	32	**9**	2	5	21	glutathione s-transferase; glutathione s-transferase ec class-sigma
G600	11	19	**9**	1	1	17	timeless protein
G660	12	22	**9**	1	2	16	ldl receptor ligand-binding repeat bearing protein; hypothetical protein; pro-epidermal growth factor
G969	12	20	**9**	1	1	16	athalia rosae coleseed sawfly/abc membrane transporter
G1440	7	18	**9**	1	3	22	peritrophic membrane chitin binding protein
G7773	2	10	**9**	0	1	23	neutral endopeptidase
G8296	1	9	**9**	0	0	23	
G8297	1	9	**9**	0	0	23	hypothetical protein
G8298	1	9	**9**	0	0	23	
G8299	1	9	**9**	0	0	23	hypothetical protein
G8300	1	9	**9**	0	0	23	
G8301	1	9	**9**	0	0	23	hypothetical protein
G8302	1	9	**9**	0	0	23	hypothetical protein
G8303	1	9	**9**	0	0	23	hypothetical protein
G8304	1	9	**9**	0	0	23	hypothetical protein
G8305	1	9	**9**	0	0	23	
G8306	1	9	**9**	0	0	23	hypothetical protein
G8307	1	9	**9**	0	0	23	hypothetical protein
G8308	1	9	**9**	0	0	23	
G8309	1	9	**9**	0	0	23	hypothetical protein

G8310	1	9	**9**	0	0	23	hypothetical protein
G8311	1	9	**9**	0	0	23	hypothetical protein
G8312	1	9	**9**	0	0	23	hypothetical protein
G8313	1	9	**9**	0	0	23	chromosome 7 scaf14703
G8314	1	9	**9**	0	0	23	hypothetical protein
G8315	1	9	**9**	0	0	23	hypothetical protein
G8316	1	9	**9**	0	0	23	c-type lectin ctl - mannose binding.; serine protease; c-type lectin ctl - mannose binding. transcript A
G8317	1	9	**9**	0	0	23	hypothetical protein
G8318	1	9	**9**	0	0	23	hypothetical protein
G8319	1	9	**9**	0	0	23	hypothetical protein
G8320	1	9	**9**	0	0	23	hypothetical protein
G8321	1	9	**9**	0	0	23	hypothetical protein
G8322	1	9	**9**	0	0	23	hypothetical protein
G8323	1	9	**9**	0	0	23	
G8324	1	9	**9**	0	0	23	hypothetical protein
G8325	1	9	**9**	0	0	23	hypothetical protein
G8326	1	9	**9**	0	0	23	hypothetical protein
G8327	1	9	**9**	0	0	23	hypothetical protein
G8328	1	9	**9**	0	0	23	chromosome 7 scaf14703
G8329	1	9	**9**	0	0	23	hypothetical protein
G8330	1	9	**9**	0	0	23	hypothetical protein
G8331	1	9	**9**	0	0	23	hypothetical protein
G8333	1	9	**9**	0	0	23	hypothetical protein; cuticular protein; structural constituent of cuticle
G8334	1	9	**9**	0	0	23	hypothetical protein
G8335	1	9	**9**	0	0	23	hypothetical protein
G8336	1	9	**9**	0	0	23	hypothetical protein
G8337	1	9	**9**	0	0	23	hypothetical protein
G166	13	49	**8**	3	10	21	cytochrome p450; corpora allata cytochrome p450; cyp4ac3
G302	14	32	**8**	2	2	11	acyl-coa-binding domain-containing protein; hypothetical protein; acyl-coa-binding protein
G604	14	24	**8**	1	3	17	transcription elongation factor spt6
G1422	6	18	**8**	1	6	25	hypothetical protein; transposase; centromere protein B
G1902	8	16	**8**	1	2	17	para-nitrobenzyl esterase
G6733	4	12	**8**	0	2	19	hypothetical protein
G7698	2	10	**8**	0	2	20	
G8215	2	9	**8**	0	1	19	hypothetical protein; transposase; centromere protein B
G8236	2	9	**8**	0	1	19	hypothetical protein
G8275	2	9	**8**	0	1	19	
G8287	2	9	**8**	0	1	19	hypothetical protein
G8295	2	9	**8**	0	1	19	
G8332	1	8	**8**	0	0	20	hypothetical protein
G8873	1	8	**8**	0	0	20	hypothetical protein
G8874	1	8	**8**	0	0	20	hypothetical protein
G8875	1	8	**8**	0	0	20	hypothetical protein
G8876	1	8	**8**	0	0	20	hypothetical protein
G8877	1	8	**8**	0	0	20	hypothetical protein

G8878	1	8	**8**	0	0	20	hypothetical protein
G8879	1	8	**8**	0	0	20	hypothetical protein
G8880	1	8	**8**	0	0	20	
G8881	1	8	**8**	0	0	20	hypothetical protein
G8882	1	8	**8**	0	0	20	
G8883	1	8	**8**	0	0	20	hypothetical protein
G8884	1	8	**8**	0	0	20	hypothetical protein; vitellogenin-1 precursor; hemelipoglycoprotein precursor
G8885	1	8	**8**	0	0	20	hypothetical protein
G8886	1	8	**8**	0	0	20	4 days neonate male adipose cdna
G8887	1	8	**8**	0	0	20	hypothetical protein
G8888	1	8	**8**	0	0	20	hypothetical protein
G8890	1	8	**8**	0	0	20	
G8891	1	8	**8**	0	0	20	hypothetical protein
G8892	1	8	**8**	0	0	20	hypothetical protein
G8893	1	8	**8**	0	0	20	
G8894	1	8	**8**	0	0	20	hypothetical protein
G8897	1	8	**8**	0	0	20	hypothetical protein
G8898	1	8	**8**	0	0	20	hypothetical protein
G8899	1	8	**8**	0	0	20	hypothetical protein
G8901	1	8	**8**	0	0	20	
G8902	1	8	**8**	0	0	20	hypothetical protein
G8903	1	8	**8**	0	0	20	sugar transporter; gastric caeca sugar transporter
G8904	1	8	**8**	0	0	20	
G8905	1	8	**8**	0	0	20	hypothetical protein
G650	14	23	**7**	1	2	13	beta-1,4-n-acetylgalactosaminyl transferase bre-4; beta-1,4-galactosyltransferase
G736	14	22	**7**	1	3	14	hypothetical protein
G779	14	21	**7**	1	1	13	regulator of g protein signaling
G825	14	21	**7**	1	2	13	zinc carboxypeptidase; zinc carboxypeptidase a; zinc carboxypeptidase a 1 precursor ec
G1339	12	18	**7**	1	1	11	low-density lipoprotein receptor ldl
G2467	4	16	**7**	1	6	24	lactosylceramide; alpha-lactosylceramide
G7261	4	11	**7**	0	2	16	hypothetical protein LOC100163706 ; src=aphid_ncbi_hmm240084
G7877	3	9	**7**	0	1	16	Lactosylceramide
G8808	2	8	**7**	0	1	16	hypothetical protein
G8850	2	8	**7**	0	1	16	hypothetical protein
G8853	2	8	**7**	0	1	16	hypothetical protein
G8889	2	8	**7**	0	0	16	hypothetical protein; src=ixodes_ISCW013637-PA
G9537	1	7	**7**	0	0	16	hypothetical protein
G9538	1	7	**7**	0	0	16	hypothetical protein
G9539	1	7	**7**	0	0	16	hypothetical protein
G9541	1	7	**7**	0	0	16	chitinase
G9542	1	7	**7**	0	0	16	vacuolar protein sorting
G9543	1	7	**7**	0	0	16	hypothetical protein
G9544	1	7	**7**	0	0	16	hypothetical protein
G9545	1	7	**7**	0	0	16	hypothetical protein

G9546	1	7	**7**	0	0	16	hypothetical protein
G9548	1	7	**7**	0	0	16	
G9549	1	7	**7**	0	0	16	bms1l protein
G9550	1	7	**7**	0	0	16	hypothetical protein
G9551	1	7	**7**	0	0	16	
G9552	1	7	**7**	0	0	16	hypothetical protein
G9553	1	7	**7**	0	0	16	hypothetical protein
G9554	1	7	**7**	0	0	16	hypothetical protein
G9555	1	7	**7**	0	0	16	hypothetical protein
G9556	1	7	**7**	0	0	16	hypothetical protein
G9557	1	7	**7**	0	0	16	hypothetical protein
G9558	1	7	**7**	0	0	16	
G9559	1	7	**7**	0	0	16	hypothetical protein
G9560	1	7	**7**	0	0	16	hypothetical protein
G9561	1	7	**7**	0	0	16	hypothetical protein
G9562	1	7	**7**	0	0	16	hypothetical protein
G9563	1	7	**7**	0	0	16	hypothetical protein
G9564	1	7	**7**	0	0	16	hypothetical protein
G9565	1	7	**7**	0	0	16	hypothetical protein
G9566	1	7	**7**	0	0	16	
G9567	1	7	**7**	0	0	16	
G9568	1	7	**7**	0	0	16	hypothetical protein
G9569	1	7	**7**	0	0	16	hypothetical protein
G9570	1	7	**7**	0	0	16	hypothetical protein
G9571	1	7	**7**	0	0	16	hypothetical protein
G9572	1	7	**7**	0	0	16	
G9573	1	7	**7**	0	0	16	hypothetical protein
G9574	1	7	**7**	0	0	16	abc transporter; atp-binding cassette sub-family a member; nod factor export atp-binding protein I
G9575	1	7	**7**	0	0	16	
G9576	1	7	**7**	0	0	16	hypothetical protein
G9577	1	7	**7**	0	0	16	hypothetical protein
G9578	1	7	**7**	0	0	16	hypothetical protein
G9579	1	7	**7**	0	0	16	hypothetical protein
G9580	1	7	**7**	0	0	16	hypothetical protein
G9581	1	7	**7**	0	0	16	
G9582	1	7	**7**	0	0	16	hypothetical protein
G9583	1	7	**7**	0	0	16	hypothetical protein
G9584	1	7	**7**	0	0	16	hypothetical protein
G9585	1	7	**7**	0	0	16	
G9586	1	7	**7**	0	0	16	hypothetical protein
G9587	1	7	**7**	0	0	16	hypothetical protein
G9588	1	7	**7**	0	0	16	glucosyl/glucuronosyl transferases; gustatory receptor; class b scavenger receptor cd36 domain
G9589	1	7	**7**	0	0	16	hypothetical protein
G9590	1	7	**7**	0	0	16	hypothetical protein
G9591	1	7	**7**	0	0	16	peroxinectin precursor
G9593	1	7	**7**	0	0	16	hypothetical protein

G9595	1	7	**7**	0	0	16	hypothetical protein
G9596	1	7	**7**	0	0	16	
G9598	1	7	**7**	0	0	16	hypothetical protein
G9599	1	7	**7**	0	0	16	hypothetical protein
G9600	1	7	**7**	0	0	16	hypothetical protein
G9601	1	7	**7**	0	0	16	
G9602	1	7	**7**	0	0	16	hypothetical protein
G9605	1	7	**7**	0	0	16	hypothetical protein
G394¥	14	28	**6**	2	3	9	scp-like extracellular protein; cysteine-rich venom protein; cysteine-rich secretory protein-2
G901	14	20	**6**	1	2	10	prolyl alpha-1 subunit precursor
G951	14	20	**6**	1	1	10	dna topoisomerase II
G1086	11	17	**6**	1	2	10	hypothetical protein
G1434	5	18	**6**	1	6	22	nfx1-type zinc finger-containing protein 1; nfx1-type zinc finger-containing protein; splicing endonuclease positive effector sen1
G5951	3	13	**6**	1	5	20	hypothetical protein
G10425	1	6	**6**	0	0	13	hypothetical protein
G10426	1	6	**6**	0	0	13	hypothetical protein
G10427	1	6	**6**	0	0	13	hypothetical protein
G10428	1	6	**6**	0	0	13	hypothetical protein
G10429	1	6	**6**	0	0	13	hypothetical protein
G10430	1	6	**6**	0	0	13	hypothetical protein
G10432	1	6	**6**	0	0	13	hypothetical protein
G10433	1	6	**6**	0	0	13	
G10434	1	6	**6**	0	0	13	
G10435	1	6	**6**	0	0	13	hypothetical protein
G10437	1	6	**6**	0	0	13	hypothetical protein
G10438	1	6	**6**	0	0	13	hypothetical protein
G10439	1	6	**6**	0	0	13	
G10440	1	6	**6**	0	0	13	atp-dependent rna helicase kurz
G10441	1	6	**6**	0	0	13	hypothetical protein
G10442	1	6	**6**	0	0	13	hypothetical protein
G10443	1	6	**6**	0	0	13	hypothetical protein
G10444	1	6	**6**	0	0	13	hypothetical protein
G10445	1	6	**6**	0	0	13	hypothetical protein
G10446	1	6	**6**	0	0	13	hypothetical protein
G10447	1	6	**6**	0	0	13	
G10448	1	6	**6**	0	0	13	hypothetical protein
G10449	1	6	**6**	0	0	13	hypothetical protein
G10450	1	6	**6**	0	0	13	hypothetical protein
G10451	1	6	**6**	0	0	13	
G10452	1	6	**6**	0	0	13	hypothetical protein
G10453	1	6	**6**	0	0	13	hypothetical protein
G10454	1	6	**6**	0	0	13	hypothetical protein
G10455	1	6	**6**	0	0	13	hypothetical protein
G10456	1	6	**6**	0	0	13	hypothetical protein
G10458	1	6	**6**	0	0	13	hypothetical protein

G10460	1	6	**6**	0	0	13	hypothetical protein
G10461	1	6	**6**	0	0	13	hypothetical protein
G10462	1	6	**6**	0	0	13	hypothetical protein
G10463	1	6	**6**	0	0	13	hypothetical protein
G10464	1	6	**6**	0	0	13	
G10466	1	6	**6**	0	0	13	hypothetical protein
G10467	1	6	**6**	0	0	13	hypothetical protein
G10468	1	6	**6**	0	0	13	hypothetical protein
G10469	1	6	**6**	0	0	13	hypothetical protein
G10470	1	6	**6**	0	0	13	
G10471	1	6	**6**	0	0	13	hypothetical protein
G10472	1	6	**6**	0	0	13	hypothetical protein
G10474	1	6	**6**	0	0	13	mannan endo-1
G10476	1	6	**6**	0	0	13	hypothetical protein
G10477	1	6	**6**	0	0	13	hypothetical protein
G10478	1	6	**6**	0	0	13	hypothetical protein
G10479	1	6	**6**	0	0	13	hypothetical protein
G10480	1	6	**6**	0	0	13	hypothetical protein; cuticle protein; cpr50cb
G10481	1	6	**6**	0	0	13	replication protein a; hypothetical protein
G10482	1	6	**6**	0	0	13	hypothetical protein
G10483	1	6	**6**	0	0	13	
G10484	1	6	**6**	0	0	13	hypothetical protein
G10485	1	6	**6**	0	0	13	hypothetical protein
G10486	1	6	**6**	0	0	13	hypothetical protein
G10487	1	6	**6**	0	0	13	hypothetical protein
G10488	1	6	**6**	0	0	13	hypothetical protein
G10489	1	6	**6**	0	0	13	hypothetical protein
G10490	1	6	**6**	0	0	13	
G10491	1	6	**6**	0	0	13	hypothetical protein; organic solute transporter alpha
G10492	1	6	**6**	0	0	13	hypothetical protein
G10494	1	6	**6**	0	0	13	hypothetical protein
G10496	1	6	**6**	0	0	13	hypothetical protein
G10497	1	6	**6**	0	0	13	hypothetical protein
G388	14	29	5	2	3	6	bombesin receptor subtype-3
G457	14	23	5	1	2	7	delta-9 desaturase 1; fatty acid desaturase; acyl-coa delta-9 desaturase
G588	14	24	5	2	2	7	transcriptional regulator atrx x-linked helicase ii; dna repair and recombination protein rad54b; lymphoid specific helicase
G589	14	23	5	1	2	7	n-ethylmaleimide sensitive fusion protein
G772	13	21	5	1	2	6	three prime repair exonuclease
G1054	13	19	5	1	2	6	dehydrogenase/reductase sdr family member
G1056	14	18	5	1	1	7	hypothetical protein; otopetrin
G1112	13	19	5	1	2	6	hypothetical protein
G1351	14	18	5	1	1	7	pre-mrna cleavage complex ii protein clp1
G1402	13	18	5	1	2	6	nucleoside-diphosphate kinase nbr-a
G208	14	34	4	2	3	4	myosin-rhogap protein; myosin heavy chain; glycyl-trna synthetase
G304	14	23	4	2	3	6	annexin x; annexin ix; anxb11

G336	14	27	4	2	3	7	hypothetical protein; phosphatidylinositol transfer protein sec14; cral/trio domain-containing protein
G391	14	19	4	1	2	5	chloride channel protein
G584¥	13	24	4	2	3	6	amp dependent coa ligase; acyl-coa synthetase
G739	13	22	4	1	2	6	carbohydrate sulfotransferase; hypothetical protein
G780	13	16	4	1	1	4	sodium/hydrogen exchanger 3 nhe3
G916	13	19	4	1	2	4	calcyphosine/tpp
G954	13	20	4	1	2	4	soluble guanylate cyclase; soluble guanylyl cyclase beta subunit
G955	14	20	4	1	2	5	valacyclovir hydrolase; serine hydrolase-like
G1002	13	17	4	1	2	4	gamma-glutamyl hydrolase precursor
G1032	14	19	4	1	2	5	peroxisomal isomerase
G1198	13	17	4	1	2	5	fumarylacetoacetate hydrolase domain-containing protein
G1205	12	15	4	1	1	4	deoxythymidylate kinase thymidylate kinase
G1209	14	17	4	1	1	4	chromosome region maintenance protein
G1225	14	18	4	1	2	5	geranylgeranyl pyrophosphate synthase/polyprenyl synthetase
G1254	13	18	4	1	1	5	hypothetical protein
G1261	14	18	4	1	2	5	sa
G1272	14	18	4	1	2	5	hypothetical protein
G1277	14	18	4	1	2	5	delta-1-pyrroline-5-carboxylate dehydrogenase
G1315	14	17	4	1	1	4	phosphatidate phosphatase
G1403	14	18	4	1	2	5	short-chain dehydrogenase
G1421	13	17	4	1	2	4	ribonuclease h1; ribonuclease H
G1482	14	15	4	1	1	4	hypothetical protein
G1553	13	17	4	1	2	4	hypothetical protein
G1642	14	17	4	1	1	4	hypothetical protein
G1745	14	16	4	1	1	4	integrator complex subunit
G2044	13	16	4	1	1	4	clip-domain serine protease; lumbrokinase-31 precursor; clip-domain serine protease subfamily D
G2116	13	16	4	1	1	4	hypothetical protein
G2210	12	15	4	1	1	4	glucosyl/glucuronosyl transferases; gustatory receptor; class b scavenger receptor cd36 domain
G2334	13	16	4	1	1	4	mrna-capping-enzyme; nadh-ubiquinone oxidoreductase flavoprotein 1 ndufv1
G2582	12	15	4	1	1	4	hypothetical protein
G2848	12	15	4	1	1	4	hypothetical protein
G3076	12	15	4	1	1	4	hypothetical protein
G4690	11	14	4	1	1	5	sodium-dependent phosphate transporter
G535	14	15	3	1	1	2	methionine-r-sulfoxide reductase
G606	13	22	3	1	2	3	potassium voltage-gated channel protein shaw shaw2; voltage-gated potassium channel
G611¥	14	21	3	1	3	4	phospholipid hydroperoxide glutathione peroxidase
G618	14	17	3	1	2	3	steroid dehydrogenase; hydroxysteroid dehydrogenase
G653	14	19	3	1	2	3	adenylsulfate kinase
G711¥	14	20	3	1	3	4	hypothetical protein
G728	13	15	3	1	1	2	possible integral membrane efflux protein efpa
G733	14	18	3	1	2	3	long chain fatty acid coa-ligase
G823	13	21	3	1	2	9	hypothetical protein

G878	14	20	3	1	2	3	dual specificity protein phosphatase; jnk stimulatory phosphatase jsp1; dual-specificity protein phosphatase
G882	14	16	3	1	1	2	glycerol-3-phosphate dehydrogenase
G911	13	19	3	1	2	4	apolipoprotein d precursor; apolipoprotein D
G957	14	18	3	1	3	4	lethal2essential for life protein; proteinlethal2essential for life protein efl21; heat shock protein
G1017	14	15	3	1	1	2	ion transport peptide precursor
G1040	14	18	3	1	2	3	sodium/nucleoside cotransporter
G1052	13	17	3	1	2	3	Pug
G1084	13	19	3	1	2	3	amine oxidase; peroxisomal n1-acetyl-spermine/spermidine oxidase; peroxisomal n1-acetyl-spermine/spermidine oxidase precursor
G1095	14	19	3	1	2	3	plasma alpha-l-fucosidase precursor
G1129	14	18	3	1	2	3	sodium/calcium exchanger
G1208	14	16	3	1	1	2	hypothetical protein
G1213	14	17	3	1	2	3	integrin alpha-ps; integrin alpha2 precursor position-specific antigen 2 alpha subunit protein inflated; integrin alpha1 precursor
G1235	13	16	3	1	2	2	hypothetical protein
G1265	12	15	3	1	2	3	chitooligosaccharidolytic beta-n-acetylglucosaminidase precursor; beta-hexosaminidase subunit beta precursor n-acetyl-beta-glucosaminidase subunit beta beta-n-acetylhexosaminidase subunit beta hexosaminidase subunit b; chitooligosaccharidolytic beta-n-acetylglucosaminidase
G1276	13	18	3	1	3	4	class b scavenger receptor cd36 domain. nb: previously described as scrb2; class b scavenger receptor cd36 domain
G1302	14	16	3	1	1	2	hypothetical protein
G1319	14	18	3	1	2	3	apis mellifera amt-2-like protein ,mrna; ammonium transporter iss; amt-1-like protein
G1373	14	17	3	1	2	3	fk506 binding protein; fk506-binding protein; fk506 binding protein fkbp
G1377	13	17	3	1	2	3	U4/U6.U5 tri-snRNP-associated protein; src=pediculus_PHUM534220-PA
G1387	14	18	3	1	2	3	intraflagellar transport homolog
G1395	13	16	3	1	2	2	l-lactate dehydrogenase
G1396	14	16	3	1	2	2	n6-adenosine-methyltransferase kda subunit
G1407	13	17	3	1	2	3	4-aminobutyrate aminotransferase
G1410	13	17	3	1	2	3	serine/threonine-protein kinase polo; hypothetical protein
G1473	12	15	3	1	2	3	leucine zipper-ef-hand-containing transmembrane protein
G1538	14	16	3	1	1	2	dna repair and recombination protein rad54b rad54 homolog b; solute carrier family glycerol-3-phosphate transporter; dna repair and recombination protein rad54
G1605	14	16	3	1	1	2	class a rhodopsin-like g-protein coupled receptor gprdop2
G1617	14	17	3	1	2	3	Tw
G1626	13	17	3	1	3	3	trehalose-6-phosphate synthase 1
G1653	14	17	3	1	2	3	dual oxidase: peroxidase and nadph-oxidase domains
G1661	14	16	3	1	1	2	nicotinamide mononucleotide adenylyltransferase
G1693	14	16	3	1	1	2	long-chain-fatty-acid-coa ligase
G1736	13	17	3	1	2	3	importin beta-4
G1795	14	16	3	1	1	2	wiskott-aldrich syndrome gene-like protein
G1799	13	17	3	1	2	3	ancient domain protein 2 cyclin m2

G1936	14	16	3	1	1	2	40s ribosomal protein s9
G1946	12	16	3	1	2	3	kinesin-like protein kif1b; kinesin heavy chain; hypothetical protein
G1975	14	16	3	1	1	2	phosphatidylinositol catalytic subunit alpha
G2006	14	16	3	1	1	2	hypothetical protein
G2025	14	16	3	1	1	2	wd repeat protein
G2035	14	16	3	1	1	2	retina aberrant in pattern; wd repeat-containing protein slp1
G2050	14	16	3	1	1	2	gtp-binding protein di-ras2
G2052	14	16	3	1	1	2	hypothetical protein
G2090	14	16	3	1	1	2	ufm1-conjugating enzyme 1 ubiquitin-fold modifier-conjugating enzyme 1
G2107	14	16	3	1	1	2	alcohol dehydrogenase class
G2125	14	16	3	1	1	2	endoribonuclease dcr-1; dicer-1
G2132	13	16	3	1	2	2	integrator complex subunit 7 int7;
G2142	13	15	3	1	1	2	branchiostoma peroxiredoxin v protein
G2143	14	16	3	1	1	2	hypothetical protein
G2184	14	16	3	1	1	2	5-aminolevulinic acid synthase
G2211	14	16	3	1	1	2	leucine-rich repeat serine/threonine-protein kinase
G2249	12	16	3	1	2	3	adenylate cyclase
G2264	14	16	3	1	1	2	tetratricopeptide repeat protein; o-linked n-acetylglucosamine transferase; sxc
G2361	11	16	3	1	2	5	phospholipid-transporting atpase
G2405	14	16	3	1	1	2	1-acyl-glycerol-3-phosphate acyltransferase
G2618	13	15	3	1	1	2	nadph oxidase
G2667	13	15	3	1	1	2	hypothetical protein
G2729	13	15	3	1	1	2	dna-directed rna polymerase iii subunit F
G2782	13	15	3	1	1	2	myo inositol monophosphatase
G2896	12	14	3	1	1	3	beta-1,3-galactosyltransferase
G3077	12	15	3	1	2	3	dna-directed rna polymerase iii subunit G
G3141	12	15	3	1	2	3	short-chain dehydrogenase
G3720	12	14	3	1	1	3	hypothetical protein
G3798	12	14	3	1	1	3	xaao aminopeptidase
G3931	12	14	3	1	1	3	name=CG6865-PA; parent=FBgn0036817; src=drosmel_CG6865-PA
G4720	12	14	3	1	1	3	karyopherin importin alpha

Table S27. Species used in the study of gene family expansions history (see Figure 1C).

Species Name	Source	File Name / Version	# of (predicted) genes
Daphnia pulex	JGI	Daphnia_FrozenGeneCatalog_2007_07_03.aa.fasta	30,940
Drosophila pseudoobscura	FlyBase	dpse-all-translation-r2.3.fasta	16,158
Drosophila melanogaster	Ref 5	http://insects.eugenes.org/arthropods/data/	13,738
Apis mellifera	NCBI Gnomon	http://insects.eugenes.org/arthropods/data/	17,182
Anopheles gambiae	Ensembl r50	Anopheles_gambiae.AgamP3.50.pep.all.fa *	12,457
Aedes aegypti	Ensembl r50	Aedes_aegypti.AaegL1.50.pep.all.fa *	15,419
Nematostella vectensis	JGI	proteins.Nemve1FilteredModels1.fasta	27,273
Homo sapiens	Ensembl r50	Homo_sapiens.NCBI36.50.pep.all.fa *	21,785
Danio rerio	Ensembl r50	Danio_rerio.ZFISH7.50.pep.all.fa *	21,322
Caenorhabditis elegans	Ensembl r50	Caenorhabditis_elegans.WS190.50.pep.all.fa *	20,176
Tribolium castaneum	Beetlebase rel3	http://insects.eugenes.org/arthropods/data/	16,422

Table S28. EvolMap reconstruction of gene gain and loss events in arthropods and four other metazoans. **Ancestor name** = the common ancestor of the species for a given row. **Sym-bets** = the number of symmetrical best alignments detected between the two descendants of the given node, as specified by the species phylogeny (Figure 1C). **Present loci** = the estimated number of genes present at the specified node, by accounting for gene families that were detected in earlier ancestors. **Loss** = the number of gene loss events estimated along the specified branch. **Paralogs** = the estimated number of duplication events along the branch, for genes having considerable sequence similarity with other members of the gene family within the same genome. **Diverged paralogs** = the number of genes that have duplicated and diverged more than the orthologous genes, and thus are assumed to have evolved under relaxed or positive selection after the gene duplication event. **Ambiguous gains** = the estimated number of genes originating at the specified branch that have no significant similarity to other gene families. **Total gains** = the sum of paralogs, diverged paralogs and ambiguous gains. **No scoring genes** is calculated only for each of the modern species = the number of genes that have no sequence similarity above a minimum threshold ($p > 10^{-4}$). **AVG and STD of sym-bet** = the average and standard deviation [S182] for the similarity estimates between orthologous members of the gene families, where a higher value indicates greater sequence conservation between the orthologous genes. **Abbreviations:** Anaga, *Anopheles gambiae*; Aedae, *Aedes aegypti*; Drome, *Drosophila melanogaster*; Drops, *Drosophila pseudoobscura*; Apime, *Apis mellifera*; cele, *Caenorhabditis elegans*; Dappu, *Daphnia pulex*; Homsa, *Homo sapiens*; Danre, *Danio rerio*; Nemve, *Nematostella vectensis*; Trica, *Tribolium castaneum.*

Ancestor name	Sym-bets	Present loci	Loss	Paralogs	Diverged Paralogs	Ambiguous gains	Total Gains	No scoring genes	AVG Sym-bet	STD Sym-bet
homsa; danre; cele; dappu; apime; trica; aedae; anoga; drome; drops; nemve	7,423									
homsa; danre; cele; dappu; apime; trica; aedae; anoga; drome; drops	6,764	8,679	17	86	877	310	1,273	0	538	104
homsa; danre	10,873	12,232	681	2,263	1,207	764	4,234	0	702	130
homsa	21,785	19,633	359	2,276	3,637	1,847	7,760	2,152		
danre	21,322	20,869	1,000	3,280	4,405	1,952	9,637	453		
cele; dappu; apime; trica; aedae; anoga; drome; drops	4,877	7,846	1,295	66	332	64	462	0	515	100
cele	20,176	15,762	2,643	2,050	2,634	5,875	10,559	4,414		
dappu; apime; trica; aedae; anoga; drome; drops	6,895	8,685	393	458	503	271	1,232	0	588	123
dappu	30,940	25,030	1,079	5,076	5,537	6,811	17,424	5,910		
apime; trica; aedae; anoga; drome; drops	7,698	9,161	756	96	819	317	1,232	0	602	122
apime	17,182	11,385	1,062	1,420	777	1,089	3,286	5,797		
trica; aedae; anoga; drome; drops	7,665	9,385	439	87	423	153	663	0	602	123
trica	16,422	12,839	914	1,824	1,566	978	4,368	3,583		
aedae; anoga; drome; drops	8,072	9,323	863	78	431	292	801	0	628	127
aedae; anoga	8,935	10,148	497	275	654	393	1,322	0	749	133
aedae	15,419	14,278	493	1,876	1,527	1,220	4,623	1,141		
anoga	12,457	11,438	720	642	707	661	2,010	1,019		
drome; drops	11,584	11,963	799	497	1,484	1,458	3,439	0	804	126
drome	13,738	13,002	196	258	560	417	1,235	736		
drops	16,158	14,626	183	728	835	1,283	2,846	1,532		
nemve	27,273	24,743	0	6,843	6,333	4,144	17,320	2,530		

Table S29. Gene duplication and duplicate gene birth rates in the *Daphnia pulex*, *Caenorhabditis elegans* and *Homo sapiens* genomes. The birth rates of gene duplicates were calculated using the number of single-pair duplicates in the youngest cohort (K_s < 0.01), the baseline number of single copy genes and the synonymous substitution rate (K_s), and giving units of duplications/gene/K_s. Birth rates are estimated by (Number of single pair duplicates < K_s .01)/(Number of single copy genes + Number of single pair duplicate gene pairs).

	Daphnia pulex	*Caenorhabditis elegans*	*Homo sapiens*
Single copy genes	16,285	13,768	15,002
Duplicate genes	14,655	6,350	7,678
Total genes	30,940	20,118	22680
Birth rate	0.0093	0.0033	0.0073

Table S30. Large fraction of *Daphnia pulex* duplicated genes. The large gene inventory is attributed to over 900 localized tandem gene duplication (TGD) clusters of 3 or more loci. Representative genomes are compared: *Drosophila melanogaster, Caenorhabditis elegans* and *Mus musculus*. The same method at identifying TGDs was applied to all species (see SOM). By using different criteria, Woollard [S183] reports 402 gene clusters for *Caenorhabditis elegans*, instead of 680 clusters by our measures.

	Total # duplicated genes	Total # 3+ tandem duplicated genes	Total # 3+ gene clusters
Daphnia pulex	13,972 / 28,093 (50%)	5,400 / 27,000 (20%)	919
Drosophila melanogaster	4,497 / 13,391 (34%)	1,500 / 13,500 (11%)	168
Caenorhabditis elegans	8,674 / 19,692 (44%)	3,000 / 20,000 (15%)	680
Mus musculus	10,244 / 18,871 (54%)		

Table S31. Gene families that are expanded and/or shared between *Daphnia pulex* and other aquatic (vertebrate) species compared to average differences found in terrestrial animals. Thirty-six eukaryotic genomes are compared by superfamily assignments [S83], including 18 invertebrates and 17 vertebrates of which 14 taxa are aquatic and 21 taxa are terrestrial. *Daphnia pulex* is the only invertebrate that exclusively lives in water and with a draft genome sequence data. Three gene families are expanded in the *D. pulex* genome and have significant aquatic versus terrestrial average differences (indicated by †), while the remaining 26 gene families have significant invertebrate versus vertebrate average differences. Significant ($p<0.05$) t-test results of root mean square deviation from expected gene count (genome × family) contingency table are listed between aquatic/terrestrial groups.

SuperFamily ID	Protein Domain	Aquatic Invertebrate Gene Count	Aquatic Vertebrate Gene Count	Terrestrial Invertebrate Gene Count	Terrestrial Vertebrate Gene Count	T Statistic	Degrees of Freedom	P-Value
sf51665†	Xylose isomerase	2.86	1.57	1.18	1	4.97	28	3.03E-05
sf10164	3 Thrombospondin C-terminal domain	8.71	10.29	2.18	6.8	4.24	15.8	0.0006475
sf52426	Cryptochrome/photolyase, N-terminal domain	1.86	2.86	1	1	3.62	18.9	0.001834
sf55528	Matrix metalloproteases, catalytic domain	19.86	27.14	4.18	24.9	3.45	26.3	0.001902
sf10364	8 TSP type-3 repeat	14	16.14	3.18	11.1	3.53	15.1	0.00297
sf55935	Guanido kinase catalytic domain	12.71	10.71	4.36	6.4	3.48	16.2	0.003031
sf82904	Noggin	1.86	3.71	1.09	2	3.34	21.1	0.003122
sf48035	Guanido kinase N-terminal domain	9	9.86	3.36	6.1	3.46	16.5	0.003139
sf81320†	Rhodopsin-like	24.3	41.9	11.8	17.1	3.44	16.5	0.003218
sf48174	Cryptochrome/photolyase FAD-binding domain	5	11	2.91	3.5	3.28	18.8	0.003981
sf52592	G proteins	162	291	104	240	3.36	14.6	0.004475
sf52769	Arginase-like amidino hydrolases	5.86	7.29	2.36	4.7	2.99	17.3	0.008169
sf53496	Prolyl oligopeptidase, C-terminal domain	3	3.86	1.73	2.7	2.66	32.8	0.01189
sf47502	Calmodulin-like	41	64.4	20.9	54	2.71	17.3	0.01472
sf10207	9 Putative alpha-L-fucosidase, catalytic domain	9.71	5.43	2.73	4	2.67	14.8	0.0177
sf11043	6 Ornithine cyclodeaminase-like (Pfam 02423)	2.57	2.14	1.18	1.8	2.51	24.2	0.01922
sf52468	Deoxyhypusine synthase, DHS	4.86	3.43	1.82	2.6	2.57	16.7	0.02022
sf63708	Ganglioside M2 (gm2) activator	2.14	2.57	1.09	2	2.46	18.3	0.02394
sf51557	Adenosine deaminase (ADA)	2.71	3.43	1.45	2.4	2.39	25.2	0.02447

sf64357	Synatpobrevin N-terminal domain	3.29	6.86	2.91	2.6	2.37	26.9	0.02543
sf53452†	beta 1,4 galactosyltransferase (b4GalT1)	14.86	12.71	5.73	11.3	2.39	20.1	0.0267
sf49266	Fibronectin type III	7.71	68	4	157.7	-2.32	28.1	0.02766
sf63608	Leukotriene A4 hydrolase C-terminal domain	3.14	2.57	1.36	2.1	2.34	20.2	0.02937
sf49600	TRAF domain	14	10.43	2.64	11	2.33	14.9	0.03421
sf46887	Methionine aminopeptidase, insert domain	3.14	3.71	1.55	3.3	2.2	31.7	0.03487
sf82283	Homocysteine S-methyltransferase	12.14	4.29	1.82	5.3	2.3	14.8	0.03649
sf81287	ML domain	2.86	2.86	1.55	1.9	2.15	21.2	0.04344
sf52002	R1 subunit of ribonucleotide reductase, C-terminal domain	2.86	3.43	1.82	2.3	2.1	26.8	0.04524
sf63984	Sir2 family of transcriptional regulators	8.29	9.29	4.73	7.8	2.07	24.1	0.04946

Table S32. Part A. Forty-six *Daphnia pulex* opsin genes belonging to 6 major clades. **Part B.** Additional Metazoan Opsins in Figure S21.

Part A.

Protein ID	Name in Figure S21	Location in genome assembly	Opsin subfamily	Major clade
Dappu-214454	BLOP	scaffold_53:628972-627385	Rhabdomeric	UV (Blue)
Dappu-303450	UVOP	scaffold_21:242254-243735	Rhabdomeric	UV
Dappu-14112	UNOP1	scaffold_95:369266-373273	Rhabdomeric	Unknown
Dappu-60874	UNOP2	scaffold_95:441206-436847	Rhabdomeric	Unknown
Dappu-307031	LOPA1	scaffold_598:27649-26145	Rhabdomeric	LongA
Dappu-307030	LOPA2	scaffold_598:19709-18148	Rhabdomeric	LongA
Dappu-67015	LOPA3	scaffold_598:16355-14836	Rhabdomeric	LongA
Dappu-306275	LOPA4	scaffold_47:938824-940341	Rhabdomeric	LongA
New	LOPA5N	scaffold_174:66413-66609	Rhabdomeric	LongA
Dappu-302464	LOPA6	scaffold_174:68557-70212	Rhabdomeric	LongA
Dappu-335676	LOPA7I	scaffold_696:761-2619	Rhabdomeric	LongA
Dappu-93838	LOPA8	scaffold_696:4556-6206	Rhabdomeric	LongA
Dappu-93844	LOPA9	scaffold_776:5823-4192	Rhabdomeric	LongA
Dappu-93844	LOPA10	scaffold_776:1944-678	Rhabdomeric	LongA
Dappu-54168	LOPB1	scaffold_40:709566-708143	Rhabdomeric	LongB
Dappu-305771	LOPB2	scaffold_40:716215-717823	Rhabdomeric	LongB
Dappu-198385	LOPB3	scaffold-40:722122-723709	Rhabdomeric	LongB
Dappu-305803	LOPB4	scaffold_40:728027-729621	Rhabdomeric	LongB
Dappu-106095	LOPB5	scaffold_40:732744-734341	Rhabdomeric	LongB
Dappu-305772	LOPB6	scaffold_40:737671-739173	Rhabdomeric	LongB
Dappu-321382	LOPB7	scaffold_40:742430-743903	Rhabdomeric	LongB
Dappu-43742	LOPB8	scaffold_6:1902006-1900546	Rhabdomeric	LongB
Dappu-216106	LOPB9	scaffold_78:111258-112698	Rhabdomeric	LongB
New	LOPB10	scaffold_78:114113-114451	Rhabdomeric	LongB
Dappu-326257	LOPB11	scaffold_78:119912-120349	Rhabdomeric	LongB
Dappu-254506	LOPB12	scaffold_78:123902-124674	Rhabdomeric	LongB
Dappu-326259	LOPB13	scaffold_78:126986-128343	Rhabdomeric	LongB
New	LOPB14	scaffold_78:133375-134342	Rhabdomeric	LongB
Dappu-326260	LOPB15	scaffold_78:142739-144179	Rhabdomeric	LongB
Dappu-24963	ARTHROPSIN1	scaffold_14:758164-761748	Rhabdomeric	Arthropsin
Dappu-47717	ARTHROPSIN2	scaffold_14:766741-771298	Rhabdomeric	Arthropsin
Dappu-24264	ARTHROPSIN3	scaffold_14:779460-783216	Rhabdomeric	Arthropsin
Dappu-23519	ARTHROPSIN4	scaffold_14:847788-844292	Rhabdomeric	Arthropsin
Dappu-2566	ARTHROPSIN5	scaffold_14:839526-835973	Rhabdomeric	Arthropsin
Dappu-47520	ARTHROPSIN6	scaffold_13:689696-688112	Rhabdomeric	Arthropsin
Dappu-223107	ARTHROPSIN7	scaffold_13:962643-964536	Rhabdomeric	Arthropsin
Dappu-47330	ARTHROPSIN8	scaffold_13:1021380-1023187	Rhabdomeric	Arthropsin
Dappu-312425	PTEROPSIN1	scaffold_6:1015520-1013655	Ciliary	Pteropsin
Dappu-312424/235776	PTEROPSIN2P	scaffold_6:1009166-1007372	Ciliary	Pteropsin

Dappu-307122	PTEROPSIN3	scaffold_6:1006658-1004665	Ciliary	Pteropsin
Dappu-97105	PTEROPSIN4	scaffold_6:767483-770451	Ciliary	Pteropsin
Dappu-51511	PTEROPSIN5P	scaffold_25:431410-435620	Ciliary	Pteropsin
Dappu-51298/103328	PTEROPSIN6	scaffold_25:446147-452002	Ciliary	Pteropsin
Dappu-51251	PTEROPSIN7	scaffold_25:460743-464047	Ciliary	Pteropsin
Dappu-243539	PTEROPSIN8	scaffold_25:484111-488573	Ciliary	Pteropsin
Dappu-303264	PTEROPSIN9	scaffold_2:3695086-3691119	Ciliary	Pteropsin

Part B.

Gene Name	Species	Accession
Bombyx UNOP	*Bombyx mori*	BGIBMGA012539-PA (silkdb.org)
Anolis pinopsin	*Anolis carolinensis*	AAD32622
Anopheles op1 4	*Anopheles gambiae*	XP_001238567
Anopheles op7	*Anopheles gambiae*	XP_001688790
Anopheles op10	*Anopheles gambiae*	XP_308329
Anopheles op8	*Anopheles gambiae*	XP_312478
Anopheles pteropsin 12	*Anopheles gambiae*	XP_312502.2
Anopheles pteropsin 11	*Anopheles gambiae*	XP_312503
Anopheles op9	*Anopheles gambiae*	XP_319247
Anopheles op6	*Anopheles gambiae*	XP_322000
Bombyx pteropsin	*Bombyx mori*	BGIBMGA008437-PA (silkdb.org)
Bombyx Lop1	*Bombyx mori*	BGIBMGA007787-PA (silkdb.org)
Apis Uvop	*Apis mellifera*	NP_001011605 XP_392791
Apis Blop	*Apis mellifera*	NP_001011606 XP_392042
Apis Lop1	*Apis mellifera*	NP_001011639 XP_397397
Apis pteropsin	*Apis mellifera*	NP_001035057
Apis Lop2	*Apis mellifera*	NP_001071293
Bombyx Lop2	*Bombyx mori*	NP_001036882
Branchinella BAG80984	*Branchinella kugenumaensis*	BAG80984
Branchinella kugenumaensis BAG80985	*Branchinella kugenumaensis*	BAG80985
Branchinella kugenumaensis BAG80986	*Branchinella kugenumaensis*	BAG80986
Branchinella kugenumaensis BAG80987	*Branchinella kugenumaensis*	BAG80987
Branchinella kugenumaensis BAG80988	*Branchinella kugenumaensis*	BAG80988
Branchinella kugenumaensis BAG80989	*Branchinella kugenumaensis*	BAG80989
Branchinella kugenumaensis BAG80990	*Branchinella kugenumaensis*	BAG80990
Branchinella kugenumaensis BAG80991	*Branchinella kugenumaensis*	BAG80991
Branchinella kugenumaensis BAG80992	*Branchinella kugenumaensis*	BAG80992
Branchinella kugenumaensis BAG80993	*Branchinella kugenumaensis*	BAG80993

Branchinella kugenumaensis BAG80994	*Branchinella kugenumaensis*	BAG80994
Branchinella kugenumaensis BAG80995	*Branchinella kugenumaensis*	BAG80995
Branchinella kugenumaensis BAG80996	*Branchinella kugenumaensis*	BAG80996
Branchinella kugenumaensis BAG80997	*Branchinella kugenumaensis*	BAG80997
Amphioxus1	*Branchiostoma belcheri*	BAC76019
Amphioxus2	*Branchiostoma belcheri*	BAC76020
Amphioxus4	*Branchiostoma belcheri*	BAC76021
Amphioxus5	*Branchiostoma belcheri*	BAC76022.1
Amphioxus6	*Branchiostoma belcheri*	BAC76024
Amphioxus melanopsin	*Branchiostoma belcheri*	Q4R1I4
Hasarius pteropsin	*Branchiostoma belcheri*	Listed in paper as AB525082 - but not found in Genbank
Amphioxus 3	*Branchiostoma belcheri*	C76023
Bufo pinopsin	*Bufo japonicus*	AAF12820
Ciona opsin1	*Ciona intestinalis*	NP_001027727
Anopheles op5	*Anopheles gambiae*	AGAP001162-RA (Anopheles genome on Ensembl)
Danio red	*Danio rerio*	AAD20549.1
Danio green1	*Danio rerio*	AAD24752
Danio peropsin	*Danio rerio*	NP_001004654
Danio Encephalopsin	*Danio rerio*	NP_001104634 XP_690306
Danio blue	*Danio rerio*	NP_571267
Danio UV	*Danio rerio*	NP_571394.1
Danio rod	*Danio rerio*	P35359.2
Drosophilia rh4	*Drosophila melanogaster*	NP_476701
Drosophilia rh5	*Drosophila melanogaster*	NP_477096
Drosophilia rh7	*Drosophila melanogaster*	NP_524035
Drosophilia rh6	*Drosophila melanogaster*	NP_524368
Drosophilia rh2	*Drosophila melanogaster*	NP_524398.1
Drosophilia rh1	*Drosophila melanogaster*	NP_524407.1
Drosophilia rh3	*Drosophila melanogaster*	NP_524411
Gallus melanopsin	*Gallus gallus*	NP_989956
Gallus pinopsin	*Gallus gallus*	NP_990740
Hemigrapsus rh1	*Hemigrapsus sanguineus*	Q25157.1
Hemigrapsus rh2	*Hemigrapsus sanguineus*	Q25158
Homo Encephalopsin	*Homo sapiens*	NP_055137
Homo melanopsin	*homo sapiens*	NP_150598
Homo RGR	*Homo sapiens*	NP_001012738.1
Homo peropsin	*Homo sapiens*	NP_006574
Homo neuropsin	*Homo sapiens*	NP_859528 XP_166440
Ictalurus parapinopsin	*Ictalurus punctatus*	O42266
Limulus ops5	*Limulus polyphemus*	ACO05013
Limulus lateral	*Limulus polyphemus*	P35360
Loligo GQ	*Loligo forbesi*	P24603
Bombyx Uvop	*Manduca sexta*	O02465

Bombyx Blop	*Manduca sexta*	O96107
Megoura rh1	*Megoura viciae*	AAG17119
Megoura UV	*Megoura viciae*	AAG17120
Mizuhopecten GQ	*Mizuhopecten yessoensis*	O15973
Mizuhopecten GO	*Mizuhopecten yessoensis*	O15974
Papillo Rh3	*Papilio glaucus*	AAD29445.1
Papillo Rh1	*Papilio glaucus*	AAD34220.1
Papillo Rh2	*Papilio glaucus*	AAD34221
Papillo Rh4	*Papilio glaucus*	AAD34224
Papilio Rh5	*Papilio glaucus*	AAD34222
Papilio rh6	*Papilio glaucus*	AAD34223
Pediculus UV	*Pediculus humanus corporis*	XP_002422743
PhLopFix	*Pediculus humanus corporis*	XP_002427337
Pediculus UNOPN	*Pediculus humanus corporis*	XP_002432663
Petromyzon pinopsin	*Petromyzon marinus*	O42490
Platynereis c	*Platynereis dumerilii*	AAV63834
Platynereis GQ	*Platynereis dumerilii*	CAC86665
Procambarus P35356	*Procambarus clarkii*	P35356
Salmo VA	*Salmo salar*	NP_001117098
Schistocerca 2	*Schistocerca gregaria*	Q26495
Schistocerca 1	*Schistocerca gregaria*	Q94741
Schistosoma GQ	*Schistosoma mansoni*	AAF73286
Takifugu TMT	*Takifugu rubripes*	NP_001027778
Tetradon RGR	*Tetraodon nigroviridis*	CAF98663.1
Fugu melanopsin	*Tetraodon nigroviridis*	CAF99228
Tertradon neuropsin	*Tetraodon nigroviridis*	CAG13006.1
Tigriopus californicus	*Tigriopus californicus*	HQ180268
Todarodes retinochrome	*Todarodes pacificus*	P23820
Tribolium pteropsin	*Tribolium castaneum*	EFA01685
Tribolium Lop	*Tribolium castaneum*	NP_001155991 XP_973147
Tribolium UV	*Tribolium castaneum*	XP_970344
Triops granarius BAG80976	*Triops granarius*	BAG80976
Triops granarius BAG80977	*Triops granarius*	BAG80977
Triops granarius BAG80978	*Triops granarius*	BAG80978
Triops granarius BAG80979	*Triops granarius*	BAG80979
Triops longicaudatus BAG80981	*Triops longicaudatus*	BAG80981
Triops longicaudatus BAG80982	*Triops longicaudatus*	BAG80982
Triops longicaudatus BAG80983	*Triops longicaudatus*	BAG80983
Triops longicaudatus BAG80998	*Triops longicaudatus*	BAG80998
Triops longicaudatus BAG80999	*Triops longicaudatus*	BAG80999
Vargula tsujii	*Vargula tsugii*	HQ180267
Xenopus melanopsin	*Xenopus laevis*	NP_001079143

F. Consequence *Daphnia*'s Genome Structure

Table S33. Summary of gene conversion features as a function of the number of genes within the genomes of *Daphnia pulex* and five selected *Drosophila* species. Conversion rate is given as converted pairs of paralogs/total pairs of paralogs analyzed.

	Daphnia pulex	*Drosophila melanogaster*	*Drosophila yakuba*	*Drosophila pseudoobscura*	*Drosophila virilis*	*Drosophila grimshawi*
No. Conversion events	7,007	190	223	313	246	377
No. Converted pairs	6,086	138	194	244	186	301
No. Converted genes	6,213	233	337	407	305	483
Events/Pair	1.15	1.38	1.15	1.28	1.32	1.25
Total pairs analyzed	55,362	1,790	2,239	2,128	1,576	2,269
Total genes analyzed	13,330	1,905	2,747	2,501	1,960	2,683
% Converted genes	46.61	12.23	12.27	16.27	15.56	18
Gene conversion rate	10.99	7.71	8.66	11.47	11.8	13.27

Table S34. Summary of genome-wide gene conversion features among *Daphnia* and five selected *Drosophila* species.

Converted	*Daphnia pulex*	*Drosophila melanogaster*	*Drosophila yakuba*	*Drosophila pseudoobscura*	*Drosophila virilis*	*Drosophila grimshawi*
Same strand	1,105	89	102	119	113	153
Opposite strand	392	33	53	43	43	49
Non-converted						
Same strand	3,881	908	1,023	829	733	852
Opposite strand	2,133	285	351	304	289	350
Fisher's 2-tail	5.51E-12	0.4382	0.0267	1	0.924	0.1773

Table S35. Summary of genome-wide gene conversion features as a function of the location of paralogs on scaffolds or Müller elements among *Daphnia* and five selected *Drosophila* species.

	Daphnia pulex	*Drosophila melanogaster*	*Drosophila yakuba*	*Drosophila pseudoobscura*	*Drosophila virilis*	*Drosophila grimshawi*
Converted intraelement/scaffold	1,497	122	155	162	156	202
Converted interelement/scaffold	4,589	16	22	40	18	18
Total converted	6,086	138	177	202	174	220
Non-converted intraelement/scaffold	6,014	1,193	1,374	1,133	1,022	1,202
Non-converted interelement/scaffold	43,262	459	480	517	357	344
Total non-converted	49,276	1,652	1,854	1,650	1,379	1,546
Total intraelement/scaffold	7,511	1,315	1,529	1,295	1,178	1,404
Total interelement/scaffold	47,851	475	502	557	375	362
% Converted intraelement/scaffold	24.6	88.41	87.57	80.2	89.66	91.82
% Non-converted intraelement/scaffold	12.2	72.22	74.11	68.67	74.11	77.75

Table S36. Summary of genome-wide gene conversion (conv.) features as a function of the size of conversion tracts among *Daphnia* and five selected *Drosophila* species. Minimum and maximum values represent the shortest and longest converted tract found by Geneconv [S99].

	Daphnia pulex	*Drosophila melanogaster*	*Drosophila yakuba*	*Drosophila pseudoobscura*	*Drosophila virilis*	*Drosophila grimshawi*
Average (bp)	169	186	192	180	182	297
Median (bp)	109	83	81	95	81	167
Minimum (bp)	20	14	11	7	11	10
Maximum (bp)	2413	2213	2837	1287	3079	2437
Total converted bp	1,180,733	35,322	42,711	56,443	44,888	111,907
Total bp converted pairs	7,004,873	385,800	518,349	644,846	518,080	861,115
Total bp screened	14,140,570	3,454,328	4,005,419	3,764,802	3,251,664	3,957,669
% Tract/conv. pairs	16.86	9.16	8.24	8.75	8.66	13
% Tract/all pairs	8.35	0.92	0.94	1.28	1.19	2.32

Table S37. Summary of genome-wide gene conversion (conv.) features as a function of the size of gene families among *Daphnia* and five selected *Drosophila* species. The asterisk indicates that the average of *Daphnia* converted families has been calculated after removing the largest family with 4,007 genes (the average would otherwise be ~11 genes per family).

	Daphnia pulex	*Drosophila melanogaster*	*Drosophila yakuba*	*Drosophila pseudoobscura*	*Drosophila virilis*	*Drosophila grimshawi*
No. Conv. gene families	942	99	144	169	131	211
Average family size conv.	7.63*	3.28	2.99	2.97	2.97	2.9
% Conv. of family size 2	26.22%	57.80%	61.10%	55.60%	55.70%	68.20%
% Nonconv. of family size 2	60.43%	80.00%	85.70%	82.90%	84.10%	83.50%

Table S38. Summary of genome-wide gene conversion features as a function of the distance of intra-element or intra-scaffold paralogs among *Daphnia* and five selected *Drosophila* species.

Converted	*Daphnia pulex*	*Drosophila melanogaster*	*Drosophila yakuba*	*Drosophila pseudoobscura*	*Drosophila virilis*	*Drosophila grimshawi*
Average distance (bp)	110,881	294,326	1,027,163	134,591	281,609	153,547
Median distance (bp)	16,060	1,797	2,601	2,325	2,360	2,530
Non-converted						
Average distance (bp)	268,781	1,508,835	1,415,872	1,434,743	1,183,649	620,793
Median distance (bp)	63,275	4,915	4,847	7,128	7,493	5,478

Table S39. Homologous di-domain hemoglobin genes (Hb) of *Daphnia pulex* and *Daphnia magna*. *Daphnia magna* hemoglobin gene cluster contig assembly NCBI accession number is AB518060.

Daphnia pulex gene	Location in genome assembly	*Daphnia magna* gene	Location in contig assembly	% identity
Dpul-Hb1 (Dappu-96311)	scaffold_4:23666681-2368249	Dmag-Hb1	553..2095	73.1
Dpul-Hb2 (Dappu-230332)	scaffold_4:2370110-2374287	Dmag-Hb2	4360..5875	73.2
Dpul-Hb3 (Dappu-311662)	scaffold_4:2372773-2374287	Dmag-Hb3	7071..8561	70.6
Dpul-Hb4 (Dappu-234836)	scaffold_4:2376081-2377561	Dmag-Hb4	10541..12024	71.7
Dpul-Hb5 (Dappu-234837)	scaffold_4:2380765-2382213	Dmag-Hb5	15384..16893	71.2
Dpul-Hb6 (Dappu-234838)	scaffold_4:2383418-2384965			
Dpul-Hb7 (Dappu-234839)	scaffold_4:2386115-2387624	Dmag-Hb7	19734..21224	70.7
Dpul-Hb8 (Dappu-230333)	scaffold_4:2388769-2390272	Dmag-Hb8	22483..24059	71.8
Dpul-Hb9 (Dappu-210408)	scaffold_17:410538-409010			
Dpul-Hb10 (Dappu-92880)	scaffold_36:522846-524214			
Dpul-Hb11 (Dappu-93831)	scaffold_452:2493-3859			

G. Evolutionary Diversification of Duplicated Genes

Table S40. The number of paralog pairs that differ unambiguously in their expression patterns among 0 to 12 conditions as a function of genetic divergence measured as nucleotide substitutions at synonymous sites (K_s).

	Number of Conditions												
K_s	0	1	2	3	4	5	6	7	8	9	10	11	12
0 - 0.05	14	7	4	3	1	0	0	1	0	0	0	0	0
0.05 - 0.1	35	31	8	9	2	1	0	0	0	0	0	0	0
0.1 - 0.5	729	468	215	118	54	40	23	10	3	4	1	2	2
0.5 - 1	940	604	414	227	163	95	61	33	25	12	8	3	22
1 - 2	1106	792	596	563	443	364	224	161	125	63	25	16	11
2 - 3	520	458	394	373	325	239	208	172	106	51	27	7	7
3 - 5	264	260	274	225	246	188	174	145	93	53	24	16	10

Table S41. Chi-square tests for associations between paralogs (K_s < 2) sharing expression patterns across 12 conditions tested on microarrays and **A.** their genomic arrangements (dispersed or clustered); **B.** whether gene conversion signatures are detected.

A.	Expression Patterns			Clustered X^2 = 0.027; p = 0.869
Genomic Arrangement	Same	Different	% Different	
Dispersed	2396	5125	68.1	
Clustered	428	932	68.5	
B.	Expression Patterns			GC X^2 =11.9; p = 0.00055
Gene Conversion (GC)	Same	Different	% Different	
No Signature	2426	5414	69.1	
Signature of GC	398	643	61.8	

Table S42. The number of paralog pairs that have the same expression patterns and that have different expression patterns among 0 to 12 conditions as a function of genetic divergence measured as nucleotide substitutions at synonymous sites (K_s), comparing sets that include and that exclude genes showing signatures of gene conversion.

	All paralog pairs				Paralog pairs excluding gene conversions			
K_s	Same	Different	% Different	P-value	Same	Different	% Different	P-value
0 - 0.05	14	16	53.3	1.0000	11	11	50.0	1.0000
0.05 - 0.1	35	51	59.3	0.2840	23	38	62.3	0.2020
0.1 - 0.5	729	940	56.3	0.0003	540	729	57.4	0.0002
0.5 - 1	940	1667	63.9	0.0000	807	1445	64.2	0.0000
1 - 2	1106	3383	75.4	0.0000	1045	3191	75.3	0.0000
2 - 3	520	2367	82.0	0.0000	504	2315	82.1	0.0000
3 - 5	264	1708	86.6	0.0000	261	1689	86.6	0.0000

H. Functional Significance of Expanded Gene Families

Table S43. Metabolic pathways (classified by KEGG and highlighted in Figure 4) containing expanded metabolic genes in the *Daphnia pulex* genome compared to insects and vertebrates. The number of gene copies is indicated for identified enzymes. “Highlighted pathway ID” refers to panels A-G in Figure 4 where pathway "H" corresponds to the enzymes not listed in any panels.

Highlighted pathway ID	KEGG map ID	KEGG name	Enzyme commission No.	Enzyme name	No. gene copies
H	map00040	Pentose and glucuronate interconversions	2.4.1.17	glucuronosyltransferase	24
-	map00040	Pentose and glucuronate interconversions	5.3.1.5	xylose isomerase	6
-	map00072	Synthesis and degradation of ketone bodies	1.1.1.30	3-hydroxybutyrate dehydrogenase	8
-	map00100	Biosynthesis of steroids	1.14.13.72	methylsterol monooxygenase	11
-	map00120	Bile acid biosynthesis	3.1.1.13	sterol esterase	28
E	map00150	Androgen and estrogen metabolism	2.4.1.17	glucuronosyltransferase	24
E	map00150	Androgen and estrogen metabolism	2.8.2.4	estrone sulfotransferase	7
H	map00230	Purine metabolism	2.7.7.6	DNA-directed RNA polymerase	105
H	map00240	Pyrimidine metabolism	2.7.7.6	DNA-directed RNA polymerase	105
G	map00330	Arginine and proline metabolism	1.14.11.2	procollagen-proline dioxygenase	12
G	map00330	Arginine and proline metabolism	1.5.1.12	1-pyrroline-5-carboxylate dehydrogenase	6
H	map00480	Glutathione metabolism	3.4.11.2	membrane alanyl aminopeptidase	26
-	map00500	Starch and sucrose metabolism	2.4.1.15	alpha,alpha-trehalose-phosphate synthase (UDP-forming)	4
-	map00510	N-Glycan biosynthesis	2.4.1.38	beta-N-acetylglucosaminylglycopeptide beta-1,4-galactosyltransferase	11
-	map00512	O-Glycan biosynthesis	2.4.1.122	glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase	16
A	map00530	Aminosugars metabolism	3.2.1.14	chitinase	15
A	map00530	Aminosugars metabolism	3.2.1.52	beta-N-acetylhexosaminidase	10
H	map00531	Glycosaminoglycan degradation	3.2.1.52	beta-N-acetylhexosaminidase	10
-	map00561	Glycerolipid metabolism	3.1.1.3	triacylglycerol lipase	37
-	map00590	Arachidonic acid metabolism	5.3.99.2	prostaglandin-D synthase	11
-	map00600	Sphingolipid metabolism	2.4.1.47	N-acylsphingosine galactosyltransferase	18
F	map00601	Glycosphingolipid biosynthesis - lactoseries	2.4.1.206	lactosylceramide 1,3-N-acetyl-beta-D-glucosaminyltransferase	9
F	map00601	Glycosphingolipid biosynthesis - lactoseries	2.4.1.65	3-galactosyl-N-acetylglucosaminide 4-alpha-L-fucosyltransferase	7

F	map00602	Glycosphingolipid biosynthesis - neo-lactoseries	2.4.1.152	4-galactosyl-N-acetylglucosaminide 3-alpha-L-fucosyltransferase	81
F	map00602	Glycosphingolipid biosynthesis - neo-lactoseries	2.4.1.206	lactosylceramide 1,3-N-acetyl-beta-D-glucosaminyltransferase	8
F	map00602	Glycosphingolipid biosynthesis - neo-lactoseries	2.4.1.65	3-galactosyl-N-acetylglucosaminide 4-alpha-L-fucosyltransferase	9
F	map00603	Glycosphingolipid biosynthesis - globoseries	2.4.1.228	lactosylceramide 4-alpha-galactosyltransferase	29
F	map00603	Glycosphingolipid biosynthesis - globoseries	3.2.1.52	beta-N-acetylhexosaminidase	10
H	map00604	Glycosphingolipid biosynthesis - ganglioseries	3.2.1.52	beta-N-acetylhexosaminidase	10
-	map00630	Glyoxylate and dicarboxylate metabolism	6.3.4.3	formate--tetrahydrofolate ligase	7
-	map00650	Butanoate metabolism	1.1.1.30	3-hydroxybutyrate dehydrogenase	8
-	map00670	One carbon pool by folate	6.3.4.3	formate--tetrahydrofolate ligase	7
-	map00680	Methane metabolism	1.1.1.284	S-(hydroxymethyl)glutathione dehydrogenase	9
-	map00680	Methane metabolism	1.11.1.7	peroxidase	38
-	map00720	Reductive carboxylate cycle (CO2 fixation)	2.7.9.2	pyruvate, water dikinase	2
D	map00920	Sulfur metabolism	2.8.2.4	estrone sulfotransferase	7
-	map00940	Phenylpropanoid biosynthesis	1.11.1.7	peroxidase	38
-	map00940	Phenylpropanoid biosynthesis	6.2.1.12	4-coumarate--CoA ligase	12

Table S44. Metabolic pathways (classified by KEGG and highlighted in Figure 4) containing expanded metabolic genes in the arthropod genomes compared to vertebrate genomes. The number of gene copies is indicated for identified enzymes. “Highlighted pathway ID” refers to panels A-G in Figure 4 where pathway "H" corresponds to the enzymes not listed in any panels.

Highlighted pathway ID	KEGG map ID	KEGG name	Enzyme commission No.	Enzyme name	No. gene copies
-	map00040	Pentose and glucuronate interconversions	2.4.1.17	glucuronosyltransferase	24
-	map00100	Biosynthesis of steroids	1.14.13.72	methylsterol monooxygenase	11
-	map00120	Bile acid biosynthesis	3.1.1.13	sterol esterase	28
-	map00150	Androgen and estrogen metabolism	2.4.1.17	glucuronosyltransferase	24
-	map00150	Androgen and estrogen metabolism	2.8.2.4	estrone sulfotransferase	7
-	map00230	Purine metabolism	2.7.7.6	DNA-directed RNA polymerase	105
H	map00230	Purine metabolism	4.6.1.2	guanylate cyclase	16
-	map00240	Pyrimidine metabolism	2.7.7.6	DNA-directed RNA polymerase	105
-	map00251	Glutamate metabolism	1.4.1.13	glutamate synthase (NADPH)	1
-	map00340	Histidine metabolism	4.1.1.22	histidine decarboxylase	7
H	map00340	Histidine metabolism	4.1.1.28	aromatic-L-amino-acid decarboxylase	7
B	map00350	Tyrosine metabolism	1.14.17.1	dopamine beta-monooxygenase	5
B	map00350	Tyrosine metabolism	4.1.1.25	tyrosine decarboxylase	7
B	map00350	Tyrosine metabolism	4.1.1.28	aromatic-L-amino-acid decarboxylase	7
-	map00361	gamma-Hexachlorocyclohexane degradation	3.1.3.1	alkaline phosphatase	6
H	map00380	Tryptophan metabolism	4.1.1.28	aromatic-L-amino-acid decarboxylase	7
-	map00480	Glutathione metabolism	3.4.11.2	membrane alanyl aminopeptidase	26
-	map00500	Starch and sucrose metabolism	2.4.1.15	alpha,alpha-trehalose-phosphate synthase (UDP-forming)	4
-	map00500	Starch and sucrose metabolism	3.2.1.20	alpha-glucosidase	11
-	map00512	O-Glycan biosynthesis	2.4.1.122	glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase	16
A	map00530	Aminosugars metabolism	2.4.1.16	chitin synthase	3
-	map00530	Aminosugars metabolism	3.2.1.14	chitinase	15
-	map00530	Aminosugars metabolism	3.2.1.52	beta-N-acetylhexosaminidase	10
-	map00531	Glycosaminoglycan degradation	3.2.1.52	beta-N-acetylhexosaminidase	10
-	map00561	Glycerolipid metabolism	3.1.1.3	triacylglycerol lipase	37
-	map00562	Inositol phosphate metabolism	3.1.3.62	multiple inositol-polyphosphate phosphatase	6
-	map00564	Glycerophospholipid metabolism	1.1.99.5	glycerol-3-phosphate dehydrogenase	7
C	map00590	Arachidonic acid metabolism	5.3.99.2	prostaglandin-D synthase	11

C	map00590	Arachidonic acid metabolism	5.3.99.5	thromboxane-A synthase	2
-	map00600	Sphingolipid metabolism	2.4.1.47	N-acylsphingosine galactosyltransferase	18
-	map00601	Glycosphingolipid biosynthesis - lactoseries	2.4.1.65	3-galactosyl-N-acetylglucosaminide 4-alpha-L-fucosyltransferase	9
-	map00602	Glycosphingolipid biosynthesis - neo-lactoseries	2.4.1.152	4-galactosyl-N-acetylglucosaminide 3-alpha-L-fucosyltransferase	81
-	map00602	Glycosphingolipid biosynthesis - neo-lactoseries	2.4.1.65	3-galactosyl-N-acetylglucosaminide 4-alpha-L-fucosyltransferase	7
-	map00603	Glycosphingolipid biosynthesis - globoseries	2.4.1.228	lactosylceramide 4-alpha-galactosyltransferase	29
-	map00603	Glycosphingolipid biosynthesis - globoseries	3.2.1.52	beta-N-acetylhexosaminidase	10
-	map00604	Glycosphingolipid biosynthesis - ganglioseries	3.2.1.52	beta-N-acetylhexosaminidase	10
-	map00790	Folate biosynthesis	2.7.6.3	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase	3
-	map00790	Folate biosynthesis	3.1.3.1	alkaline phosphatase	6
-	map00920	Sulfur metabolism	2.8.2.4	estrone sulfotransferase	7
-	map00940	Phenylpropanoid biosynthesis	6.2.1.12	4-coumarate--CoA ligase	12

Table S45. Ninety-six (96) *Daphnia pulex* genes from three lineage-specific gene family expansions that are part of the glycosphingolipid biosynthesis neo-lactoseries metabolic pathway.

Enzyme 2.4.1.152 (Alpha-1,3-fucosyltransferase C, Glycosyl transferase, family 10)	Location in genome assembly
Dappu-104196	scaffold_29:516567-518088
Dappu-106945	scaffold_46:224049-225221
Dappu-107642	scaffold_51:725267-726313
Dappu-111600	scaffold_87:116388-117590
Dappu-116054	scaffold_173:170541-171871
Dappu-13230	scaffold_356:21642-22727
Dappu-13713	scaffold_68:176859-177731
Dappu-15329	scaffold_356:7770-8810
Dappu-19438	scaffold_10031:420-1163
Dappu-198878	scaffold_46:232818-234380
Dappu-219820	scaffold_1396:115-1342
Dappu-221393	scaffold_4:3007607-3011002
Dappu-227431	scaffold_76:225531-227364
Dappu-23160	scaffold_14:1100902-1106527
Dappu-236411	scaffold_7:1456987-1458048
Dappu-241186	scaffold_18:353961-355160
Dappu-244685	scaffold_29:499533-504745
Dappu-24623	scaffold_7:2167295-2168158
Dappu-248921	scaffold_46:242704-243900
Dappu-251980	scaffold_61:720217-721224
Dappu-25363	scaffold_29:601986-605294
Dappu-253741	scaffold_72:252130-254074
Dappu-25935	scaffold_34:56787-57719
Dappu-260055	scaffold_132:222649-224187
Dappu-260935	scaffold_145:50678-51847
Dappu-266638	scaffold_332:14980-18089
Dappu-266923	scaffold_356:19512-21162
Dappu-266928	scaffold_356:31731-32948
Dappu-272135	scaffold_2299:6593-7516
Dappu-302400	scaffold_17:1454152-1456089
Dappu-302457	scaffold_173:167951-169157
Dappu-302634	scaffold_18:1062894-1063732
Dappu-302891	scaffold_19:1404773-1406891
Dappu-308012	scaffold_72:235616-236979
Dappu-311402	scaffold_4:1049624-1050748
Dappu-312894	scaffold_7:1448864-1450036
Dappu-313010	scaffold_7:2093696-2094982
Dappu-313025	scaffold_7:2168985-2170223
Dappu-315506	scaffold_14:1098141-1099795
Dappu-315514	scaffold_14:1115817-1116902
Dappu-316372	scaffold_17:640256-641443
Dappu-316572	scaffold_17:1457020-1458114
Dappu-316587	scaffold_18:11511-12227
Dappu-316980	scaffold_18:1352486-1353631
Dappu-318584	scaffold_25:756640-757413
Dappu-319378	scaffold_29:538303-540175
Dappu-325563	scaffold_71:175007-191884
Dappu-325685	scaffold_72:213716-214915

Dappu-328684	scaffold_107:160375-162469
Dappu-331779	scaffold_173:181964-182991
Dappu-331784	scaffold_173:195174-196654
Dappu-334524	scaffold_356:27899-29071
Dappu-336888	scaffold_1221:450-1589
Dappu-3750	scaffold_356:11019-11912
Dappu-3751	scaffold_66:364793-365776
Dappu-3818	scaffold_183:114080-114946
Dappu-4083	scaffold_52:670620-671612
Dappu-4136	scaffold_13:1294763-1295593
Dappu-4141	scaffold_18:1250655-1251494
Dappu-41601	scaffold_3:2506282-2507324
Dappu-48653	scaffold_17:1277138-1280333
Dappu-49176	scaffold_18:13927-17142
Dappu-49339	scaffold_18:1253253-1254113
Dappu-52155	scaffold_29:856285-857325
Dappu-53630	scaffold_36:644234-645364
Dappu-55591	scaffold_50:689973-690977
Dappu-56240	scaffold_54:717402-718004
Dappu-58299	scaffold_72:221080-222066
Dappu-58316	scaffold_72:239493-240638
Dappu-58354	scaffold_72:216595-217752
Dappu-60056	scaffold_87:9135-10897
Dappu-60476	scaffold_90:365213-366334
Dappu-63087	scaffold_132:309467-310211
Dappu-64359	scaffold_173:192910-193980
Dappu-65379	scaffold_216:159136-160098
Dappu-66309	scaffold_332:35690-36814
Dappu-66315	scaffold_332:27798-28952
Dappu-67044	scaffold_604:27130-28257
Dappu-67045	scaffold_604:41549-42619
Dappu-67046	scaffold_604:35948-36971
Dappu-68594	scaffold_1936:6389-7141

Enzyme 2.4.1.206

(Beta-1,3-galactosyltransferase 5, Glycosyl transferase, family 31)

Dappu-111641	scaffold_87:215283-216322
Dappu-14718	scaffold_59:689162-689806
Dappu-241308	scaffold_18:724825-725889
Dappu-241507	scaffold_18:1240340-1242342
Dappu-314238	scaffold_10:1875122-1876652
Dappu-316941	scaffold_18:1246144-1247416
Dappu-325474	scaffold_70:418642-420376
Dappu-56803	scaffold_59:685045-685659

Enzyme 2.4.1.65

(alpha 1,3-fucosyltransferase, Glycosyl transferase, family 10)

Dappu-202947	scaffold_118:212190-214813
Dappu-26234	scaffold_18:1256198-1257003
Dappu-58283	scaffold_72:244180-245176
Dappu-58437	scaffold_72:241005-242003
Dappu-61832	scaffold_107:156253-157660
Dappu-64347	scaffold_173:173776-180217
Dappu-64409	scaffold_173:184551-190756

Table S46. Alignment of Enzyme 2.4.1.65 *Daphnia* proteins, with *Tribolium castaneum* and *Ixodes scapularis* orthologs, using MUSCLE [S58].

Protein ID	Description

```
Dappu-61832        ------------------------------------------------------------
Dappu-64347        ------------------------------------------------------------
Dappu-26234        ------------------------------------------------------------
Dappu-58283        ------------------------------------------------------------
Dappu-64409        ------------------------------------------------------------
Dappu-58437        ------------------------------------------------------------
Dappu-202947       ---------------------------------MLALHVQQQNPLWLRVLEQSSAQGGGY
Ixodes_ISCW003580  MPGHFDVRRMRDQGACNSRSPRVQTWRSMWPGKKLRLRRLLATAIGLSCCTLLLISFKEV
Tribolium_TC014343 MP------------------------------PRLSARRLCLVIFFFGGVTVLVTLHHRL


Dappu-61832        ------------------------------------------------------------
Dappu-64347        ------------------------------------------------------------
Dappu-26234        ------------------------------------------------------------
Dappu-58283        ------------------------------------------------------------
Dappu-64409        ------------------------------------------------------------
Dappu-58437        ------------------------------------------------------------
Dappu-202947       YQQNSKELDSDDSANFVVINNDESKIPSAD--NNVRINNTTISSNILRRHGL---PWYIK
Ixodes_ISCW003580  QQDADLQDQPNVLEPKLLLQQQSYAVPQDSHKGTVRRDGRQMEKRVESGSGSSERPWYMK
Tribolium_TC014343 TWPSTKSRIPSSDEDELLIHTTAPSLPVVE------------EHETEQPPKSQEKAWFFG


Dappu-61832        ------------------------------------------------------------
Dappu-64347        ------------------------------------------------------------
Dappu-26234        ------------------------------------------------------------
Dappu-58283        ------------------------------------------------------------
Dappu-64409        ------------------------------------------------------------
Dappu-58437        ------------------------------------------------------------
Dappu-202947       NDGYRPSQGDLVDNIWPVGRLKGDRIEEQLMIPSRDIDQHATPGDGTSFPVKPKLKKIFL
Ixodes_ISCW003580  GGLRRPLPGD-TGSLWP-HEDAGDRIEAQLMFVPEDYKRNSSR-----------LKKILL
Tribolium_TC014343 GGTLFPTASKGLPRLFP-DQTDGDRIIEQLMYVPEDYQGFDTP-----------EKVILA


Dappu-61832        ---------------------CPVTTCLATDDPYLLDSVDQYDAVMFH------------
Dappu-64347        ------------GRDAFRKWGCPVWQCETSTNR---TDVHDYDAVIFHMR-SWNRNDLPK
Dappu-26234        --------------------------------------VQNYDAVIFHLR-SWKSNDLPQ
Dappu-58283        ---------------------CPVWQCETSTNR---TDVHDYDAVIFHMRGSWNPNDLPH
Dappu-64409        --------------------------CETSTNR---TDTHDYDAVIFHMR-SWNPNDLPK
Dappu-58437        ---------------------CPVWQCETSTNR---TDVHDYDAVIFHMRGSWNPNDLPH
Dappu-202947       PNGLGSWQT-KSGQKVFTEQKCPVDRCSLTSNR---DEAANADAIMFKDF----FSTPSH
Ixodes_ISCW003580  MHGMGGWGELPRGRTVFLRDKCPVDTCEIVTSQ---DDAAEADAILFKDR----FTPPRH
Tribolium_TC014343 YNGLGTWGQ-RSGPGSF--HGCPVSRCSLTDDR---SRAADADAILYKDH----FIHPPV
                                                          . : **::::

Dappu-61832        ------------------WP--------SLCDY---------------PYV---------
Dappu-64347        RRT-PQQRYVFWLLESAGWPEYLPMHTSSLGNFFNWTLTYRWDSDMVMPYG-Y-------
Dappu-26234        NRSLHQQRYIFWLLESAGWPEYL--DTKPLGNFFNWTLTYRWDSDMVMPYG-YVRPTGNV
Dappu-58283        RRS-PQQRYVFWILESAEWREYL--NTSTLGNFFNWTLTYRWDSDMIMPYG-YVRPTGNV
Dappu-64409        RRT-PQQRYLFWLLESAGWPEYL--NTSQLGNFFNWTLTYRWDSDMVMPYG-Y-------
Dappu-58437        RRS-PQQRYVFWILESAGWPEYL--NTSTLGNFFNWTLTYRWDSDMVMPYG-YVRPTGNV
```

```
Dappu-202947       PRP-PHQIWIMYMLECPLHTQYI-----REKDVFNWTATYKSDSELVTPYEKWVYFDDKV
Ixodes_ISCW003580  RRP-WQQVWILYLLECPYHTQTF----AHFRDTFNWTATYRHDSDIVAPYEKFVRYDDLD
Tribolium_TC014343 SRP-FNQVWIMYFLECPYHTQSI-----KFPDVINWTATYRRDSDLVAPYERWTYFDPQV
                                  .           :                **

Dappu-61832        -------------------ASRSEHQLFVMVSDESPQHSR-----------IDIFGKC
Dappu-64347        ------DQLKQLMSVQKMNYAAGKTKMASWMVSNCGAHSNRLQMVKILQKYIQVDVYGVC
Dappu-26234        PLHPSDDQMKELLSNQKVNYATAKTKMAAWMVSNCGSHSSRNEMVNIIKKYIQVDVYGAC
Dappu-58283        PLHPSENQLKQLMSNQKVNYAAGKAKMASWMVSSCFSHSSRHEMVKILQKYIQVDIYGAC
Dappu-64409        ----------QLMSVQKMNYAAGKTKMAAWMVSNCGSHSNRKEMVSLLQKYIQVDVYGAC
Dappu-58437        PLHPSENQLKQLMSDQKVNYAAGKTKMAAWFVSNCVAKSNRNEMVKILQKYIQIDVYGVC
Dappu-202947       RRKP-----------VTTNFAANKTKKVAWFVSNCGAKNNRLEYAHALQKHIDVDIYGSC
Ixodes_ISCW003580  PVAEA----------SRVLPNHNKTKKVAWFVSNCAARNQRLQYARKLGAHIEVDIFGAC
Tribolium_TC014343 RQKV-----------QNRDYSANKTKKVAWFVSNCGARNGRLAYARELSKYIQVDIYGMC
                                     . :    :**.  ..  *           :*::* *

Dappu-61832        GKPFCSFDQL-----NDCYQRIEIDYKFYLSFENSLCRDYITEKFF-NLLDRNIVPIVYG
Dappu-64347        GNLTCPKENS-----DRCNNLLD-EYKFYLSAENSLCADYVSEKFY-RALKTDIIPVVYG
Dappu-26234        GTMSCPKEAGVDNSSEDCRDMVGKTYKFYMSLENSLCRDYISEKLF-GMLHRPIIPIVCG
Dappu-58283        GTKTCPKKEDENNSSEECRDVAGGNYKFYMALENSLCHEYISEKFF-GMLHRPIIPVVFG
Dappu-64409        GPLKCPKEVGVDNSSEDCRDMAGQNYKFYMALENSLCRDYISEKFF-GMLQRPVIPVVFG
Dappu-58437        GNLTCPKEVGVDNSSEDCRDMAGENYKFYMALENSLCHEYISEKFF-GMLHRPVIPVVFG
Dappu-202947       GTKNCPRHSG-----DHCLDILSTEYKFYLAFENSNCRDYITEKFYVNGLGSKVLPIVMG
Ixodes_ISCW003580  GPLKCPRARA-----GHCFDILDREYKFYLAFENSNCKDYITEKFFVNGLGRDVVPIAMG
Tribolium_TC014343 GPLACPRSD------KKCFDLLDREYKFYLAFENSNCRDYITEKFYVNGLGQNVLPIVMG
                   *   *.          * :     ****:: *** * :*::**::   *   ::*:. *

Dappu-61832        A--GNYEAIAPPHSYIDALKY-TPVQLAKYLDILDKNDTLYNEYFWWKPFYKLMA-----
Dappu-64347        G--ADYAAYAPPHSYIHVADFASPKQLAEYLLLLDKNEALYLKYFEWKKDYDVLRGPLD-
Dappu-26234        L-HDYYDKIAPPHSFINAAKFENMQKLADYLILLDKNDTLYNEYFWWKPH----------
Dappu-58283        L-HDHYDKIAPPHSYINAAKFENMRQLADYLILLDRNDTLYNEYFWWKPHFESRYKQKDV
Dappu-64409        L-HNHYDQMAPAHSFINAAKFENMRQLADYLILLDRNDTLYNEYFWWKPHFESRYKQKDV
Dappu-58437        L-HDHYDKIAPPHSFINAAKFENMRQLADYLILLDRNDTLYNEYFWWKPHFESRYKQKDV
Dappu-202947       APRADYEKHAPEHSFIHVDDFATPKELADYLHLLNSNDTLYNEYFEWKETGQFIN-----
Ixodes_ISCW003580  GRPEDYRRASPDHSFVHVEDFPSEKALADYLHVLDRNDSLYNEYFRWKGSGEFIN-----
Tribolium_TC014343 ARPEDYQRSAPEGSYIHVDEFAGPAELAAYLNRLDKDSTLYNSYFKWKGTGQFIN-----
                        *   :*  *:: . .:     ** **  *: :.:** .** **

Dappu-61832        NIAFCQLCQQLNQ-PRTHVQWYHDIDAWYDGGNHCHKPNRFKVPYTFSYFIIIGRM----
Dappu-64347        --GWCDLCAKLND-PQEPAKVYQSMAEWWYDEVPCYPGESFIKTVLNHIQ----------
Dappu-26234        ------------------------------------------------------------
Dappu-58283        NIGMCHLCASLHN-KDLPPKVYPNMTDWWESKSSCISTPLIS------------------
Dappu-64409        NIGMCHLCASLHN-KDMPPKVYANMTQWWDEQSFCINSPPIS------------------
Dappu-58437        NIGMCHLCASLHN-KEMPAKVYPNMTHWWDEQSSCINSPPIL------------------
Dappu-202947       TYFFCRLCSMLHEAPYSPPRYYDDFNEWWRGGTNCIKGSWRDLENHQRNKKNKKKVGGDG
Ixodes_ISCW003580  TYFWCRLCAMLHA-PPVP-KVYPDIGAWWSGPGTCNSNRWSKFKTKKDSVGYVFT-----
Tribolium_TC014343 TFFWCRLCAMLHA-PRVH-RHYDDINDWWRGPGVCSSKSWRNADFV--------------
```

Table S47. Alignment of Enzyme 2.4.1.206 *Daphnia* proteins, with *Tribolium castaneum* and *Ixodes scapularis* orthologs, using MUSCLE [S58].

Protein ID	Description

```
Ixodes_ISCW018107  -----------------------------------MRIVFLRRIPGDVTMHYQIDAWL
Tribolium_TC014213 --------------------------MLLPCMGVPWLAQVRLLHMAILVLATASCLL
Tribolium_TC008953 -------------------------------------------------MPNTEVIVEGWD
Ixodes_ISCW003730  ---------------------------------------MATALVRLLPRITVIALCLL
Dappu-325474       SLSLMYNNGSSKDPSTSCQNQTALATVLQTKSNCESNLTNCLGQKDNVTIELKNMANDWE
Dappu-241507       ----------------------------------MTNKGAGRLDNDSTADVTININRK
Dappu-241308       -------------------------MRLPIFLALLAIAKTQLSQNHFYKKLNVINEERR
Dappu-111641       -------------------------MRNHSHPEHITNISLNNKASIQAETNLTIIDRE
Dappu-316941       ------------------------------------------------MATHLRNT-----
Dappu-314238       VFFESHERLNPTSEFRRAQNDSHRFTQALNSCAHSNDVQPPIKEQQHWSSDENLTVINRD
Dappu-56803        ------------------------------------------------------------
Dappu-14718        ------------------------------------------------AEITNNNLA

Ixodes_ISCW018107  -----------LLYVFGVP----------------KSKHWRTHAHFRQHRFSSRATPAMA
Tribolium_TC014213 -----------YVAYITSPQLTTTASPLRTLVSSEIRAFQGNTTQAEVAKNMTVAPPSSN
Tribolium_TC008953 -----------F------------------------NTTRDTTHYVLN-----------
Ixodes_ISCW003730  -----------FGYGLLYRPLSFGS----------LAGRPRPDMSWLLAQQDIRQL----
Dappu-325474       IRYNKVVIEKKFLSDYMVSVWDTRIIDEEAEKAKPIKDRMQDYIRYSVARLGLHEL----
Dappu-241507       -----------LFDYLAFH-LRDK-----------EYDGIENYIRFMTANLGLKSLPISS
Dappu-241308       -----------IISSLVVPSLINT-----------PYPGVANYTLYETARLGLLI-----
Dappu-111641       -----------FFDYLANHLRDT------------PYPGVGNYIRYTVARLGLAPL----
Dappu-316941       ----------------------------------PYPGAENYTRYTVARLGMLPL----
Dappu-314238       -----------FFEYLASQLRDT------------RYPGVETHTRYVVAKTRRKYL----
Dappu-56803        ------------------------------------------------------------
Dappu-14718        ------------------------------------------------------------

Ixodes_ISCW018107  SHPPGLHEDVNP---YPFGYVLNKPDL---C-------------------ATGSKILVLI
Tribolium_TC014213 SGGSSEAPQLPAVRTLTNATNSSQPDLTRGVAAEIIYEAGHVDVSSQICPELGRDLKLLI
Tribolium_TC008953 ----------------TNLSAHIWPEH--FC-------------------DLNSFLLVMV
Ixodes_ISCW003730  ---------------ISNGSLLLSPKD--PC--------------------PSFLAVVI
Dappu-325474       VDVPKLKPTFGVV--YNDILSFQYPINTPGC--------------YKNGSASSRFPSLFV
Dappu-241507       AGNRMLPGMEGLV--VNDISWFRYPIDIGPC------AAAGLDGSINSQNASLHRRSLFV
Dappu-241308       SNANSIVPEFGPV--LNAVTSLNYPITIKRC-----------------GDIEQRRQSAFI
Dappu-111641       VGVEPLLAEFGPV--INDVLSFTYPISIPPC-----------------QHHFLANQTIVL
Dappu-316941       AGIEPLKPEYGPV--INNFTSFRYPITISPC-----------------QKVKTDYPSVFI
Dappu-314238       LNVKPLRPDFGPV--LNDVTSFNYPIQISRC---------------RDPIVRGGPSLFV
Dappu-56803        ------------------------------------------------------------
Dappu-14718        ------------------------------------------------NPSVFI

Ixodes_ISCW018107  AVM--TASGNFNQRRAIRDTW------GKESLHRGFKLV--FLLGLPR-YDVLQRSILAE
Tribolium_TC014213 AIT--SAPSHESARMAIRETW------GHFASRKDVAIA--FMLGSIS-NETVNANIEKE
Tribolium_TC008953 C----SGPANFEARSAIRDTW------GHERIILGNNVSLFFLLGETT-NSSLQYDIMLE
Ixodes_ISCW003730  C----SAVNNFVARRAIRDTW------GQDARSPLVRAF--FLLGRTD-NETLQEDVVRE
Dappu-325474       AVTGLATDNNQKRRDEIRTLFKKEI--PELKHNLFSSINYAFFLGQPG-DLVHQSLIEQE
Dappu-241507       SVI--SGPNNFERRAAIRRTWPAHLRNQSNLNHPLDVVGFGFLIGLTN-DSVVQQKVKEE
Dappu-241308       AVI--SAADNFEKREKIRQTWKSHI--DFVRKFKLFNIQFSFILGQSE-DAFTQRKIQEE
Dappu-111641       LVN--SAPGNFDRRKIIRQTWKNHFKAPHIDADRLGIAGFAFVLALTD-NNVTQNQIEQE
```

```
Dappu-316941       AVV--SAPENFEKRNIIRQTWRTHLN-LEYHEKLMNIIGFAFILGMSD-KNVTQIKIEEE
Dappu-314238       AVI--SAPKYFHKRDIIRRTWQRHLQ-MQSDLNSMNLAGFGFIVGLTQGDDGIQKRIEDE
Dappu-56803        ----------------------------------MARFGFFLGQTR-NDSIQKRIEEE
Dappu-14718        ALI--SAPDHFKERNDIRETWLIHLK-SVLEKNLLGMARFDFFLGQTR-NDSIQKRIEEE
                                                          *.:.       :  :  *

Ixodes_ISCW018107  DSLHADIVQGNFTDCYRNLTFKSVMMVRWASASCPG-AEFVLKIDDDVLLNVWDFAPTLS
Tribolium_TC014213 QYLYGDIIRGKFRDTYDNLTLKTISMLEWVDNYCPK-AAFVLKTDDDMFINVSRLLAFIA
Tribolium_TC008953 SDRFGDIIQERFIDSYNNLTLKSVFMLKLVSSYCANSTKYLLKIDDDMFVNMIPVVRMLR
Ixodes_ISCW003730  SRLFGDVIQADFMDTYNNLTVKSVVLLKWTGQQCPQ-TRYILKTDDDMYVNVPNLVSYLN
Dappu-325474       SDKNKDVVQVDMMDNGKNDSLKLAAIFNWVQQFCTN-VDVVFKMDENF--EIATLKKFGS
Dappu-241507       SETFGDILQVNMIDRYVDLSVKLASLFNWVDTYCPR-VDFVLKVDDDVYVNVHNLATVLH
Dappu-241308       SKTHDDIIQFEMLDTHRNLPLKMAGLFNWVNTICPK-LDFLLKLDDEMYLNVHVLANFVN
Dappu-111641       ANTHGDMIQIGISDFYRNLSLKVAGLFHWLYSNCAR-VDFVAKLDDDVYVNVRNLARFVQ
Dappu-316941       SKTHKDILQIEIPDIYYRLAVKVAGLFNWLHRYCAQ-IDFLLKVDDDVYVNVRNLAHFVN
Dappu-314238       GKTYGDILQIEMIDDYYNLTFKVVGLLNWVNDHCSR-VDYVLKVDDDVYVNTHNLVAVMN
Dappu-56803        SQKHGDIVQIEMDDSYRNLTLKGIAVLNWVRQHCAK-VDLVFKVDDDVYVNVHNLVHFVR
Dappu-14718        SQKHGDIVQIEMDDSYRNLTLKGIAVLNWVRQHCAK-VDLVFKVDDDVYVNVHNLVHFVR
                        *::.  : *     ..*   :.      *.     : * *::.  :   .

Ixodes_ISCW018107  ALHGV---DRTIWGL-------LAQ---RWTPERNPRSKWYVSWGMYQNATYP-DFLTGP
Tribolium_TC014213 KHSPE---QRTIYGR-------LAK---KWKPIRNKKSKYYISPNQYKPAVFP-DFTTGP
Tribolium_TC008953 DRNST---TDLLMGK-------LIC---RARPIKDTTSKWYSPRYMYPHHVYP-NYVSGT
Ixodes_ISCW003730  KKGG----RKMLLGC-------LIS---GATPIRDWTSKWYVPPFVYPHHTYP-DYLSGT
Dappu-325474       ALTEKEIPDTFVYG--------VKG---DIRPQR-EAGKRMITMEEFPWTTFP-AYFNGL
Dappu-241507       SLTVA---DQSIYGR-------QCG---GMIPDR-KGGKWMTSYENWPWHKFP-IYFQGA
Dappu-241308       TYRQLG--KMTIFGQSPRKGYPFIN---NWGPQR--SGMHEIALEEWPWNTYP-NYVNGP
Dappu-111641       TYRHQS--NQSMFGS-------AAG---NLWPAR--DGKWNMTFEDWPWNEYP-PYFLGP
Dappu-316941       EQKVQPSINQTLFGS-------YIGYGRDYIPDR--EGKHFISYEEWPWTRYP-RFFNGP
Dappu-314238       NLNSS---EHSMYGS-------FA----EGLPNR--GGKWYISFEDWPWSNYP-TYFRGA
Dappu-56803        SNYQS---NNSVFGH-------AWG---ETYPHRYKDSKYYISLEEYPWSNYPYNWLSGP
Dappu-14718        SNYQS---NNSVFGY-------VWS---EPYPNRYKDSKYYIPLEEYPWRHYP-NYVNGP
                           : *                 * .   .    .   :    :*  :  *

Ixodes_ISCW018107  SYLLSGDSVPLLARASDSVPYLYLEDVFLTGLVAEKAGVRRVHNDGFLN-----YRKFFT
Tribolium_TC014213 AYLLPARLSKELYVAALNHTYFKLEDVFVTGIVANSLKIKRVHAPEFLN-----KRVSLT
Tribolium_TC008953 GYVMSVDVAEKLYKAALKTPIFHLEDVYTTGLCAKRAGVRPKNNPLFTY-----QSMNYD
Ixodes_ISCW003730  GYVMSGDVLGQLFRTALETPFFYMEDIFVTGMVAQKVGIKPVNYDAFKF-----YKRKNN
Dappu-325474       AYFITGNMIVPLMAAFQTVPMLPLEDVYL-GICIIKSDMKRYTYCG--------RDINNS
Dappu-241507       GVVIAGSAVRPILSAMQVTPYFIWEDMYLVGLCAAKAKVQLRTSNQ----------
Dappu-241308       AYLIHQTAILPLLAAIQTTPIMPFEDIYITGICSEKAGVVTQYSSGYNR----------
Dappu-111641       AVLFPSSTILPLLAALQTTPMMPIDDVYYSGMCTEKAGVVLRFSTNSTR----------
Dappu-316941       GVVISGNSILSLLAAMQTTPIMTSDDVYYIGICTEKTNITLHFSSKSTSVFSMECPDLSR
Dappu-314238       AILMPGITIGPLLAASQTTPFLPFDDTFLTGLCTAKAAITVRISDRFFV---GGATEVPE
Dappu-56803        AYFMHASVVIPLLAASQTTPLHPFEDVFLTGMCREKAGVKIRNSIDQRQ------QLWFM
Dappu-14718        AYFMHASVVIPLLAASQTIPFNPFEDVFLTGLCTEKASV----------------------
                   . .:       :  :    .    :* :  *:      :

Ixodes_ISCW018107  PCTTPRVIASHGYTPLYLRHVW-----------------------------------
Tribolium_TC014213 PCSVQKGISIHMVKGVEQYDLWKKLHDVAAKCKK-------------------------
Tribolium_TC008953 VCLYMRLYTAHRFTPSDIRKTYTLLKD--SNVTRECTYHRGRSNLSVNWLMNNILKVNKP
Ixodes_ISCW003730  PCVFRKLITAHIMTPSELRSMWSRVRDRRIKCS-------------------------
Dappu-325474       PCF-----------------------------------------------------
Dappu-241507       -----------------------------------------------------------
Dappu-241308       -----------------------------------------------------------
Dappu-111641       -----------------------------------------------------------
Dappu-316941       PCNLRLFVSWLTSSGSLMNKSHVAIEDFYQNKTQCVVSTGSNGTNTTINQNEPVHFYFDP
Dappu-314238       PCHVYTSITWLTDSVAQLNNSRWATENFYNNLTHCTLNDPGGANQTVNSKKDKFHFIFST
Dappu-56803        KPYFEGTHRFLFYRPIHQSPK-------------------------------------
Dappu-14718        -----------------------------------------------------------
```

Table S48. Alignment of Enzyme 2.4.1.152 *Daphnia* proteins, with *Tribolium castaneum* and *Ixodes scapularis* orthologs, using MUSCLE [S58].

Protein ID	Description

```
Dappu-328684       ------------MHIFKTLFSQLDRHRVFYLLWLLFLLNVFTFKQLTLNEDDVDSKELVI
Dappu-244685       ------------------------------------------------------------
Dappu-104196       -----MNNRRRFQNVFYKLNLHRVLYPLWLFFLFNVFTLKQLTIDETENEVKELDIVKHI
Dappu-319378       ----MNNRRRFQNVFYKLNRHRVLYPLWLFFLFNVFTLKQLTIDETENVEVKELDIVKHI
Ixodes_ISCW004236  SLPRKGRSCPIWPKMTPSVRTSIFILTSLLLLLWLFSFSAFRPTFRQIVGVWSPVKWSYY
Ixodes_ISCW024758  ----------------LGTFWYSKSKFQRCTQESVLVNGSNENVYMNRLENVMYNYSLWT
Ixodes_ISCW023318  -------------------MVHSFPVIMVTRKFAIKFIFVAVLICACFVILYVAPRLLT
Dappu-41601        ------------------------------------------------------------
Dappu-48653        ------------------------------------------------------------
Dappu-331779       ------------------------------------------------------------
Dappu-67046        ------------------------------------------------------------
Dappu-55591        ------------------------------------------------------------
Dappu-334524       ------------------------------------------------------------
Dappu-302891       ------------------------MSPYRRFVVGILIIVLLTRFYNKVAFYKNEENEK
Dappu-60476        ------------------------------------------------------------
Tribolium_TC008651 ---------------------MVTFHKLLFIILGLLIATIFLLSQYRTELTPPQTL-KR
Tribolium_TC008652 ----------------------------MAQTIQLEQATIFLLSQYRTESTPAQTLIKR
Dappu-227431       -------------------MSNNLLNTVSQKRIALFATLLCILILTFFVFNTPIFHSHD
Dappu-251980       ------------------------------------------------------------
Ixodes_ISCW003590  ---------------------------------------------MFFRGLVAVFLVTTC
Dappu-318584       ------------------------------------------------------------
Dappu-316980       ---------------------------------------------MAPAFAYHFLIIEKG
Dappu-312894       -------------------MVFILFLCISSVVAFLVYFHETNVAPSEFVNQTSTGEIVS
Dappu-236411       ------------------------------------------------------------MN
Dappu-60056        ------------------------------------------------------------M
Dappu-3751         ------------------------------------------------------------
Dappu-107642       ------------------------------------------------------------
Dappu-253741       AAFSTKMKWSESSFQLVAALWRRKHKVLVLLFCFVFVIGVRQLDFNQEDKVIETEEKLMP
Dappu-13230        ------------------------------------------------------------
Dappu-219820       ------------------------------------------------------------
Dappu-315514       ------------------------------------------------------------
Dappu-25363        ------------------------------------------------------------
Dappu-315506       TSFIFVVCFLAFINYQHLGVTHSSILPKFSISRASQQQVSHANDAENTNNNTFKNLNKKT
Dappu-25935        ------------------------------------------------------------
Dappu-260935       --------------------------------MYANNALRSTADHHVAKEEKDSDHLL
Dappu-19438        ------------------------------------------------------------
Dappu-52155        ------------------------------------------------------------MQ
Dappu-302634       ------------------------------------------------------------
Dappu-66315        ------------------------------------------------------MGDNV
Dappu-66309        --------------------------------------------------------MVDN
Dappu-302400       FHRHWIAIQSSLSKQTGFSSKIVSYWLASSFLIAIIFLYYLLLAFWDNKLLFRQNQQVVP
Dappu-56240        ------------------------------------------------------------
Dappu-65379        ------------------------------------------------------------
Dappu-4136         ------------------------------------------------------------
Dappu-266638       LLRKRIRSKMMCRRYPGICAGRRRPWVIRPRPVKRRPLVSRPTFRPTMTRFPCVSCGRRG
Dappu-13713        ------------------------------------------------------------
Dappu-4141         ------------------------------------------------------------
Dappu-266923       ----------------------------MHFSNEKGSSASNDPTSEKIHKGPVIVRNKR
Dappu-272135       ------------------------------------------------------------MYR
Dappu-116054       LKLTRLSLAKRIILVAIGIVFFLAALIRRDDDGRTSPALPNPFDSISFQPIRETNLVVNR
Dappu-331784       ----------MGLSAAGLFLTSAAFLYWNEMNNQQQLITFSQSTNKDVTGKIAANVNMKI
Dappu-67045        ------------------------------------------------------------
```

```
Dappu-316572       ------------------------------------------------------------
Dappu-302457       ------------------------------------------------------------
Dappu-64359        ------------------------------------------------------------
Dappu-4083         ------------------------------------------------------------
Dappu-53630        ------------------------------------------------------------
Dappu-111600       ----------------------------------MTSILHNKIGGEPTLNNWVNRPNN
Dappu-15329        ------------------------------------------------------------
Dappu-58299        ------------------------------------------------------------
Dappu-248921       ------------MSLLTTRQLVVNRLDLVYKNKRRLAAIFRTSVLAVGIFTLVIYYQSL
Dappu-325563       PEIEDEPSASEDEGQKAGTEEEEPVVEPLIPKDLRVQLERLVLSPATLQRLKRVQKAKKR
Dappu-23160        ------------------------------------------------------------
Dappu-221393       SEPCKIATKTTQCQQEKGDQHERSVSGLLPVPTKLKTMVKGVARNSVPEERYSYRDVILN
Dappu-3818         ------------------------------------------------------------
Dappu-58354        ----------------------------------MTVIVHLRPERAERSGNKQDNFVIP
Dappu-311402       ------------------------------------------------------MSLDKS
Dappu-106945       ----------------------------MKTKKSLVIFGLTNLVFLLQLYTPVEYSVFH
Dappu-198878       ------------------------------------------------------------
Dappu-313025       ---------MKQGTISISKLNCKKIFVIFLILNVVLAVVLFGRQNENFSFQQLYSPFNTS
Dappu-313010       ---------MFLKDNNRWSEKNETAMEINQVVINALFGIALFGNHDTSVNVRQFYSSFDK
Dappu-24623        ------------------------------------------------------------
Dappu-58316        ------------------------------------------------------------
Dappu-308012       ---------MAANIKKCCKLTAAAFIVSTFFLLCADRLARPDSNPHRPVELFQQPNHQCI
Dappu-67044        ---------MAANIKKCCKLTAAAFIVSTFFLLCADRLARPDSNPHRPVELFQQPNHQCI
Dappu-260055       SECDLSSPPPIQPNANASAIVYRGLRRWRPPPYGLTDPDRSVSFRLVQPVQQQENIHPQF
Dappu-63087        ------------------------------------------------------------
Dappu-316372       --------------------MWLLLLMILPEHIRNRRSVKVAAFLVVVGTLVFWWYNFP
Dappu-68594        ------------------------------------------------------------
Dappu-325685       -----------------MDISRRLRFSLELFTFLAVIFIWIYWRFTLPPPIEDDRKCNT
Dappu-3750         ------------------------------------------------------------
Dappu-336888       ------------------------------------------------------MVRETDQN
Dappu-266928       ----------------MKSLYAKRISVQKLATLLVIVALCYLASSSRNQAASDYRQSVE
Dappu-49339        ------------------------------------------------------------
Dappu-316587       ------------------------------------------------------------
Dappu-241186       ----------------------MCSLFTNRITKVKILAAFLVLGTLVFIHSLNNGSLLT
Dappu-49176        ------------------------------------------------------------


Dappu-328684       IKRIDRSAFRQLDMKKILMWNP-WYGDFG-----FALDDD---------FAFSRVGCK-F
Dappu-244685       -----------------MWNP-WYGDFG-----FTLDDD---------SSFSHMGCK-V
Dappu-104196       EHQSSASFQESEKVKTILMWNP-WYGDFG-----FTLDDD---------SSFSHMGCK-V
Dappu-319378       EHQSSTSLQESEKVKTILMWNP-WYGDFG-----FTLDDD---------SSFSHMGCK-V
Ixodes_ISCW004236  PWYNRDSANVTKEVPRILLWTS-FYGTWIGSLNNSRTGE-----------MLTTKCS--
Ixodes_ISCW024758  PFFDRKGYYGGKDLPRILLWTA-IYGKWH-----SGLTDQRV-------DELEFAGCP--
Ixodes_ISCW023318  KPSPRGGLGLNLTPKYVLTWTPFFGDVDY-----IPSGQ-----------LESTKCGGI
Dappu-41601        ------------------------------------------------------KLNCP-I
Dappu-48653        ------------------------------------------------------CP-V
Dappu-331779       ---------------------------------------------MXXXXXXXXXXXX-X
Dappu-67046        ------------------------------------------------------CP-V
Dappu-55591        ------------------------------------------------------CP-V
Dappu-334524       ---------MNTTFKRILFWNEGNNKNKN-----YGVGH----GR----DVLQKLGCP-V
Dappu-302891       LLSVNNVKPSQEPVKKILFWYEHYTLRKGQNRIRVGVGQ-GPAGISSDSQALRLSHCP--
Dappu-60476        -------MQSDIQWKTILFWNEAYAGNKT-----FDIGI----GK----DKFLKANCP-V
Tribolium_TC008651 NVDDKPTFLDNNDTKTILYWTP-MFQSLN-----FYLGL----GS----KIFEK--CA-Y
Tribolium_TC008652 SVDGKPTFLGNNVTKTILYWTP-MFQSPH-----FYLGT----GS----KIFEK--CA-Y
Dappu-227431       LSEYETIKKNCSGKQLVLFWTK-FFETDD-----FYVGL----GI----KPFKQ--CT-V
Dappu-251980       -------------------MKV-FLRSLP----------------------NNNSCP-F
Ixodes_ISCW003590  VVVYNLSGDHNPKPLTILLWTT-WSGKES-----YPYFK----------EDVVDSKCP--
Dappu-318584       ------------------------------------------------------------
Dappu-316980       FSNEIEKLTHFLPTKNILLWPNSYG---------YRFGI----GR----QPFVDNGCR-V
Dappu-312894       QTETWKRISVINGTKNILLWKN----AWG-----FRFEL----GR----AAFLNSGCR-V
```

```
Dappu-236411      DTDNWKRISVIKGTKNILLWNE----LWG-----YRFGL----RK----AAFLNSGCR-V
Dappu-60056       AYVQRAVDSSDGEPKIILFWTK-YHGSAS-----FDFGL----GS----RPFETAGCR-V
Dappu-3751        ------------STKTILIWNP-YSRFEL-----EVFGE----GA----DTFTSHHCP-I
Dappu-107642      -------MEQPAEFKTILYWND-FFGIED-----FTFGL----GR----EPFIQAQCP-I
Dappu-253741      FNDIEEEEAGPSALKTILYWNS-FFAFKD-----FNFGF----GQ----QPFLDAKCP-T
Dappu-13230       ---------------ILYWNS-FFGKKN-----FAFGF----GQ----QPFVNAKCP-T
Dappu-219820      ------------------------------------------------------------
Dappu-315514      -----------------------------------M----GD----QPLIKAKCP-V
Dappu-25363       ------------------------------------------------------------V
Dappu-315506      NQIDLSNQSSSDALKIILLWSTWSSTMAD-------------------EPLVKARCP-V
Dappu-25935       ------------------------------------------------------------I
Dappu-260935      TTTPNNTTANESPLKIILMWNAWTNHIAD-------------------KPLVKGQCP-V
Dappu-19438       ------------------------------------------------------------
Dappu-52155       HDTCEEIRPLSHRRKTLLLWTPVSFTWSI--------------------RWSLKNCPAV
Dappu-302634      -------------------------------------------------MDLNEQCP-S
Dappu-66315       VSPPLACPIQKDKKIIILYWTTFYHFVDF-----ANAGL----GG----KPFAHCESLGV
Dappu-66309       LSPPPATPIHFDKKFIILYWTKYYHFLDF-----GNIGL----GG----KPFVTCDRAGI
Dappu-302400      IGNVSGLSAPCQRKILILYWTK-YFTSVD-----FEYGL----GR----TPFATCDDNRI
Dappu-56240       ------------------------------------------------------------
Dappu-65379       ------------------------------------------------------------
Dappu-4136        ------------------------------------------------------------
Dappu-266638      GRFRQQCPNVNYIQIIILYWTK-YFGASD-----FGFGI----GR----KSFAQCDQTCS
Dappu-13713       ----------------------------------PFGS----------CIDRVTASS-S
Dappu-4141        ------------------------------------------------------------
Dappu-266923      FAQETIKQAESNKPKIILYWNK-YFNHSD-----MGFGV----GQ----EPFIKAGCK-V
Dappu-272135      FAQETIKQAESNKPKIILYWNK-YFNHSD-----MGFGV----GQ----EPFIKAGCK-V
Dappu-116054      EARQVEKVEVKSQPKIVLYWNT-FYNQTD-----MTFGF----GR----QPFVDAGCQ-I
Dappu-331784      TTIKSTTVKQLETLKIVLYWNT-FFNQTD-----MTFGF----GR----QPFVDAGCQ-I
Dappu-67045       ----------NREPKRIVYWTG-YYDRAD-----MIFGM----GQ----EPFIKAGCK-V
Dappu-316572      -MQEIVETKPSPPDKVILFWTP-YYNDSD-----YTVGF----GR----DPFVKNGCQ-F
Dappu-302457      --MGAAKKEPESPPKTILYWTP-YYNRTD-----FTVGL----GQ----DPFIKFGCK-V
Dappu-64359       --------------KTILFWTP-YYNHSD-----YTFEL----GQ----DPFIKFGCK-V
Dappu-4083        ---------------ILFWNE-YFQSKN-----YEFGL----GH----EPFVKAGCQ-F
Dappu-53630       MPSRLASEPKWEKLKRILYWTE-YFGTKD-----YPFKL----GD----QTFREAKCR-V
Dappu-111600      NRHQRNKKDHSLMFKKILFWNS-YFSSKD-----FELGL----GR----TAFKDAGCR-I
Dappu-15329       ---------------ILFWTP-FFGVKD-----YGFGI----GR----DAFTKAKCP-V
Dappu-58299       ------------------------------------------------PFMTAGCR-Y
Dappu-248921      RKLEIAGTQQYRDVKSILIWNA-PERAEV-----VAFVNSARDGR----NVF--NSCP-I
Dappu-325563      KEREDSSESGSSPPKRTRKSGKTAEHTQLIETAVFGFGH----------QPFLDHGCE-V
Dappu-23160       ------------KIKSILFWNG-PRRSEM-----TIFGT----GH----DAFVQQECP-I
Dappu-221393      EKEISANNQTSLGYKNILIWNE-ADRTET-----ANFGI----GH----DPFVEHKCE-V
Dappu-3818        ------------------------------------------------------CE-F
Dappu-58354       NRTANVTRNNNNRYQSILIWNS-PDRIET-----SAFGL----GH----EPFIRNGCQ-V
Dappu-311402      VIDNLFPSLFLSKRKTILIWNS-AHRIET-----AAFGI----GH----EPFVQYGCE-I
Dappu-106945      TFSKITGLSNNRKNKTILIWNS-PQILDT-----APFGF----GH----EPFIAHGCE-V
Dappu-198878      -MIPWIAA---RDIKTILIWNS-AHRIET-----AAFGF----GR----QTFFQHGCD-I
Dappu-313025      LFDVFQSAPTLRENKTILIWNSAHRIETA----AFGFEL----------DSFRRHGCE-V
Dappu-313010      SVSDFLSRVTYRGNKTILIWNSAHRIETA----AFGFGY----------QPFIQHGCE-V
Dappu-24623       ------------------------------------------------------CE-V
Dappu-58316       --------------------------M-----TIFGT----GH----DAFVQHGCP-V
Dappu-308012      DLVENNSTTVQGRIKTILLWNA-PQRPEV-----VIFGT----GH----DAFVQRGCP-V
Dappu-67044       DLVENNSTTVQGRIKTILLWNA-PQRPEV-----VIFGT----GH----DAFVQHGCP-V
Dappu-260055      QNTFQNRVLVSKRGKSILLWNS-NENERF-----FRHHS--------------GSCG-S
Dappu-63087       ------------------------------------------------------------
Dappu-316372      FQKVHDYILQSKRGKSILLWNS-NENERF-----FRQHS--------------GSCG-S
Dappu-68594       ------------------------------------------------------------
Dappu-325685      NNIIIPKGNNNYPVKKILLWNA-SQRKEV-----RAFGV----GQ----DVFARKRCA-F
Dappu-3750        ------------------------------------------------------CP-V
Dappu-336888      ESVESSGFIQKNVTKTILLWNG-VRRKEV-----RVFGQ----GD----QVFVNQSCP-V
Dappu-266928      SKEHHQEETEKNATKRILLWNG-SRRVEV-----QVFGK----GQ----DAFAKQNCT-Y
```

```
Dappu-49339        ------------------------------------------------------------
Dappu-316587       ------------------------------------------------------------
Dappu-241186       RSAQKAVKIIRNETKTILVWNG-SGRKEV-----RNFGW----GK----DAFINKNCP-Y
Dappu-49176        ------------------------------------------------------------


Dappu-328684       --TNCILSK-NKNI---VTPERADAIVFLYTNLCE--------LPKVHG--RQEYQRFVL
Dappu-244685       --TNCVLSK-NKTK---VIPEQADAIVFLYTNLCE--------LPKIHG--RQGFQRFVL
Dappu-104196       --TNCVLSK-NKTK---VIPEQADAIVFLYTNLCE--------LPKIHG--RQGFQRFVL
Dappu-319378       --TNCVLSK-NKTK---VIPEQADAIVFLYTNLCE--------LPKIHG--RQGFQRFFL
Ixodes_ISCW004236  --RSCILTN-DRSL-----LESSDAIVFHIRDIDM--AN----LPQR----RSPFQKWVF
Ixodes_ISCW024758  --DRCYITN-DRRL-----LHSSDAVVLYGTDLDL--AD----MPWR----RYRGQKWVY
Ixodes_ISCW023318  SIPPCVVTS-NRSL-----LNESDLVIFHMRDIRA--DD----LPAE----RPPGQRWAL
Dappu-41601        --TDCFTTD-NRSM--LKTAAEFDAIVFHLRTFNI--ED----LPPT----RGQNQRWIF
Dappu-48653        --WQCETSD-NRNH-----VQDYDAVVFHLRSWSR--ND----LPQR----RSPHQRYIG
Dappu-331779       --XXXXTST-NRTD-----VHDYDAVIFHMRGSWD-PND----LPQR----RSPHQRYVF
Dappu-67046        --WQCETST-DRTN-----VHDYNAVIFHMRGSWN-PNE----LPQR----RSPHQRYVF
Dappu-55591        --WQCETSD-NRTN-----VQEYDAVVIHLRTWNK--KD----LPKL----RSPHQRYVF
Dappu-334524       --WQCEISA-NRTD-----VHTYDAVLFHLRTWSK--ND----LPHP----RLANQRYVF
Dappu-302891       --YQCDIFN-RERIVDWDTLKYYDAIVFHQHGWTP--ND----VPMK----RWPHQHYIF
Dappu-60476        --WQCKTIS-DRNI---LLIESYDAVVFNQRKWTP--TD----LPVN----RSGHQRYIF
Tribolium_TC008651 --NNCYATY-VKNE---RPVEKFDAIIFHGVEYQEKWFG----KPQK----RNPNQVYIF
Tribolium_TC008652 --KNCYATY-VKNE---LPVEKFHAIMFHAVEYQEKLFG----KPQK----RNPNQFYIF
Dappu-227431       --FACCSTN-NRNF-----LDVSDAIIFHIRDLDL--ND----MPPR----RSVRQRWIF
Dappu-251980       --RNCRVTT-DRRS---NLLGKFDAIIFNMAVLHQLATDK---LPPADT--RESHQYYIF
Ixodes_ISCW003590  --QVCLFTR-QRRH-----LKSSAAILFHGKDIYL--ND----MPSY----RSPQQRWIF
Dappu-318584       ----------------------------------------MP--------------
Dappu-316980       --SNCLITY-NSTL---MPHWQFDAFLVHPPTING----------PYILKDRRPDQMFVM
Dappu-312894       --TNCLITY-NNTL---MTHDKFDAFVIHSPTQHT--------PWILKD--RRPDQMFVM
Dappu-236411       --DNCFITN-NASL---MPHENFDAILVHPPTQKT--------PKEFKN--RRADQIFVM
Dappu-60056        --SNCKTTT-DRLL-----LNESHAIIFHSGNLNM--SD----MPPV----RFDHQRWIF
Dappu-3751         --NNCFITK-NRTW---APLHQFDSIIFNMPPLSL--YK----FPVDEH--RRPEQRYIF
Dappu-107642       --STCQVTN-DRSQ-----FNGSQVVVFSAQNLNF--SD----LPPH----RFPHQRFVF
Dappu-253741       --ATCFLTN-DRTL-----FNQSDVVIFSVQQMNL--TD----LPPY----RFAHQRFVF
Dappu-13230        --ATCYVTD-DRSL-----FNRSDVVIFSIQGMNL--TD----LPTH----RFPHQRFVF
Dappu-219820       ----MNLTD----------------------------------LPTH----RFPHQRFVF
Dappu-315514       --TACLFTP-DLTL-----FNQSDVVVLSVETT----PD----FLVN----RLPHQRFVF
Dappu-25363        --TSCIFTP-DRSL-----LNHSHVVLFFANNETKRNDA----LPEH----RQPHQRFVF
Dappu-315506       --TSCIFTA-DMSL-----IHQSDVVVLYVDTLTD--------FPLN----RRPHQRFVF
Dappu-25935        --RSCVFTT-DMSL-----INQSDVIVLHFDTLED--------FPLN----RQPHQRYVF
Dappu-260935       --KSCLFTT-DMSL-----MQQSDVVVLHFDTLED--------YPVN----RQPHQRFVF
Dappu-19438        ----CLFTT-DMSL-----LQQSDIVVLHFDTLED--------YPIN----RQPHQRFVF
Dappu-52155        --KGCRLIS-DRRL-----LNESDAVIFHFRNGSF--DR----LPTC----RRPDQRYVY
Dappu-302634       --HNCRLST-DRRL-----LNESDAVIFHFWNDKL--DR----IPTY----RSPHQYYVY
Dappu-66315        NGDGCVVTT-DRNL-----LNQSDAVMFHFRCFDL--ND----MPPPAW--RRPRQHFIL
Dappu-66309        N-SGCMATT-DRNL-----LNESDAVIFHFRTINV--SD----MPPPEW--RRPQQHFIF
Dappu-302400       ---VCLTTM-DRGL-----VNESDAVIFHSRDLRD--ND----LPPPGW--RLPHQHYVF
Dappu-56240        ------------------------------------------------------------
Dappu-65379        ----CVTTT-DRRL-----LNDSDAVIFHARDLHP--ND----LPPPGQ--RRPHQNFVF
Dappu-4136         ----CWTTT-DRGL-----LNRSQAVIFHARDLDP--DD----LPPPGW--RRPHQQFIF
Dappu-266638       --ENCLTTS-DRNL-----LNKSDAVIFHGRDLKD--SD----LPPPEW--RLPHQHFVF
Dappu-13713        --SNCLTTT-DRGL-----LNDSNAVIFHGRDLHV--QD----LPLPEW--RRPHQIFIF
Dappu-4141         ----CLTTT-DRGL-----LNDSGAVIFHGRDLHV--ED----LPPPGW--RRPHQMFIF
Dappu-266923       --NNCIATS-DRSL-----LKESDGVIIHAGDYSE--ND----LPIY----RSPHQRFIF
Dappu-272135       --NNCIATS-DRSL-----LKESDGVIIHAGDYSE--ND----LPIY----RSPHQRFIF
Dappu-116054       --SNCIATN-DRRL-----FNRSDGVIIHAGDYLE--HD----LPTY----RLPHQRFIM
Dappu-331784       --SNCIATN-DRRL-----FNRSDGVIIHAGDYLE--HD----LPTY----RLPHQRFIF
Dappu-67045        --TNCWATG-DRTL-----LEQSDAVIFHAGQFNL--SD----LPSK----RLQRQRYIF
Dappu-316572       --TNCITTA-DRNS-----LDKSDAVIFHAFQVNS--RD----LPAQ----RHPRQRFVF
Dappu-302457       --TNCIATA-DRKL-----LNQSDAVIFHALQVNS--RD----LPTH----RHPHQRFIF
```

```
Dappu-64359         --TNCIATA-DRKL-----LNQSDAVIFHALQVNS--RD----LPTH----RHPHQRFIF
Dappu-4083          --SNCMTTD-DRQL-----LNVSDAVLFHAMDFDE--LD----FPSLVN--RRPDQRFIF
Dappu-53630         --SNCQLTD-DRSL-----LNSSDAVIFHINDFDD--RD----LPDPLD--RLAHQRFIF
Dappu-111600        --TNCLLSD-DRRL-----LDTSDAIIFHANDFNE--RD----LPDPHR--RRPNQRFIF
Dappu-15329         --NNCMTTT-DRNL-----VNQSDAIIFHPFDVNV--KD----LPTY----RTAHQRYIL
Dappu-58299         --TNCLTTT-DKSL-----ANQSDALIFHPNDFDV--DN----LPRH----RLAAQRYVF
Dappu-248921        --TECRIDL-EASG----TLDTYDAIVVNFNDQFR--LID---LPEFR---RKPHQRMVF
Dappu-325563        --SDCAIFD-NETS---LPIEEYDAIVMHMCLIWL--SE----IP--------------
Dappu-23160         --SDCEIVN-SPHQYPYRPLSSFDAVIFNFNDEFW--------LTKRPHFQRQPHQRFIF
Dappu-221393        --SDCAIFTRDTSM---LPYEEYDAVIIHMLFLKM--FQ----LPNFE---RRRHQRFIF
Dappu-3818          --SDCAVFN-QQSA-ASLPLEEFDAVIVQISTMWL--SD----LPENRT--RSKHQRFIF
Dappu-58354         --SDCVIFD-NETA---LPLKEYDAIVMNMHVIWL--TE----LPYFK---RRQHQRFIF
Dappu-311402        --SDCILFD-NATSPDLLPIEDYDAIILHMHELWI--TG----HPIYN---RQKYQRLIF
Dappu-106945        --SNCIVFD-QPSI---LPLEEYDAILVHVHELWK--TR----MPDFH---RQKHQRFVF
Dappu-198878        --KECVVFD-NKTS--ILPLEEYDAIIIHMHELWQ--TQ----MPNFT---RRAHQRLIF
Dappu-313025        --SDCIVFD-NATSHELLPLEDYDAIIIHMHELWL--TH----LPEFQ---RKSHQRLIF
Dappu-313010        --SDCVVFD-NATTPELLPLEDYDAIIIHMHELWL--TQ----LPEFK---RQAHQRLIF
Dappu-24623         --SDCVVFD-NATTPELLPLEDYDAIIIHMHELWL--TQ----LPEFK---RQARQRLIF
Dappu-58316         --SDCEIVN-SPHQYPGRPLDSYDAIIFNFNDEFWL-TK----RPIFN---RQPHQRFIF
Dappu-308012        --SDCELVN-SPYQYPGRSVESYDAIVFNINDQFGVGSR----RPYADGNQRPATQRYVF
Dappu-67044         --SDCELVN-SPYQYPERSVDSYDAIVFNINDQFGVGSR----RPYADGNQRPATQRYVF
Dappu-260055        --IRCEIIS-NRSE---RPIESYDAIVVIFDDQFS--PVDPMELAEFQSESNNTNQKFVF
Dappu-63087         ------------------------------------------------------------
Dappu-316372        --IRCEIIS-NRSE---RPIESYDAIVVIFGDDFS--PVDPMELAEFQSESNNTNQKFVF
Dappu-68594         ----------------------------------ME----LAEFQSESNNTNQKFVF
Dappu-325685        --TQCEIFT-DRWE---HPLDYYDAIVVVFNDEFL--SKEDMAMPEFESG-RNPNQRLVF
Dappu-3750          --NRCEIVTSSRTE---RPIESYDAIIVVFHDELI--TSYELKMPEFPNG-RNPNQRLIF
Dappu-336888        --NGCEIVT-SRTE---RPIESYDAIIVVFHDELI--TPYELKMPEFPNG-RNPNQRLIF
Dappu-266928        --SRCEISD-NRTE---RPLEHYDAIVVVLNNEFI--SPDQLKLPEFDNK-RNASQRLVF
Dappu-49339         ----------------------------------------MPQFPNK-RNASQRVVF
Dappu-316587        ----------MTRL------------------------------------------
Dappu-241186        --TRCEMTD-NRSE---RPLEHFDAIVFVLNDEFT--SPDQMMMPDFKNK-RNASQHLVL
Dappu-49176         ----------------------------------------MPEFQYK-RNQSQRLVF


Dappu-328684        LTDDP-PMCYPRNYFE-RNNLF---GSFFNWTISYRENADV--TWKRGWIEK--------
Dappu-244685        LTDDP-PMCYPRNYFE-RDNHF---GSFFNWTISYRENADI--TWKRGWIEK--------
Dappu-104196        LTDDP-PMCYPRNYFE-RDNHF---GSFFNWTISYRENADI--TWKRGWIEK--------
Dappu-319378        LTDDP-PMCYPRNYFE-RDNHF---GSFFNWTISYREKADI--TWKKGWIEK--------
Ixodes_ISCW004236   WSMEP-PPYS-------VFAGFKYMMNMFNWTMTYRFDSDI--PVQYGQLER--------
Ixodes_ISCW024758   WSLEP-PPHCVLR----SLTYL---NNTFNWTMTYRQDSDVLDSYVLSLTKK--------
Ixodes_ISCW023318   LDYEA-PPHTP------RVPDV--LKGTFNWTITYRQDSDV-------NVLP--------
Dappu-41601         WSLES-PQYNMQ-----DIYPL---DGLFNWTMTYRRDSDV--IQPYGWIQP--------
Dappu-48653         WIMES-AAWREYMV---DNSPM---VNFFNWTFSYRWDSDI--VSPYGYVKP--------
Dappu-331779        WNLES-AEWREYL----DTSQL---GNFFNWTLTYRWDSDM--VMPYGYVRP--------
Dappu-67046         WILES-AGWFKFL----DTSPM---GNFFNWTLTYRWDSDM--VMPYGYVRP--------
Dappu-55591         FSMES-SAWRAYS----VVKSM---ENLFNWTMTYRWDSDI--VYPYGYINP--------
Dappu-334524        WSMES-AAWRIY-----SVAPM---AEFFNWTMTYRWDSDV--VAPYGYVRP--------
Dappu-302891        LSMES-SAWRFV-----DTKSM---ANFFNRTMTYRRDSDI--FNPYGWFKS--------
Dappu-60476         WSRES-PGWRYV-----NTNTM---AEFFNWTMTYRWDSDI--AYPYGWI----------
Tribolium_TC008651  SNQES-PVNT----PS-FIRDF---DNFYNWTMTYRLDSDI--LRPYGFLVK--------
Tribolium_TC008652  SNKES-PVNT----PS-FIKDF---NNFYNWTMTYRLDSDI--LRPYGFLIK--------
Dappu-227431        YLQES-PLHTPN-----ILYDL---SNVFNWTMTFRMDSDI--YTPYPVVES--------
Dappu-251980        FSQES-PFYHKENV---QIKDY---IGYFNWTMSYLPESNI--PYPYGRIER--------
Ixodes_ISCW003590   FSLEP-PTATSLS----MLEKL---DELFNMTMTYRQDSDI--TTFYGYTVQ--------
Dappu-318584        -----------------RLDAF---ENYFNWTMTYLPESDI--PLPYGRIEQ--------
Dappu-316980        FSTEP-PVHMY------HLQKY---ENYFNWTISYRTGSTF--QLKYGEIIA--------
Dappu-312894        FTTEP-PPHMP------KLDKF---ENFFNWTMTYRSGSTF--QLKYGEIVP--------
Dappu-236411        FSNEP-PDHMPS-----DMKSF---DNYFNWTMTYRRGSDF--HLKYGEIIP--------
Dappu-60056         YSFTS-PVNLA------PIPKF--LQDKFNWTMTYRRDSDIIHRYPFGAMVA--------
```

```
Dappu-3751        FSQEP-PTYIGE-----EVKLF---NHRFNWTMSYATHADI--RYHYGEIIP--------
Dappu-107642      FEMES-PVNTDPQSMLDPRTRF----SFFNWTMTYRLDSDIVQRDSYGFVVP--------
Dappu-253741      YEMES   TTDPLPLLYNRTRY----GFFNWTMTYRLDSDIVNRDAYGLVVP--------
Dappu-13230       YEMES   TTDYRPLLHNQTRF----GFFNWTMTYRLDSDIVNRDPYGIVLP--------
Dappu-219820      YEMES   TTDYRPLLHNQTRF----GFFNWTMTYRLDSDIVNRDPYGIVLP--------
Dappu-315514      FVMES   NTVDIPML-RNNLT--RYNYFNWTMSYRRDSDIVLRDFLGAVVSKNNLNDQY
Dappu-25363       VARHASIESDSLISALTEDDRI--RYNFFNWTMTYRRDSDIVFRESFGAIKN--------
Dappu-315506      AQLES-PDNTKMATI--NDPRL--RYDYFNWTMTYRRDSDIFLRDYYGSVIK--------
Dappu-25935       YHFES-PENTASDFM--DDPRF--RYGYFNWTMTYRRDSDIFLRDYYGSLVA--------
Dappu-260935      YHFES-PDNTASELM--NDSNF--RYDYFNWTMTYRRDSDIYLRDYYGSLIA--------
Dappu-19438       FHFES-PENTASTLM--NDPRI--RYDYFNWTMTYRRDSDIFLRDFYEKLNF--------
Dappu-52155       LNFES-AIRSRSSYPWGKLP-----RHFFNLTATYNLDSDFV-GLAFGGFQF--------
Dappu-302634      LNFES-AIRSRNHFPWRKIP-----HDFFNLTATYRLDSDFFGKMFYGFQFE--------
Dappu-66315       FEQES-PVHTAYYTGL-KLPLL---KDFFNRTMTYRRDSDIAYLNTHGRLRF--------
Dappu-66309       FEVES-PVHTYLPAL--RWPSL---KSYFNRTMTYRRDSDV--SNIRIDSDP--------
Dappu-302400      FNHES   HTDLNLL--RRPVF---WNYFNRTMTYRRDSDIVDLHPYET-----------
Dappu-56240       ----------------------------------------------------------
Dappu-65379       FLLES-PMHTDLKML--QMPLF---QNYFNRTMTYRLDSEV--VNTYGRIRT--------
Dappu-4136        FNYES-PVHTDLA----KLRLY--FNHYFNRTMTYRRDSDVVSLHPYGRLKC--------
Dappu-266638      FLYES   HTDLEVL--QRPVF---RNYFNRTMTYRRDSDVVDLHPYGRIKC--------
Dappu-13713       FLLES-PIHTDLGLL--QQPVF---RHYFNRTMTYRRDSDVVELHAYVFSAS--------
Dappu-4141        FLLES-PVHT-DLELL-QRPVF---RNYFNRTMTYRRDSDVVELHAYDSAVV--------
Dappu-266923      FNLETLPGLR-------HLPCF-SRRHFYNWTMTYRRDSDIYDARPYGALRL--------
Dappu-272135      FNLETLPGLR-------HLPCF-SRRHFYNWTMTYRRDSDIYDARPYGALRL--------
Dappu-116054      LLFETLPGGY-------HLPFF-ARPHFYNWTMTHRRDSDVYLSKSYGALRR--------
Dappu-331784      NNYETLPGGN-------GLPCF-SRQHFYNWTMTHRRDSDVYVNRPYGALRR--------
Dappu-67045       FLFETLPLSRDYAVYFSRAVDY-----YFNWTMTHRRDSDVYCAQHYGKIRR--------
Dappu-316572      FLYET-IPNTSIPCVGKCLPERQYLPHYFNWTMTHRRDSDVYVAEQYGAITP--------
Dappu-302457      FLQYA--------------------PHYFNWTMTHRRDSDVYVAEPYGAIAP--------
Dappu-64359       FLQYA--------------------PHYFNWTMTHRRDSDVYVAEPYGAIAP--------
Dappu-4083        YNYET-CVGEK------DMPVFVWTKDFFNWTMTYRRDSDIYDPHPYGSIRR--------
Dappu-53630       YNFETMDGFQ-------DYPFFKKTKHFFNWTMTYRRDSDIYDAWTYGAIRR--------
Dappu-111600      YNYETMVTAS-------DMPMFTQTKHFFNWTMTYRRDSDIYDVRTYGALQR--------
Dappu-15329       FFYEA------------------------------------------------------
Dappu-58299       LYYEAMASERERLSV--FTEPL---KHFFNWTMTHRRDSDIFSSHPYGSLRR--------
Dappu-248921      FTQEP-PPAL-KGY---DFRRY---ANYFNWTMTYRTDSDI--PLTYGRITK--------
Dappu-325563      -----------------NFQSM---RNYFNWTMSYRLNSDI--RLLYGRIEP--------
Dappu-23160       FTIEP-PPSNEPM----NVTGY---TNYFNWTMTYRLDSDV--PFPYGRIRP--------
Dappu-221393      LTQET-PVMMPL-----YISSL---DNYFNWTMTYKRNSDV--QFLYGRIEP--------
Dappu-3818        FAQES   SMTESLP--DIFSM---RNYFNWTMSYRSNSDI--QFLYGRIQP--------
Dappu-58354       MTQES   SMLFL----RVKTL---KNYFNWTMSYRRNSDI--QFRYGRILP--------
Dappu-311402      LTQEA-PTTLAI-----DVNEM---GNYFNWTMSYRFNSDI--QLLYGRIHP--------
Dappu-106945      LTQES-PISMHTI----DVAKM---GNLFNWTMSYKFNSDV--RLLYGRIHP--------
Dappu-198878      LSQES-PTTI--P----------------------------------------------
Dappu-313025      LSQES-PTTLPI-----DVTKF---GNYFNWTMTYKLNSDV--QLLYGRVSP--------
Dappu-313010      LTQES-PTTMPI-----DITIL---GNYFNWTMSYRLNSDV--QLLYGRVSP--------
Dappu-24623       LTQES-PTTMPI-----DITEF---GNYFNWTMSYRLNSDV--QLLYGRVSP--------
Dappu-58316       FTQEP-PPSIKQM----NISGY---RNYFNWTMTYRMDSDV--RFLYGRIRP--------
Dappu-308012      LTQEP-PPALVDQ----NLAQY---RNYFNWTMTYRMDSDV--RLLYGRIRP--------
Dappu-67044       LTQEP-PPALVDQ----NLAQY---RNYFNWTMTYRMDSDV--RFLYGRIRP--------
Dappu-260055      YTRKS-PQSLASYH---NLSEF---TGVFNWTMTYRRDSDI--PLLYGRIEP--------
Dappu-63087       -------------------------GVFNWTMTYKRDSDI--PLLYGRIEP--------
Dappu-316372      YTRKS-PQSLASYH---NVSEF---TGVFNWTMTYRRDSDI--PLLYGRIAP--------
Dappu-68594       YTRKS-PQSLASYH---NVSEF---TGVFNWTMTYRRDSDI--PLLYGRIAP--------
Dappu-325685      FTQES-PPALRSHY---NMTRF---VHFFNWTMTYALDSDI--PLLYGRIIP--------
Dappu-3750        LTQEP-PTSLKRYY---NTSQL---TNFFNWTMTYRMDSDI--PFLYGRVLP--------
Dappu-336888      LTQEP-PTSLKRFY---NTSQL---KHFFNWTMTYRMDSDI--PFLYGRVLP--------
Dappu-266928      FTQEP-PPALMPYY---NTSRF---ANFFNWTMTYRMDSDI--RLLYGRFIP--------
Dappu-49339       FTQEA-PPALRPLF---NMSQL---VDIFNWTMTYRFDSDI--PFIYGRVIP--------
Dappu-316587      ------------------------ANFFNWTMTYRINSDI--QLLYGRIIA--------
```

```
Dappu-241186       FTQES-PPALKSYY---NMTQL---AHFFNWTMTYRMDADI--RFLYGRIIP--------
Dappu-49176        LTQEA-PPALKPYY---NMTRL---ANFFNWTMTYRSDADI--RLRYGRIIP--------


Dappu-328684       ------------LDKPTKKNS--------------------------------FSQMR
Dappu-244685       ------------LLDQPLKIR--------------------------------HSFPQIG
Dappu-104196       ------------LLDQPLKIR--------------------------------HSFPQIG
Dappu-319378       ------------LLDQPLKIR----------HSFSQKLEKEE---------------
Ixodes_ISCW004236  ------------KEDLAPKKN-----------------------------------H
Ixodes_ISCW024758  ------------PEPTPYSID-------------------AL---------------
Ixodes_ISCW023318  ------------QLRRADAPD--------------------------------QMAKP
Dappu-41601        ------------IGSIGLQPE----------VEEINREMELA-------------IK
Dappu-48653        -----------VQGRVPLHPN----------EKQMKEYLSNS--------------K
Dappu-331779       ------------TGKVPLHPS----------EDQLKQLMSVQ--------------K
Dappu-67046        ------------TGQVPLHPS----------ENQLKHLMSSS------------DQK
Dappu-55591        ------------IGNVPLHPS----------ESQMKYFLSHP------------KVK
Dappu-334524       ------------IGNVPLHPS----------EAQMKFYLSNH-----------RNSS
Dappu-302891       ------------ANQAVIFDS----------SQFDSQTLPRL-------LHETHLDSA
Dappu-60476        ------------TLTNPIIGT----------KLSDTHELEQL-------IAETQFDST
Tribolium_TC008651 ------------QKTGYKLPT-----------------------------------
Tribolium_TC008652 ------------KKTDYELPT-----------------------------------
Dappu-227431       ------------TPGSALELQ--------------------------IWNTTFNRNKL
Dappu-251980       ------------AAAVDLKSK-----------------------------------
Ixodes_ISCW003590  ------------ARKRPRVTV--------------------------------KSK
Dappu-318584       ------------LASAPNGET----------ERRDMQESVRL-------------SN
Dappu-316980       ------------LESAPQTEE----------ELATLRLSAAQ------------QSA
Dappu-312894       ------------LETAPTTEK----------EAEAMRLTVAQ-------------SG
Dappu-236411       ------------LDTAPKTEA----------EATTMREKMVH-------------FG
Dappu-60056        ------------SKTMRKSYG--------------------------KTRPTANSV
Dappu-3751         ------------LPSAPATHH----------TRKAYIQSTNN--------------G
Dappu-107642       ------------RSKHPISAM---------AYPTFQHSGDVTT------------TSS
Dappu-253741       ------------IPNSTASSV----------YPKSRPAGSAP--------TPHRRNYA
Dappu-13230        ------------TQNSTSYPK----------LRRGNNAAALH---------DLKNKDE
Dappu-219820       ------------TQNSTSYPK----------LRRGNNAAALH---------DLKNKDE
Dappu-315514       PPRFETNVSRRQYNSNTTRKLS--SVDALKHLVTTREDRRKVDK-------NPFVDESRH
Dappu-25363        ----------------APELMD-----------------------------------
Dappu-315506       ------------TNHYSKLNT--------------------------AKISNISQEM
Dappu-25935        ------------KTSFNNNTL--------------------------VSRYNYENK
Dappu-260935       ------------KASIKNNSREMRYNYGNLTSNNYNDQIITDAIHILPGMELVKTDLNF
Dappu-19438        ------------RLELATFIR-----------------------------------
Dappu-52155        ------------EPKEKLTIQ--------------------------RKAFDRTNYY
Dappu-302634       ------------RLESIQPTS--------------------------DDLTNYY
Dappu-66315        ------------DPAIPMAFN-----------------------------------
Dappu-66309        ------------ADRFPVNLT-----------------------------------
Dappu-302400       ------------SASLPFQID-----------------------------------
Dappu-56240        -----------------------------------------------------------
Dappu-65379        ------------FNPATIASS-----------------------------------V
Dappu-4136         ------------RPSSGAGGN-----------------------------------DF
Dappu-266638       ------------IHPSPSCLN----------FPRLNRSVVQD---------ISSSVNF
Dappu-13713        ------------TNVTSINNI-----------------------------------
Dappu-4141         ------------   TDMNVTS-----------------------------------
Dappu-266923       ------------KRDSEIMSQ--------------------------DEPKPNNS
Dappu-272135       ------------KRDSEIMSQ--------------------------DEPKPNNS
Dappu-116054       ------------RKDKEIVNQ----------LPPRLNPGERPP-------KPEDLFTRK
Dappu-331784       ------------RNHSQLVNQ----------LPPKLNPGQRPP-------KPEDLFTRK
Dappu-67045        ------------KASSPLLEQ----------LPPALPPGQRPI-----APAKLMEHVTN
Dappu-316572       ------------KLSTFPAQL----------PDELPPGTLPA-------NPAELLNRN
Dappu-302457       ------------KYWTLPAQL----------PDELPPDTLPA-------NPAVLLNRK
Dappu-64359        ------------KYWTLPAQL----------PDELPPDTLPA-------NPAVLLNRK
Dappu-4083         ------------RRDSRSLVT-----------------------------------
```

```
Dappu-53630        ------------RTNALPPPA-------------------------------------
Dappu-111600       ------------RTNALPPPT---------SIPVRLSPEVLPPDPASMMMPNNKSFSRH
Dappu-15329        ------------------------------------------------------------
Dappu-58299        ------------KSTSVVANV--------------------------------------
Dappu-248921       ------------REKTLTNTE----------MAS-----------------------
Dappu-325563       ------------KVIAPRTGE----------EIRQLIKETHQ------------PSL
Dappu-23160        ------------KRSEDNLVS--------------------------------------
Dappu-221393       ------------EATAPKTPE----------EVEEMMAKTRH------------PLA
Dappu-3818         ------------GPTAPKTRE----------ETLRLMKANKN-------------KN
Dappu-58354        ------------GPSAPKTRA----------ETRKLIKSTRQ------------SSA
Dappu-311402       ------------GPTAPKTRE----------ETDQMVKGTTA---------------
Dappu-106945       ------------ELTAPEKLE----------ETHRMIEAMHL------------PSA
Dappu-198878       ------------------------------------------------------------
Dappu-313025       ------------LPTAPKTSE----------ETRKMIEETHL------------SST
Dappu-313010       ------------RPTAPTTAD----------ETRKMIEETDL------------SST
Dappu-24623        ------------RPTAPTTAD----------EIRKMIEETHL------------SST
Dappu-58316        ------------KPSAPKTME----------ETEVRMKESTR-------QLMLNKRKV
Dappu-308012       ------------KPSAPKTMG----------ETEVRMKESTR-------QLMLNKRKI
Dappu-67044        ------------KPSAPKTME----------ETEVRMKESTR-------QLMLNKRKI
Dappu-260055       ------------EELSFLSPE----------DVLRHIERARK--------TFRPRPKS
Dappu-63087        ------------EELSFLSPE--------------------------------------
Dappu-316372       ------------EELSFLSPE----------DVLRHIERARK-------TFRPRPKSR
Dappu-68594        ------------EELSFLSPE--------------------------------------
Dappu-325685       ------------KKSSRLSPE----------QILQLRKKSRK------------SFRP
Dappu-3750         ------------KETAPRTPN----------EIAHYREMAKN----------ISKPLL
Dappu-336888       ------------KETAPRTPN----------EIAHYREIARN----------ISKPLL
Dappu-266928       ------------KENAPITAE----------DVSRCREKARN------------NTS
Dappu-49339        ------------VGRNTPLDPG-----------------------------------
Dappu-316587       ------------KENAPRTPE----------EISNLREKARV------------SPP
Dappu-241186       ------------KEHAPEEIS----------NLREKARASLP---------------
Dappu-49176        ------------KENAPRTPE----------EISNLREI-----------------


Dappu-328684       LRTNKKKKKLVGWYVARCG-SM-SK----REGYINELKEY-I---------Q-VDSF---
Dappu-244685       MRDNWKKKKLVGWYVARCG-SK-SK----REGYINELREY-V---------E-VDSY---
Dappu-104196       MRDNWKKKKLVGWYVARCG-SK-SK----REGYINELREY-V---------E-VDSY---
Dappu-319378       AGRQPTRQPTVGWYVARCG-SK-SK----REGYINELREY-I---------E-VDSY---
Ixodes_ISCW004236  SALWKSKSVMAVWMVSHCN-TD-GR----REAYVKELKKY-L---------K-VDVY---
Ixodes_ISCW024758  RRLWEGKSTMAVWPVSHCH-TF-GR----REDYVEELQKY-I---------A-VDVF---
Ixodes_ISCW023318  FNWWTDKTRHVLWLVSNCK-TP-SN----REGFVQELRKF-I---------Q-VDVV---
Dappu-41601        RKFKNKKTKMVAWFVSNCQ-SK-SQ----REQYANVLAKY-V---------Q-VDVY---
Dappu-48653        VDYANGKSKMAAWFVSNCL-SK-SN----RNEMVNELQKH-M---------Q-IDVY---
Dappu-331779       VNYAAGKTKLAAWMVSSCI-SH-SR----RHEMVKILQKY-V---------Q-IDVY---
Dappu-67046        VNYAAGKTKMAAWFVSNCR-TQTSN----RNELIKVLKKY-I---------E-IDVY---
Dappu-55591        VNYAQGKTKMATWFVSNCKTVH-SS----RNKLIELLQRHNI---------Q-VDVY---
Dappu-334524       RNYAKRKTKMAVWFASNCR-TVVSS----RNELVKELQKY-I---------A-IDVY---
Dappu-302891       VNYATGKTRKVAWFVSNCK-SL-SA----RNEYVDRLKTF-I---------D-VDIY---
Dappu-60476        ISYTAGKTKKVAWFVSNCK-SL-SA----RNEYVDRLKTF-I---------D-VDIY---
Tribolium_TC008651 VEEIQKRPKKIAWFVSNCG-TS-SE----RELLVNEIQKE-I---------H-VDVY---
Tribolium_TC008652 VEEIQKRPKKIAWFVSNCK-TS-SQ----RELLVNEIQKE-I---------H-VDVY---
Dappu-227431       KHNVRRKKKLVAWFVSNCF-TT-SG----REKYVLELQKY-V---------E-VDIY---
Dappu-251980       -SYPRKKTKLVAWFVSHCNTTQ-SR----RENYVRELQKH-I---------P-IDIF---
Ixodes_ISCW003590  EFWRAPKPGAAVWMVSHCK-TD-SR----RETYVSELQKV-L---------P-VDIF---
Dappu-318584       VNPAKGKSKLAIWMVSNCH-AR-SN----RMKYVRQLQKY-V---------D-VDIISTG
Dappu-316980       VNPAKGKTKLAIWMVTNCK-AR-SN----RQGYVRALRKH-LTANEKNQTSSLLDIFSR-
Dappu-312894       LNPAKGKTKLAIWLVSNCN-AR-SN----RQGYVKVLKNY-M---------D-VDIISKK
Dappu-236411       MNPALGKTKLAIWLVSNCY-AR-SN----RQGYVKVLKNY-M---------D-VDIFSKD
Dappu-60056        QRRIPHKKKLVAWITSTCP-TS-VR----RENYVRQLARH-I---------S-VDIY---
Dappu-3751         ENFAAGKTKLVAWFVSHCF-TQ-SR----REKYVTIMRQY-I---------P-VDIY---
Dappu-107642       PSPTSGKKKLVAWFVSNCV-TS-IH----REDYVKQLGRH-V---------P-VDIY---
```

```
Dappu-253741      PVNIFGKKKLAAWFVSNCV-TS-GR----REDYVKELIQY-I---------P-VDIY---
Dappu-13230       LVNISSKKKLVAWFVSHCV-TS-NR----REDYVRELSKY-I---------Q-VDIY---
Dappu-219820      LVNISSKKKLVAWFVSHCV-TS-NR----REDYVRELSKY-I---------P-VDIY---
Dappu-315514      YPWIKSKTKLVAWFVSHCD-TP-IQ----REEYARQLGQH-I---------P-VDIY---
Dappu-25363       RGRKAKKIKLVTWFVSHCS-TQ-IR----REEYARQLGQY-V---------P-VDIF---
Dappu-315506      TALIRGKSKLVAWFVAHCS-TP-IR----REEYVRQLSDF-V---------T-VDIY---
Dappu-25935       NDDDLGKTKMITWFVGHCS-TP-IR----REEYVHKLSQY-V---------P-IDVY---
Dappu-260935      TALIRGKSKMVTWFVGHCT-TP-IR----REEYVRQLSQY-V---------P-VDIY---
Dappu-19438       -----KKSKMVTWFVGHCQ-TP-VR----REEYVRQLSLY-V---------S-VDIF---
Dappu-52155       GINITSKTKTAAWFASNCK-TS-IN----REGYVRELQKY-I---------P-VDVF---
Dappu-302634      GVNITAKTKLAAWFVSNCQ-TS-IN----REGYVRELGRH-I---------P-VDVF---
Dappu-66315       ---LTLKNRTIAWFVSKCHTTH-SGGGSYREYLVQNLSSF-I---------P-VDIY---
Dappu-66309       -----LKNRTVAWFVSNCH-SD-GVGGSWREFLVRNLSQF-I---------P-VDIY---
Dappu-302400      ---LKVKNKTVAWMVSNCK-TD-SR----RESLVSHLSLL-I---------P-VDVY---
Dappu-56240       -MNLTVKNRTAAWFVSNCK-TS-SQ----CELLVRNLSNF-I---------P-VDVY---
Dappu-65379       QMNLTVKNRTAAWFVSNCK-TS-SQ----REFLVRNLSNF-I---------P-VDVY---
Dappu-4136        RMDLTRKKRTAAWFVSNCV-TD-SR----RESLVRNLSLF-I---------P-VDIY---
Dappu-266638      QIDLILKNRAVAWFVSNCE-TD-SR----RELLARNLSRF-I---------P-VDIY---
Dappu-13713       ------KDRAVAWFVSNCH-SN-SQ----WESLVRRLSEF-I---------S-VDIY---
Dappu-4141        --ISSSKNRTVAWFVSNCN-SN-SQ----RESVVRRLSQF-I---------A-VDIF---
Dappu-266923      QVDISNRNKLMVWFNSHCS-TH-SR----REDYVKKLAEF-M---------P-VDIY---
Dappu-272135      QVDISNRNKLMVWFNSHCS-TH-SR----REDYVKKLAEF-M---------P-VDIY---
Dappu-116054      YPELAKRTKLMAWFNSHCP-TH-SQ----REDYVKKLSEF-I---------P-VDIY---
Dappu-331784      YPELAKRTKLMAWFNSHCP-TH-SQ----REDYVKKLTEF-I---------P-VDIY---
Dappu-67045       HPRLAKKDKLLAWFCSNQK-TH-GR----REDYIGELGKY-M---------A-IDVY---
Dappu-316572      YPQLANKTKMVAWFASHCP-TH-SQ----REDYVQELAKF-V---------Q-VDIY---
Dappu-302457      YPHLANRTKMVAWFASHCP-TH-SQ----REDYVKELANF-V---------Q-VDIY---
Dappu-64359       YPHLANRTKMVAWFASHCP-TH-SQ----REDYVKELANF-V---------Q-VDIY---
Dappu-4083        -----KKTKMVAWFVSHCH-TD-GL----REEYFGQLGKY-V---------G-IDVY---
Dappu-53630       ----SMPTKMVAWFVSHCR-TE-SL----REKYFQWLGQH-V---------P-IDTY---
Dappu-111600      HFLVAKKTKMVAWFVSHCR-TD-SL----REKYFQLVGQH-V---------A-IDTY---
Dappu-15329       ------------YVSHRN---------------------------------------
Dappu-58299       ---LGNKTKLVAWFNSNCD-TL-GG----REKYFRQMAQY-T---------P-IDTY---
Dappu-248921      --NSGNKTKLVAWMATQCL-TD-GR----RESYVKELKRH-I---------D-IDVY---
Dappu-325563      KNFANKKNYLVVWMASHCK-TP-GL----RETYIRQLSTF-I---------R-VDIY---
Dappu-23160       -----KKTKKIAWFVSKCW-TQ-SR----RENFFRQLFEF-Y---------PSIDVY---
Dappu-221393      KNYAANKTRPIVWMVSHCR-TS-GQ----RETYVRQLSQY-I---------A-VDVY---
Dappu-3818        KNYAANKIKLVAWMVGHCD-TL-GL----REVYVRQLAKF-I---------P-IDVY---
Dappu-58354       KNYAANKIHLAVWMASHCE-TP-SL----RETYVRQLSKF-I---------P-VDVY---
Dappu-311402      KNYAAHKTQQIAWMVSHCD-TH-GL----RETYVAQLSKF-I---------P-VDIY---
Dappu-106945      RNYAANKTRLVVWMASHCA-TN-SL----RETYVKELSKY-I---------P-VDIY---
Dappu-198878      ------------------------------------------------------------
Dappu-313025      KNYAANKTSPVVWMASHCG-TH-SL----RETYVRQLGKF-I---------P-VDVY---
Dappu-313010      KNYAANKTQPVGWMVSHCD-TN-SL----RETYVRQLSKF-I---------P-VDVY---
Dappu-24623       KNYAANKTQPVGWMVSHCD-TN-SL----RETYVRQLSKF-I---------P-VDVY---
Dappu-58316       TRRKKKKKKLVAAMISHCT-TD-GQ----REQYIKQLRKH-V---------K-VDVY---
Dappu-308012      TRRKKKEKKLVAAMISHCT-TD-GQ----REQYVKQLKKH-V---------K-VDVF---
Dappu-67044       TRRKKKEKKLVAAMISHCT-TD-GQ----REQYVKQLKKH-V---------K-VDVF---
Dappu-260055      ISKPNKGSPIVA-----CD----GL----RLQHNQSAGINYI---------E-VDVY---
Dappu-63087       ----DVHPLLPSWMASDCN-TT-SQ----RELYVKELKNY-I---------E-VDVY---
Dappu-316372      ISKLNKVSPIVAWMASDCN-TT-SQ----RELYVKELKNY-I---------E-VDVY---
Dappu-68594       ---DVLLSPIVAWMASDCN-TT-SQ----RELYVKELKNY-I---------E-VDVY---
Dappu-325685      RSAFRQKTKKIAWMVSHCW-TH-SM----RELYAKELNKY-M---------D-VDIY---
Dappu-3750        KSELRNKTKKIAWMVSHCE-TH-NQ----REKYVAELQKY-V---------D-VDIY---
Dappu-336888      KPELRNKTKKIAWMVSHCE-TH-NQ----REKYVAELQKY-V---------D-VDIY---
Dappu-266928      MKSVHKKTKAVAWMVTHCN-TH-SQ----RETYIKELGKY-I---------E-VDTY---
Dappu-49339       --QKHNKTKSLAWMVSHCD-TQ-SQ----RETYAKALAKH-I---------D-VDIY---
Dappu-316587      SDAVRNKTKSVAWMVSHCK-TH-GQ----REKYVEELSKY-I---------D-VDIF---
Dappu-241186      PNSKRNKTKTVAWMVSHCN-TH-GQ----RETYVKELSKD-I---------D-VDIY---
Dappu-49176       --ALRNKTKTVAWMVSHCK-TH-GQ----REKYVKELSKY-I---------D-VDIY---
```

```
Dappu-328684       GPC-----GN----LS----CPE----TNGSPGEALQPCLDM--LADNYKFV-LAFERFI
Dappu-244685       GPC-----GT----KS----CPE----TNGTPTAAILPCLDM--LAENYKFV-LAFEHNV
Dappu-104196       GPC-----GT----KS----CPE----TNGTPTAAILPCLDM--LAENYKFV-LAFERFI
Dappu-319378       GPC-----GT----KS----CPE----TNGTPTAAILPCLDM--LAENYKFV-LAFERFI
Ixodes_ISCW004236  GLC-----GD----HK----CSR----SR------GTSCYSD--FERKYFFM-LAFENSI
Ixodes_ISCW024758  GKC-----GK----HR----CER----DT------TPRCHTL--FANNYFFL-LSFENAV
Ixodes_ISCW023318  GQC-----GH----LS----CLP----KM------SADCYHN--ASKVYFFY-LALENSI
Dappu-41601        GDC-----GS----MA----CDR----DN------AANCYEM--LEQDYKFY-LSFENSF
Dappu-48653        GNC-----GT----MT----CPR----NI------EDECREM--AAKNYKFY-MALENSL
Dappu-331779       GAC-----GT----LE----CPK----ELGVDNS-SEECRDM--AGQNYKFY-MALENSL
Dappu-67046        GSC-----GN----KK----CPKEVGVDNS-----SEDCRDM--AGQNYKFY-MALENTL
Dappu-55591        GKC-----GN----MT----CPK----KQDKSFESSDECREM--AAQRYKFY-FALENSL
Dappu-334524       GTC-----GN----LT----CPK----KLDDSYESSEECRDL--AASEHKFY-LSLENSL
Dappu-302891       GQC-----GN----MS----CSR----SN------PEFCRQM--LESDYKFY-LSLENTL
Dappu-60476        GEC-----GN----MS----CSR----SN------PELCRKM--LERDYKFY-LSLENTL
Tribolium_TC008651 GRC-----GT----LH----CEK----NN------KEGCYDM--MERKYKFY-LSFENSI
Tribolium_TC008652 GKC-----SA----LH----CEK----DN------TEACYDK--MERDYKFY-LSFENSI
Dappu-227431       GTC-----GN----LT----CSH----SD------HIECYKM--LERDYKFY-LAFENSI
Dappu-251980       GLC-----GP----LK----CNW----NSDTGIS-HPECYDM--LEKEYKFY-LSFENSL
Ixodes_ISCW003590  GKC-----GK----HV----CEP----KA------SDACYQD--AAKNYSFY-LSFENSI
Dappu-318584       GKC-----GG----KDL---CPK----LKN-----DELCYDM--IEKTYKFY-LAFENSI
Dappu-316980       DGC---EGGR----NI----CPR----EKN-----GQECYDS--IERDYKFY-LSFENSI
Dappu-312894       GKC-----GG----QDV---CPR----EKN-----SDVCYDM--IETTYKFY-FSFENSI
Dappu-236411       GQC-----GG----EDR---CPR----SQN-----EDVCYDM--IEKTYKFY-FSFENSI
Dappu-60056        GGC-----GH----KY----CGS------------HEQVRDI-----PFNFFVLAFENSL
Dappu-3751         GGC-----YS----LR----CPM----NESAFLS-TEPCYDL--LDSSYKFY-LAFENSF
Dappu-107642       GKC-----GN----LS----C--------------GDRCLEM--IRSDYKFY-VAFENSF
Dappu-253741       GKC-----GN----LS----CAD------------QTRGREM--VRDHYKFY-IGFENSL
Dappu-13230        GKC-----GN----LT----CSN------------RNHCKEM--IRRDYKFY-IAFENSL
Dappu-219820       GKC-----GN----LT----CSN------------RNHCKEM--IRRDYKFY-IAFENSL
Dappu-315514       GRC-----GK----EQVTSICDS----ADD----   NCEEIRALRAQYKFY-LAFENSW
Dappu-25363        GNC-----SS-----E----CPY--------------DCYAM--LRAEYKFY-LAFENSW
Dappu-315506       GRC-----GK-----D----CPS--------------NCDDL--LRTDYKFY-LAFENSW
Dappu-25935        GNC------T----KQ----CPS--------------HCDDM--LRTDYKFY-LAFENSW
Dappu-260935       GNC------T----QD----CPY--------------HCDEM--LRAEYKFY-LAFENSW
Dappu-19438        GSC------T----KK----CPY--------------NCDEM--LRAEYKFY-LAFENSW
Dappu-52155        GKC-----------LKNPKTCPR----KK------QKECDDM--LKREYLFY-LSFENSF
Dappu-302634       GNC-----------LENHKSCPRKKDANNQPLYYVRTECDEA--LERDYLFY-LSFENSF
Dappu-66315        GGCATKPENK----------CNT------------PRDCNLM--LSQYYRFY-LSFENSL
Dappu-66309        GGCATEEEKK----------CPN------------RPACNPM--LGQYYRFY-FCFENSL
Dappu-302400       GFC-----GNGS--HQ----CPS------------RADCDRF--LGQNYRFY-LSFENSL
Dappu-56240        GSC-----RNNGSNHT----CVN------------RADCNVM--LGRYYRFY-LSFENSL
Dappu-65379        GSCRNNGSNQ----HT----CVN------------RADCNVM--LGRYYRFY-LSFENSL
Dappu-4136         GEC----HGG----HQ----CRN------------RPECDRM--LSRHYRFY-LSFENSL
Dappu-266638       GKC-----GDGR--HS----CQN------------RVGCDRI--LSRHYRFY-LSFENSL
Dappu-13713        GKCAN---GK----HS----CPN------------KSECDQM--LSRHYRFY-LSFENSL
Dappu-4141         GKCANAA-GS----QQQHHSCPA----N-------QSECDRM--LSRHYRFY-LSFENSL
Dappu-266923       GKC-----GT----LE----CLP----RN------TPRCDSR--VLMKYKFY-LAAENSL
Dappu-272135       GKC-----GT----LE----CLP----RN------TPRCDSR--VLMKYKFY-LAAENSL
Dappu-116054       GKC-----GS----LE----CLP----YN------DPRCDTK--VLVNYRFY-LAAENSL
Dappu-331784       GKC-----GS----LE----CLP----RN------DPRCDTK--VLVNYRFY-LAAENSL
Dappu-67045        GNC-----GN----LT----CLP----RN------SDRCDNL--LDE-YKFY-LSAENSL
Dappu-316572       GKC-----GT----ME----CLP----RN------SYRCENL--LDN-YKFY-LAAENSL
Dappu-302457       GKC-----GT----ME----CLP----RN------SQRCESL--LDD-YKFY-LAAENSL
Dappu-64359        GKC-----GT----ME----CLP----RN------SQRCESL--LDD-YKFY-LAAENSL
Dappu-4083         GRC-----GK----LN----CLP----SR------SSKCDQL--LDS-YKFY-VAAENAI
Dappu-53630        GSC-----GS----LA----CVP----VR------SDKCDVE--LLDSYKFY-VAAENAL
Dappu-111600       GSC-----GS----LT----CVP----LR------SEKCDKL--LDS-YKFY-VAAENAI
```

```
Dappu-15329        ----------------------------------FPNCDQI--LDD-YKFY-VSAENSI
Dappu-58299        GAC-----GH----LK----CDP----AE------GTQCDKL--LGN-YKFY-IAAENSL
Dappu-248921       GLC-----GR----LS----CAR----HPVD-IS-HPRCYDK--LESTYKFY-LSLENSI
Dappu-325563       GRC-----GN----LS----CPR----NTTHSYS-NPTCYDL--LEAKYKFY-LSFENSI
Dappu-23160        GEC-----GIKG--LD----CQP----WK------SLDCDKI--IGD-YKFY-IAAENSF
Dappu-221393       GKC-----GN----FS----CPR----NEANWIS-DPKCYDM--LQTRYKFY-LSFENAF
Dappu-3818         GKC-----GN----LS----CSR----NTMHAYS-DPQCYQM--LEAKYKFY-LSFENSI
Dappu-58354        GGC-----GN----FS----CIR----NDSHWLS-DPKCYDM--LEAKYKFY-LSFENSI
Dappu-311402       GGC-----GN----LT----CDR----NESHWLS-DPICYTM--LEKKYKFY-LSFENCI
Dappu-106945       GGC-----GN----LS----CLH----SKTNYVS-DPKCYDQ--LETQYKFY-LSFENSI
Dappu-198878       --------------------CPR----NGMNWLS-FPKCYDE--LETKYKFY-LSFENSI
Dappu-313025       GGC-----GN----LS----CSR----NHSHWLS-FPHCYNV--LDEKYKFY-LSFENSI
Dappu-313010       GGC-----GN----LS----CSR----NDDHWLS-YPECYTM--LEEKYKFY-LSFENSI
Dappu-24623        GGC-----GN----LS----CSR----NDDHWLS-YPHCYTM--LEEKYKFY-LSFENSI
Dappu-58316        GWCNHGERSSKT--LH----CDT----DELLSS--TPECYNM--LDSNYKFY-LSFENAI
Dappu-308012       GWCDDDGLGSG---LR----CDT----HELLTS--TPECYNM--LDSNYKFY-LSFENAI
Dappu-67044        GWCDDDGLGSG---LR----CDT----HELLTS--TPECYNM--LDSNYKFY-LSFENAI
Dappu-260055       GEC-----GN----LT----CD-------------GPQCYDI--LLRSYKFY-LSFENSL
Dappu-63087        GEC-----GN----LT----CD-------------GPQCYDI--LLRSYKFY-LSFENSL
Dappu-316372       GEC-----GN----LT----CD-------------GPQCYDI--LLRNYKFY-LSFENSL
Dappu-68594        GEC-----GN----LT----CD-------------GPQCYDI--LLRNYKFY-LSFENSL
Dappu-325685       GGC-----GN----FS----CPT----HVLHSS--DPQCYDM--LQSDYKFY-LSFENSL
Dappu-3750         GKC-----GNGK--LF----CPR----HGMFHS--EPHCNKV--IESTYKFY-LSFENSF
Dappu-336888       GKC-----GNGK--LF----CPR----HGMFHS--EPHCNKV--IESTYKFY-LSFENSF
Dappu-266928       GPC-----GN----RY----CPR----HVLYSS--DPKCYEM--LESTYKFY-LSFENAI
Dappu-49339        GKC-----GN----MS----CAS----HDLHTS--APHCYTM--LESTYKFY-LSFENSL
Dappu-316587       GKC-----GH----LT----CTK----NPLHIS--DPQCYNM--IESTYKFY-LSFENAI
Dappu-241186       GGC-----GN----LS----CAL----DALHHS--DPQCYNM--IESTYKFY-LSFENAI
Dappu-49176        GRC-----GN----LS----CAK----HSLYHS--DPKCYDM--IESTYKFY-LSFENAI
                                                                 . *  .  *.

Dappu-328684       CDDFVTKRFFDLLS-RDTVPIVFG-GADYTRIAPPHSFIDALSFN-PRQLADRLLE----
Dappu-244685       ---------FQPEV-KDVV----G-QRLAAKFVHQFQIVKAVGHGEG-------------
Dappu-104196       CDDFVTKRFFDLLS-RDTVPIVFG-GADYKRIAPPYSFIDALSFN-PKELADHLLK----
Dappu-319378       CDDFVTKRFFDLLS-RDTVPIVFG-GADYKRIAPPYSFIDALSFN-PKELADHLLK----
Ixodes_ISCW004236  CRDYITEKFFTALR-YDMVPVVFG-GANYTRVAPSRSFIDALSFKSPKHLAEHLTR----
Ixodes_ISCW024758  CKDYVTEKLYYTLL-YDIIPVVFG-GANYSAVAPAGSYIDALSFESPKHLAVHLTS----
Ixodes_ISCW023318  CTDYITEKFYNALT-WGMVPIVMS-GANYTSVAPPRSYIDALSFQNVRHLADHLKQ----
Dappu-41601        CDDYVTEKFFSVLR-LDVVPIVFG-GGNYSAISPPFSYINAQDFETAVQLADYLKM----
Dappu-48653        CQDYVTEKFFAMLH-QPIIPIVYGVHDHYDQIAPTHSFINAAKFETMKQLADYLIL----
Dappu-331779       CRDYITEKFFGMLQ-RPVIPVVFGLHNHYDQMAPPHSFINAAKFENMRQLADYLIL----
Dappu-67046        CRDYITEKFFGMLH-RPIIPVVFGLHDHYDQMAPPHSFINAAKFENMRQLADYLIL----
Dappu-55591        CRDYVTEKFFENIR-RPILPIVFGLHGDHEKLAPPHSFINAANFKNMKALANHMNL----
Dappu-334524       CRDYVTEKLFAMMH-RPIIPVVYGLHDDQEKLAPPHSFINAAKFENTKALADYLIL----
Dappu-302891       CEDYVTEKFFDQMR-YHIIPIVFDLHGHHARMAPSHSYINAADYQSVRELADYLTL----
Dappu-60476        CEDYVTEKFFDQMR-YHIIPIVFDLHGHHARMAPPHSYINAADYQSVRELADYLTL----
Tribolium_TC008651 CEDYVTEKLYNVLQ-RNIVPIVYG-GADYNTLAPPKSVINVMDFMSVKHLVKHLKY----
Tribolium_TC008652 CEDYVTEKLYNVLQ-RNIVPIVYG-GADYNTLAPPKSVINVMDFMSVKDLVKHIKY----
Dappu-227431       CKDYVTEKFFNALL-FNVVPVVYG-GANYHALAPKNSYIDVRDFSSVHHLVKYLKF----
Dappu-251980       CSHYVTEKFYSILK-LDVVPVVMG-RANYSGIAPPYSFIDALRYS-PKQLADYLLL----
Ixodes_ISCW003590  CRDYVTEKFFRPLL-FDLVPVVLG-GGDYVSVAPPGSYINALDFRSPAELGEYLKR----
Dappu-318584       CREYVTEKFFNSIA-RNLVPIVLG-GANYSAIAPEHSYIDALAYS-PRQLAAYMKR----
Dappu-316980       CDDYVTEKFFEMMS-RNVVPVVLG-GANYTALAPPHSFINALDFT-PRELANYLKQ----
Dappu-312894       CEEYVTEKFFEMMG-RNIVPVVLG-GADYSAIAPPHSYISALDYT-PKQLAKYLKE----
Dappu-236411       CEEYVTEKFFEMMG-RNIVPVVLG-GADYSAIAPPHSYISALDYT-PKQLAKYLKE----
Dappu-60056        CTDYVSEKLYTALE-NGVVPVVYG-EADYRAYAPSYSYVNARDFGSPKELAEYLWL----
Dappu-3751         CNDYVTEKFFDVLQ-RRIIPIVMG-GANYSAIAPPHSYIDALQYS-PRELAEYLKL----
Dappu-107642       CTDYVTEKLTRALL-YDAVPIVMG-GVDYNRFAPPHSFIDVKDFDSPEQLGNYLLL----
Dappu-253741       CTDYVTEKLMVGLL-YDAVPIVKG-GVDYTEFAPPHSFIDVNDFTSPKQLADYLLL----
Dappu-13230        CTDYVTEKLAIGLI-YDAVPIVMG-SVDYTKFAPPHSFIDVNDFPSPKQLARYLLL----
```

```
Dappu-219820       CTDYVTEKLAIGLI-YDAVPIVMG-SVDYTKFAPPHSFIDVNDFPSPKQLARYLLL----
Dappu-315514       CPDYVTEKFYRTLQ-FDTVPIVLG-GAEYDRFAPPHSFINALDFSSPKQLAEYLLL----
Dappu-25363        CPDYVTEKFTRPLF-HDAVPIVLG-GADYSHFGPPHSYINARDFASPKALADYLIL----
Dappu-315506       CPDYITEKFIRPLV-YDSVPIVLG-GANYSHFAPPHSYINARDFDSPKELADYLIL----
Dappu-25935        CPDYVTEKFIRPYL-YEAIPIFLG-GADYSKYAPRNSYINARDFDSPKQLAEYLIL----
Dappu-260935       CPDYVTEKFIRPFV-YDAIPIFLG-GADYSQFAPPHSYINARDFKSPKELAHYLIL----
Dappu-19438        CPDYVTEKFIRPFL-YDAVPIVLG-GADYNQFAPSNSYINAMDFGSPK------------
Dappu-52155        CPDYVTEKFYRAFE-TGTVPVVFG-GANYSLFAPPHSYINARDFKTPKLLAEYLIQ----
Dappu-302634       CPDYVTEKFYRAVE-MGTVPVVFG-GANYSLFAPPHSFINARDFQTPKLLAEYLVK----
Dappu-66315        CPDYVTEKLYRPMA-YDTVPVVYG-GSDYSFYLPAGSYINAMDFDSPQSLANYLKK----
Dappu-66309        CPDYVTEKCYRPLA-YDTVPVVYG-GSDYSLFFPAGSYINALDFDSPESLANYLKK----
Dappu-302400       CPDYITEKLYRPLA-HGVVPVVYG-GSDYSFYLPAGSYVNARDFDSPQSLAEYLEK----
Dappu-56240        CPDYVTEKLYRTLM-HDTVPVVYG-GANYSLYLPEGSYVNARDFDSPENLANHLKE----
Dappu-65379        CPDYVTEKLYRTLM-HDTVPVVYG-GANYSLYLPEGSYVNARDFNSPENLVNHLKE----
Dappu-4136         CPDYVTEKLYRALA-HDTVPVVYG-GADYSLYLPAGSYVDARDFESPQSLADHLKK----
Dappu-266638       CPDYVTEKLYRPMA-YDTVPVVYG-GSDYSFYLPAGSYINAMDYDSPQSLANHLKK----
Dappu-13713        CPDYVTEKLYWPLA-HDTVPVVYG-GADYSDFFPARSYVDGRHFENPEALADHLKK----
Dappu-4141         CPDYITEKLYRPLA-HDTVPVVYG-GADYSLYLPVGSYVNARDFKNPEALANHLKK----
Dappu-266923       CPDYVTEKFYRGFL-NDIVPVVYG-GADYSQYAPPHSYINIADFRSPKELADYLLL----
Dappu-272135       CPDYVTEKFYRGFL-NDIVPVVYG-GADYSQYAPPHSYINIADFRSPKELADYLLL----
Dappu-116054       CPDYVTEKFYRALM-NDIVPVVFG-GADYAQYAPPNSYVNIADFQSPKQLAEYLLL----
Dappu-331784       CPDYVTEKFYRALM-NDIVPVVFG-GADYAQYAPPNSYVNIADFQSPKELAEYLLL----
Dappu-67045        CADYVSEKFYRALK-TDIIPVVYG-GADYAAYAPPHSYIHVADFASPKQLAEYLLL----
Dappu-316572       CPDYVSEKFYRALN-QNIVPIVYG-GADYAEYAPPHSFINIADFKSPQDLAAYLKL----
Dappu-302457       CPDYVSEKFYRALT-NDIVPIVYG-GADYTDYAPPHSFINLADFASPKDLAAYLKL----
Dappu-64359        CPDYVSEKFYRALT-NDIVPIVYG-GADYTDYAPPHSFINLADFASPKDLAAYLKL----
Dappu-4083         CTDYVTEKFYRALS-SDIVPIVYG-GADYSSYAPPLSYIDVSDFKSPKDLADYLKL----
Dappu-53630        CPDYVTEKFYRALA-ADIVPIVYG-GADYSAYAPPSSYIDAGDFKSPKALADYLKL----
Dappu-111600       CPDYVTEKFYRAMA-ADIVPIVYG-GADYSEYAPPMSYIDAGDFKSPKALADYLKL----
Dappu-15329        CPDYITEKFYRALE-MGVVPVVYG-GADYSAYAPPHSYINAADFESPQALADYLLL----
Dappu-58299        CADYVTEKFYRALE-ADVVPIVYG-GADYSAYAPAHSYINTADFASPKALAEYLYV----
Dappu-248921       CRDYVTEKFFKIIQ-RRIVPVVYG-GADYERIAPAGSYIDARRYH-PAQLADYLRR----
Dappu-325563       CEDYVTEKFFEIMK-RDLIPIVYG-GAKYINIAPHHSYIDATQYT-PEGLARYLKLGRHY
Dappu-23160        CPDYVTEKFYRALQ-VGAVPIVYG-GSDYSAYAPPYSFIHAADFQSPKDLADYLIL----
Dappu-221393       CTDYVTEKFFEIMD-HDMIPIVYG-AANYSEIAPPHSYINALDFT-PEGLARYLQM----
Dappu-3818         CEEYVTEKFFEIAN-RDIVPIVYG-GADYKRIAPPHSFIDALEFT-PEALAQYLTI----
Dappu-58354        CEDYVTEKFFEIMN-HDIIPVVYG-GANYSRIAPPHSYIDALQFT-PETLAQYLKV----
Dappu-311402       CTDYVTEKFFELLN-YDIIPIVYG-GANYSQLAPLHSFINALDFT-PETLAQYLKI----
Dappu-106945       CNDYVTEKFFEIIN-HNIVPIVYG-GANYSQFAPHHSYINALDFT-PEKLAQYLLL----
Dappu-198878       CNDYVTEKFFEIMN-HNIVPIVYG-GANYSQFAPHHSYINALDFT-PEKLAQYLLL----
Dappu-313025       CTDYATEKFFEILT-HNMVPVVYG-GANYSYIAPPHSYINALDFT-PEKLAEYLKL----
Dappu-313010       CTDYATEKFFEILK-HNMIPVVYG-GANYSQIAPPHSYINALDFT-PEKLAEYLKL----
Dappu-24623        CTDYATEKFFEILK-HNIVPVVYG-GANYTQIAPPHSYIDALDFT-PEKLAEYLKL----
Dappu-58316        CPDYVTEKFFQIMSLRDIVPVVYG-GADYAQLAPEHSYIDARQFE-PQQLAAYLKK----
Dappu-308012       CQDYVTEKFFHIMSLRDIVPVVYG-GADYAQLAPGHSYIDALQFE-PKQLAAYLEM----
Dappu-67044        CQDYVTEKFFHIMSLRDIVPVVYG-GADYAQLAPGHSYIDALQFE-PKQLAAYLEM----
Dappu-260055       CPDYVTETFFTMMD-RDVVPVVYG-GADYTRYAPTHSYIDARQIK-PEELATYLKL----
Dappu-63087        CPDYVTETFFTMMD-RDVVPVVYG-GADYTRYAPTHSYIDARQIK-PEELATYLKL----
Dappu-316372       CPDYVTDTFFTMMD-RDVVPVVYG-GADYTRYAPTHSYIDARQFK-PEELATYLKI----
Dappu-68594        CPDYVTDTFFTMMD-RDVVPVVYG-GADYTQFAPIHSYIDARQFK-PEELATYLKF----
Dappu-325685       CKDYVTEKFFKVMD-HDIVPIVYG-AADYARHAPPHSYIHAGKFK-PKELADYLKL----
Dappu-3750         CKDYVTEKFFKILD-LYMIPIVYG-GADYTQHAPPHSYIDARKFK-PKELAAYLKI----
Dappu-336888       CKDYVTEKFFKILD-LYMIPIVYG-GADYTQHAPPHSYIDARKFK-PKELAAYLKI----
Dappu-266928       CPDYVTEKFFKILG-QNLVPIVYG-GADYTQHAPAHSYIDALKYK-PKELAAYLQL----
Dappu-49339        CPDYVTEKFFKIMG-HDIVPIVYG-GADYSRHAPPHSYIDARHFK-PKELAAYLKQ----
Dappu-316587       FPDYVTEKFFKIMG-HHIVPVVYG-GADYTQHAPPHSYIDARKFK-PEELAAYLKL----
Dappu-241186       CPDYVTEKFFKIMG-HHIVPIVYG-GADYTQHAPPHSYIDARKFK-PKELATYLKL----
Dappu-49176        CPDYVTEKFFKIMG-HHIVPVVYG-GADYSQYAPPHSYINAREFK-PKELAAYLKL----
                                           :    .           . :
```

```
Dappu-328684       --------LDKSDRQYYRHFWWKDFYQVTPLLSTR----------------PF-CDLCEK
Dappu-244685       --------LVGGDS--------RPVHQVGTVKGQQAKK-------------GHAVTEVAQ
Dappu-104196       --------LEKDEKHYFRHFWWKDVYKVIYTK-------------------------
Dappu-319378       --------LEKDEKHYFRHFWWKDVYKVIYSR------------------PF-CDLCEK
Ixodes_ISCW004236  --------VAKNVSAYKSYFDWKSRYKILSWSE-----------------EF-CTLCSK
Ixodes_ISCW024758  --------VAKDFNLYKSYFNWKGKYDLIPWTEI----------------TF-CNLCSK
Ixodes_ISCW023318  --------IAKDPLLYNSYHAWRQRYKIVWRPFQ-------------------CNLCQM
Dappu-41601        --------LDSNDDLYNQYFWWKPHYRVRNHIQDLKL-------------SM-CGLCSR
Dappu-48653        --------LDKNDTLYNEYFWWKPHFEVRYSQKDKNK-------------SM-CHLCAA
Dappu-331779       --------LDRNDTLYNEYFWWKPHFESRYKQKDVNI-------------GM-CHLCAS
Dappu-67046        --------LDRNDTLYNEYFWWKPHFESRYKQKDVNI-------------GM-CHLCAS
Dappu-55591        --------LDMNDTLYNEYFWWKPYFQVRDSQKDRNQ-------------AM-CHLCAR
Dappu-334524       --------LNNNDTLYNEYFWWKPYFKVHDSEDEKNK-------------SM-CRLCAA
Dappu-302891       --------LDGNDTLYNEYFWWKKHYASVGPGDTRE--------------GM-CRLCDL
Dappu-60476        --------LDGNDTLYNEYFWWKKHYVVNNNDDGIKR-------------SM-CELCRM
Tribolium_TC008651 --------LDSHPEEYLKFLEWKKDYIVETASTQ----------------TL-CTLCQK
Tribolium_TC008652 --------LDSHPEKYLKFLEWKKDYIVETSSTR----------------SL-CTLCQK
Dappu-227431       --------LARNDSAYLHYFDWRKTPPGLSLLPRTNQ-------------GW-CTLCSM
Dappu-251980       --------LDGNQTLYERYLKWKTSYIIRSGYEEMGGQ-------------AL-CSLCAQ
Ixodes_ISCW003590  --------VAGDPEWYESFFLWKNHFKLKYEHLG-------------------CKLCSK
Dappu-318584       --------VDQNDSLYAEFFWWKPHYRVVNLPQTNKE-------------SF-CNLCAA
Dappu-316980       --------LDADDRLYAEYFWWKPHYQVANLYHTNRQ-------------VF-CHLCQA
Dappu-312894       --------LDSNDTLYAEYFWWKPHYRIRNLYDTNRK-------------AF-CDLCEA
Dappu-236411       --------LDSNDTLYAEYFWWKPHFTVRNLYGTSRQ-------------TF-CDLFEA
Dappu-60056        --------LHQNDHLYQNYFSWNQDYMVDRFPTD----------------GW-CNLCQM
Dappu-3751         --------LASDDKLYNEYFWWKPHFQVVKRYPFLAAN-------------AL-CSLCDK
Dappu-107642       --------LNASDSLYGRYFEWKRHFNVQLTTKQ----------------GW-CHLCKL
Dappu-253741       --------LDKADSLYARYFDWRRDFTVELYQKR----------------GW-CRLCQL
Dappu-13230        --------LSETDALYMRYFDWKRDFTVHLNLKL----------------SW-CRLCQL
Dappu-219820       --------LSETDALYMRYFDWKRDFTVHLNLKL----------------SW-CRLCQL
Dappu-315514       --------LNSSEELYVGYFQWKNHYRVSLPAMD----------------GW-CDLCRM
Dappu-25363        --------LNNSDALYASYFDWKKDFRVVKTDMS----------------GW-CDLCQL
Dappu-315506       --------LDKSDDLYARYFDWKRDHDITLLDLS----------------GW-CDLCEM
Dappu-25935        --------LDKSESLYASYFSWKNHYYVSVPDMY----------------GW-CELCRM
Dappu-260935       --------LDKSDDLYARYFDWKRDYYVSVPDFY----------------GW-CELCRM
Dappu-19438        ------------------------------------------------------------
Dappu-52155        --------LSRNLDLYSHFFDWKGKFNLRKSS------------------GWACKLCEM
Dappu-302634       --------LS------------------------------------------------
Dappu-66315        --------LMADDELYLSYFRWREHYAVDTAPKD----------------AF-CQLCQM
Dappu-66309        --------LMTDDELYLSYFRWRRKYVVDLAPKD----------------SW-CQLCEM
Dappu-302400       --------LMLDDELYLSYFRWRNRFTVDPKPVD----------------GM-CQLCRL
Dappu-56240        --------LMINDELYLSYFRWKQRYTVELGHLN----------------GW-CSLCRL
Dappu-65379        --------LMINDELYLSYFRWKQRNTVELGHLN----------------GW-CSLCRL
Dappu-4136         --------LMSDDRLYLKHFTWRRNYVVD----------------------------
Dappu-266638       --------LMADDELYLSYFRWRQKFAVDPSPID----------------GM-CRLCQL
Dappu-13713        --------LIANDTLYSSYFEWKSQYVV----------------------------
Dappu-4141         --------LMANDTLYASYFQWRIKYVVD----------------------------
Dappu-266923       --------LDKNDALYRKYFDWKKNFEVINRPLN----------------GW-CDLCEK
Dappu-272135       --------LDKNDALYRKYFDWKKNFEVINRPLN----------------G--------
Dappu-116054       --------LANNDALYSKYFDWKKDYEVIRKPLN----------------GW-CDLCAK
Dappu-331784       --------LANNDALYSKYFDWKKDYEVINRPPD----------------GW-CDLCAK
Dappu-67045        --------LDKNEALYLKYFEWKKDYDVVRSPLD----------------GW-CDLCEK
Dappu-316572       --------LDSHDALYLEYFRWKKHYEVVRKPKK----------------GW-CDLCAK
Dappu-302457       --------LASNEALYVEYFQWKKHYAVVRSPKK----------------GW-CDLCAK
Dappu-64359        --------LASNEALYVEYFQWKKHYAVVRSPKK----------------GW-CDLCAK
Dappu-4083         --------LDENDGLYLKYFDWKKDYEVVSRPVT----------------GW-CELCEK
Dappu-53630        --------LDQNDGLYLKYFDWKKDYQVVNGPVG----------------GW-CQLCEK
Dappu-111600       --------LDENDGLYLKYFDWKKDYEVVRRPVG----------------GW-CELCEK
Dappu-15329        --------IERNPRLYSEYLDWNKDWEINKQKQSE----------------GW-CRLCEK
Dappu-58299        --------LDTNPGLYSKYFDWKKDWEVIRSPTD----------------GW-CDLCEK
```

```
Dappu-248921       --------LDADDTLYQEFFRWKKDYAVEAGVASMARR------------GF-CHLCSR
Dappu-325563       HCPRITAEISLASSILPLLAYHSSPHILESSIPSLLAYHSSLVLSSRTQLRNF-NSFSQQ
Dappu-23160        --------LDQNPKLYARYFEWKKDWIVDREPFD----------------GW-CSLCEK
Dappu-221393       --------LDANDTLYNEYFWWKNHYRVESGEPQMARH------------GF-CDLCKK
Dappu-3818         --------LDANDELYNEFFWWKSHYKVEAGLQQMARH------------GF-CDLCKK
Dappu-58354        --------LDANDQLYNEYFWWKGHYAVESGVEQMARH------------GF-CDLCKK
Dappu-311402       --------LDANDTLYNEYFWWKDHYRIESGIEQMARH------------GF-CDLCKK
Dappu-106945       --------LDANDNFYNEYFWWKDHYRVESGVEQMARH------------AF-CDLCKK
Dappu-198878       --------LDANDNFYNEYFWWKDHYRVETGVEQMARH------------GF-CDLCKK
Dappu-313025       --------VDSNDTLYNEYFWWKDHYEVEAGVDQMASH------------GF-CDLCKK
Dappu-313010       --------VDSNDTLYNEYFWWKDHYEVEAGVDQMASH------------GF-CDLCKK
Dappu-24623        --------VDSNDTLYNEYFWWKDHYEVEAGVDQMASH------------GF-CDLCKK
Dappu-58316        --------LAANETLYNEYLWWKDDYVVEAGMEEMVRR------------GF-CDLCRK
Dappu-308012       --------LAANETLYNEYLWWKDDYTVEAGLEQMVRH------------GF-CDLCRK
Dappu-67044        --------LAANETLYNEYLWWKDDY---------------------------------
Dappu-260055       --------LDANDTLYGEYFWWKDHYRNIT------------------------------
Dappu-63087        --------LDANDTLYGEYFWWKDHYRVTSSKENMWRN------------SF-CDLCQK
Dappu-316372       --------LDANDTLYGEYFWWKDHYRVTSSEENMWHN------------SF-CDLSQK
Dappu-68594        --------LDANDTLYGEYFWWKDHYQVTSSEENMWRN------------SF-CDLC--
Dappu-325685       --------LDANQTLYEEYFWWKDHFRVESSVDDMSRH------------GF-CDLCQK
Dappu-3750         --------LDADDALYNEYFWWKDHYHVEFITENTSRH------------GF-CSLCQK
Dappu-336888       --------LDADDALYNEYFWWKDHYHVEFITENTSRH------------GF-CSLCQK
Dappu-266928       --------LITNETLYNEYFWWKDYYKVEFTLEDRSRH------------AF-CDLCQK
Dappu-49339        --------LDADDALYNEYFWWKDHYRVEYSVDDRSRH------------AF-CDLCQM
Dappu-316587       --------LDANDTLYNEYFWWKDYYDVEYSIEGTTRH------------GF-CDMCQK
Dappu-241186       --------LDANDTLYNEYFWWKDYYRVEYSVEDMTRH------------GF-CDLCQK
Dappu-49176        --------LDANDTLYNEYFWWKDYYDVEYSIEDMSRH------------GF-CDLCRK


Dappu-328684       LN-------SNLPR-----KVYRDIDAW----WYNS---TKCSGPEDRGIVIRNKGNEDL
Dappu-244685       LG-----------------EIRVDIR----------------------------------
Dappu-104196       ------------------------------------------------------------
Dappu-319378       LN-------SNLPR-----KIYRNLDDW----WYNN---TKCSAPEDRGIVIRHNGTVDD
Ixodes_ISCW004236  LH-----G-KDFREQ----TTYNDMRVW----W-EQ-E-GRCRSWNL-------------
Ixodes_ISCW024758  LY-----S-EHFRRS----TVYEDILYW----W-NA-T-SQCRVWDRYSNQLLQ------
Ixodes_ISCW023318  LH-----E-QSPP------RMYEDINAW----WFM------------------------
Dappu-41601        LH-----Y-DRAL------KIYDDMEKW----W-VQ-D-SHCHTPRSDNVFHIPFWKN--
Dappu-48653        LH-----N-TTLPP-----KIYRDMTEW----W-ET-K-SKCADSPHIS-----------
Dappu-331779       LH-----N-NELPA-----KSYSNMTDW----W-EK-Q-SYCVTSPPIS-----------
Dappu-67046        LH-----N-KDMPS-----KTYTNMTDW----W-DE-R-SACINSPPIS-----------
Dappu-55591        LH-----D-EKLPR-----KIYSNLTDW----W-EK-K-STCIYSPTIS-----------
Dappu-334524       LH-----D-ETLPP-----KIYHNLTDW----W-DT-Q-STCIFSPKIS-----------
Dappu-302891       LH-----D-STIPS-----KTYRNMTDW----W-DV-Q-SKCRSLTFVDKNTSKNDSNFY
Dappu-60476        LH-----V-PNKPS-----KIYSDMTNW----W-DI-Q-ATCQTITFSEETDFAEESDGE
Tribolium_TC008651 LN-----E-PIKQ------KIYNDITKW----WAGK-DLDKCMVSKNGFLDKYLLQS---
Tribolium_TC008652 LH-----E-PIKQ------KIYSDITKW----WAGK-NKNKCMVNKNGFLDKYLLQS---
Dappu-227431       LN-----N-DSLPS-----HSYSNIHSW----W-FE-K-GQCEKDRTSIQKLAI------
Dappu-251980       LN-----L-SDDRKEI---MPAADVLST----W-NP-T-TRCLNPRYVKAFHSIDRNNNR
Ixodes_ISCW003590  LH-----S-DIASGRT---FTYNKFRK-----WFLE-D-ARCANWKQLLHGRA-------
Dappu-318584       LH-----E-TPMQE-----RKAQGLQK-----WYVD-D-SHCLVKPNFNSTQ--------
Dappu-316980       LH-----Y-QVGQPLLANGSTLQDVKK-----WYMD-D-SHCLDIPKFDET---------
Dappu-312894       LH-----T-SPIQS-----SVAKGLHQ-----WYHK-D-AKCRHNPKFDET---------
Dappu-236411       LH-----STPLKR------GTVNGLEK-----WYMK-E-SHCANMPIIIRN---------
Dappu-60056        LH-----R-PIES------QAYSDVQRW----W-AE-E-VTCTSNYHFNLTVSNNLEPVA
Dappu-3751         LH-----H-NKTE------SIYHDLAAG----W------------------------
Dappu-107642       AN-----D-DRLPS-----RTYDDIFQW----WVDD-P-ETCNLKVGDTPIQRTS-----
Dappu-253741       AN-----D-DRLPP-----RVYDDILKW----WVDD-P-VNCNLPS-------------
Dappu-13230        AH-----DSQVISS-----TTYKDILEW----WVSN-----------------------
Dappu-219820       AH-----DSQVISS-----TTYKDILEW----WVSN-TPANCSNLPPHTKFPFPEFAIKF
Dappu-315514       AH-----D-ETLPA-----KVYPDIKQW----W-ID-D-GPCENDRTNYF----------
```

```
Dappu-25363       AH-----N-DSLPS-----KVYPDIKRW-------------------------------
Dappu-315506      AH-----D-DTLPA-----KTYRDIKQW----WMLD-D-GECETDSNKYF----------
Dappu-25935       IH-----D-SKLPP-----KVYPDIKKW----WMS------------------------
Dappu-260935      AH-----D-NTLPI-----KVYHDIKQW----WMLD-A-GECESNSTKYF----------
Dappu-19438       ------------------------------------------------------------
Dappu-52155       LH-----T-DLRVTAA---KSYEDIGE-----WFFD-K-NTCENYQWSNVRS--------
Dappu-302634      ------------------------------------------------------------
Dappu-66315       LR-----N-PDVKA-----KTYANMSAW----WLGETINHTCMYAPPKSLVFNQTG----
Dappu-66309       LR-----D-PKP-------KMYDDIGAW----WSGE-T-INQTCLMTPPKSLVNVT----
Dappu-302400      LS-----D-KKTEE-----KIYPDIAEW----WHGG-N-HTCLTPPPSLV----------
Dappu-56240       LN-----DRNDAEK-----KSYAVIAAW----WSGQLNNQTCFTPPPTSLV---------
Dappu-65379       LN-----DRNDAEK-----KSYADIAAW----WSGQLTNQTCFTPPPTSLV---------
Dappu-4136        ------------------------------------------------------------
Dappu-266638      LS-----D-TQTEA-----KSYPDISSW----LAGNVANQTCFPPPTTK-----------
Dappu-13713       ------------------------------------------------------------
Dappu-4141        ------------------------------------------------------------
Dappu-266923      LN-----D-PTQKS-----KSYENVAKW----W-YD-D-IPCLAGSSFINSIATM-----
Dappu-272135      ------------------------------------------------------------
Dappu-116054      LN-----D-PTLAS-----QSYASVAKW----W-YD-D-SPCLPGSSYITSLIRSS----
Dappu-331784      LN-----D-ESLPR-----KSYSNMGMWCIRKWYLH-E-EKERRESLADALGIFD-----
Dappu-67045       LN-----D-PHEPT-----KIYESMAEW----W-YD-D-VPCYPGESFIKTRLNHIQ---
Dappu-316572      LN-----D-PQLETVT---KSYADVGHW----W-IR-K-LPCYPGSSFLMSHT-------
Dappu-302457      LN-----D-PHWQSQR---KSYEDVAEW----W-VR-K-LPCYPGSSFLLGHM-------
Dappu-64359       LN-----D-PHWQSQR---KSYEDVAEW----W-VR-K-LPCYPGSSFLSGHTSIPAS--
Dappu-4083        LN-----D-PHQRP-----KVYKDISDW----W--------------------------
Dappu-53630       LN-----D-AQQKP-----KVYADMTDW----WFHT-N-IPCLSGYDYLDHLLQQDAKDN
Dappu-111600      LN-----D-PQQKP-----KVYKDMTDW----WYHK-D-IACLSGYDYLDNLLQQNMTTF
Dappu-15329       LNAAAKSK-EQQPKNNSASKVYRDMAKW----YYE------------------------
Dappu-58299       LN-----R-PEEPE-----KSYEDIGT-----WFYD-K-VPCLPGSSLKNLYGEM-----
Dappu-248921      LH-----H-DQTV------KTYVDLTSH----WQHP-S-DECQSPLEMNEFIFSLY----
Dappu-325563      LS-----S-YISSIIV---AIFVSSGPG----F-LV-V-GPHKVIQSGILWATIVS----
Dappu-23160       LN-----D-PDANQTS---KSYRDIAKW----W--------------------------
Dappu-221393      LH-----Q-DESV------KYYPEIRSE----W-HP-N-SQCRHLSSTWENSPQNYLTPV
Dappu-3818        LH------------------------------------------------------------
Dappu-58354       LH-----QEDEGVV-----KFYPQLVSE----W-DP-K-KKCKYFDSWETQS--------
Dappu-311402      LH-----Q-DDGVT-----KYYPELLTD----W-NP-D-TVCEKVESWDIPTYPVTHRFF
Dappu-106945      LH-----H-EEGVT-----KFYPELESE----W-HP-K-TQCRYFSSWETSA--------
Dappu-198878      LH-----Q-EEGVT-----KFYSDLVSE----W-HT-K-TQCKQMSNWETSTTTQSTTTT
Dappu-313025      LH-----Q-DQGVI-----KYYSELVSE----W-HY-N-TQCHQFTSWETQS--------
Dappu-313010      LH-----Q-DQGVI-----KYYHELVSE----W-DP-E-TKCKQMSSWEKN---------
Dappu-24623       LH-----Q-------------------------------------------------
Dappu-58316       LH-----EANQEP------KMYTSMASR----WNP----ARCQRPSKHGDQIKPEQNLPG
Dappu-308012      LH-----V-DNERI-----KSHPSLFPK----W-HP---GRCSRPTYKLKKSPKKFPFLK
Dappu-67044       ------------------------------------------------------------
Dappu-260055      ------------------------------------------------------------
Dappu-63087       LH-----R-DFES------KSYQDLISY----W-DD-N-NQCVPFDPKWIF---------
Dappu-316372      LH-----R-DFES------KSYQDLISY----W-GD-Y-NQCVPFDPKWIF---------
Dappu-68594       ------------------------------------------------------------
Dappu-325685      LH-----E-DSEF------KSYAEMASD----WGDD-S-RQCA----------------
Dappu-3750        LH-----D-------------------------------------------------
Dappu-336888      LH-----D-IQTPF-----QSYADEGVLT---DLGD-D-SKCLPFDPNWIS---------
Dappu-266928      LH-----E-QDDR------KSYPDLSAE----W-GD-G-NKCKPFDPTWI----------
Dappu-49339       LH-----E-SDDDGHF---QTYPDMESF----W--G-N-ETCQPFDPKWIS---------
Dappu-316587      LH-----E-LKDVDY----QSYKRSG------F--------------------------
Dappu-241186      LH-----Q-QQDGDF----RTYKELESE----W-GD-G-NKCQPFDPSWLS---------
Dappu-49176       LH-----Q-QQDGNF----QSNKELESE----W-GD-G-NKCQQFDSSWL----------
```

I. Expanded and Unknown Genes are Ecoresponsive Genes

Table S49. Counts of unique gene transcripts sampled from cDNA libraries partitioned into three ecological conditions. Biotic challenge includes *Daphnia pulex* exposed to bacterial infection, predators, juvenile hormone and varying diets. The abiotic challenge includes animals exposed to environmental toxicants, elevated UV, hypoxia, acid, salinity and calcium starvation. Standard non-ecological conditions include animals at various stages of life-history within a controlled laboratory environment. The transcribed gene counts with homology to proteins from other species, without homology to other proteomes are tabulated here, with Chi-square statistical analysis of the effects. The transcribed gene counts for loci found within tandem duplicated gene (TDG) clusters and outside of TDG clusters are tabulated below, with Chi-square statistical analysis of the effects.

Homology vs no homology	Biotic challenge			Abiotic challenge			Standard conditions		
	Homology	No homology	Total	Homology	No homology	Total	Homology	No homology	Total
Count	1,184	1,393	2,577	2,895	3,700	6,595	3,599	2,632	6,231
Expected Values	1,284.6	1,292.4		3,287.4	3,307.6		3,106	3,125.0	
Chi-square contribution	7.873	7.826		46.847	46.562		78.254	77.778	
Row Percent	45.95%	54.06%	16.73%	43.90%	56.10%	42.82%	57.76%	42.24%	40.45%

Chi-square statistics for all table factors = 265.1399; d.f. = 2; p = $2.664438e^{-58}$

Within vs outside of TDG clusters	Biotic challenge			Abiotic challenge			Standard conditions		
	In TDG cluster	Not in TDG cluster	Total	In TDG cluster	Not in TDG cluster	Total	In TDG cluster	Not in TDG cluster	Total
Count	936	1,641	2,577	2,462	4,133	6,595	1,999	4,232	6,231
Expected Values	902.9	1,674.1		2310.8	4284.2		2183.3	4047.7	
Chi-square contribution	1.21	0.653		9.894	5.336		15.55	8.388	
Row Percent	36.32%	63.68%	16.73%	37.33%	62.67%	42.82%	32.08%	67.92%	40.45%

Chi-square statistics for all table factors = 41.03073; d.f. = 2; p = $1.231094e^{-09}$

Table S50. Differential expression (DE) of the genome of *Daphnia pulex* with four treatments measured on genome tiling path microarrays. Counts of tiles with DE per genome feature (gene, intron, unknown). Tiling DE is ascertained from statistical analysis of balanced treatment × three-replicate design using the LIMMA package in R [S16, S37, S38]. Counts of the tiles with up-regulation, down-regulation and no difference in each genome feature are tabulated here, with Chi-square statistical analysis of the effects.

Cadmium exposure	Up-regulated				Down-regulated			
	Gene	Intron	Unknown	Total	Gene	Intron	Unknown	Total
Count	9,539	2,118	26,226	37,883	16,461	2,493	31,242	50,196
Expected values	9,659	2,189	26,035		12,798	2,901	34,497	
Chi-square contribution	1	2	1		1,048	57	307	
Row Percent	25%	6%	69%	1%	33%	5%	62%	2%

Cadmium exposure	No differential regulation			
	Gene	Intron	Unknown	Total
Count	717,889	164,017	1,947,692	2,829,598
Expected values	721,432	163,537	1,944,628	
Chi-square contribution	17	1	5	
Row Percent	25%	6%	69%	97%

Chi-square statistics for all table factors = 1441.834; d.f. = 4; p = $5.863123e^{-311}$

Kairomone exposure	Up-regulated				Down-regulated			
	Gene	Intron	Unknown	Total	Gene	Intron	Unknown	Total
Count	48,569	10,405	12,7001	18,5975	39,292	8,238	118,583	166,113
Expected values	47,416	10,748	12,7810		42,352	9,601	114,160	
Chi-square contribution	28	11	5		221	193	171	
Row Percent	26%	6%	68%	6%	24%	5%	71%	6%

Kairomone exposure	No differential regulation			
	Gene	Intron	Unknown	Total
Count	656,028	149,985	1,759,576	2,565,589
Expected values	654,121	148,279	1,763,189	
Chi-square contribution	6	20	7	
Row Percent	26%	6%	69%	88%

Chi-square statistics for all table factors = 662.5405; d.f. = 4; p = $4.494261e^{-142}$

Mixed metal exposure	Up-regulated				Down-regulated			
	Gene	Intron	Unknown	Total	Gene	Intron	Unknown	Total
Count	53,806	10,954	194,138	258,898	104,842	6,881	95,965	207,688
Expected values	66,008	14,963	177,926		52,952	12,003	142,733	
Chi-square contribution	2,256	1,074	1,477		50,849	2,186	15,324	
Row Percent	21%	4%	75%	9%	50%	3%	46%	7%

Mixed metal exposure	No differential regulation			
	Gene	Intron	Unknown	Total
Count	585,241	150,793	1,715,057	2,451,091
Expected values	624,929	141,662	1,684,501	
Chi-square contribution	2,520	589	554	
Row Percent	24%	6%	70%	84%

Chi-square statistics for all table factors = 76829.46; d.f. = 4; p = 0

Sex differences	Up-regulated				Down-regulated			
	Gene	Intron	Unknown	Total	Gene	Intron	Unknown	Total
Count	142,616	4,737	68,803	216,156	93,665	6,398	126,267	226,330
Expected values	55,111	12,493	148,552		57,705	13,081	155,544	
Chi-square contribution	138,940	4,815	42,813		22,409	3,414	5,511	
Row Percent	66%	2%	32%	7%	41%	3%	56%	8%

Sex differences	No differential regulation			
	Gene	Intron	Unknown	Total
Count	507,608	157,493	1,810,090	2,475,191
Expected values	631,073	143,054	1,70,1064	
Chi-square contribution	24,155	1,457	6,988	
Row Percent	21%	6%	73%	85%

Chi-square statistics for all table factors = 250502.2; d.f. = 4; p = 0

SUPPLEMENTAL REFERENCES

S1. Colbourne, J.K., et al., *Phylogenetics and evolution of a circumarctic species complex (Cladocera : Daphnia pulex).* Biological Journal of the Linnean Society, 1998. **65**(3): p. 347-365.

S2. Lynch, M., et al., *The quantitative and molecular genetic architecture of a subdivided species.* Evolution, 1999. **53**(1): p. 100-110.

S3. Lynch, M., *The Origins of Genome Architecture*. 2007, Sunderland, MA: Sinauer Associates, Inc. 389.

S4. Qi, W.H., et al., *Comparative metagenomics of Daphnia symbionts.* Bmc Genomics, 2009. **10**: p. -.

S5. Aparicio, S., et al., *Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes.* Science, 2002. **297**(5585): p. 1301-1310.

S6. Batzoglou, S., et al., *ARACHNE: A whole-genome shotgun assembler.* Genome Research, 2002. **12**(1): p. 177-189.

S7. Huang, X.Q., et al., *PCAP: A whole-genome assembly program.* Genome Research, 2003. **13**(9): p. 2164-2170.

S8. Cristescu, M.E.A., et al., *A microsatellite-based genetic linkage map of the waterflea, Daphnia pulex: On the prospect of crustacean genomics.* Genomics, 2006. **88**(4): p. 415-430.

S9. Salzberg, S.L. and J.A. Yorke, *Beware of mis-assembled genomes.* Bioinformatics, 2005. **21**(24): p. 4320-4321.

S10. Kurtz, S., et al., *Versatile and open software for comparing large genomes.* Genome Biology, 2004. **5**(2): p. -.

S11. Choi, J.H., et al., *A machine-learning approach to combined evidence validation of genome assemblies.* Bioinformatics, 2008. **24**(6): p. 744-750.

S12. Wu, T.D. and C.K. Watanabe, *GMAP: a genomic mapping and alignment program for mRNA and EST sequences.* Bioinformatics, 2005. **21**(9): p. 1859-1875.

S13. Singh-Gasson, S., et al., *Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array.* Nature Biotechnology, 1999. **17**(10): p. 974-978.

S14. Doyle, J.J. and J.L. Doyle, *A rapid DNA isolation procedure for small quantities of fresh leaf tissue.* Phytochem. Bull., 1987. **19**: p. 11–15.

S15. Buchanan-Carter, J. and X. Wang, *Quant-iT™ PicoGreen® dsDNA Protocol for 454 Genome Sequencer*, in *CGB Technical Report*. 2007: doi: 10.2506/cgbtr-200702.

S16. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biology, 2004. **5**(10): p. -.

S17. Tsuchiya, D., B.D. Eads, and M.E. Zolan, *Methods for meiotic chromosome preparation, immunofluorescence, and fluorescence in situ hybridization in Daphnia pulex.* Methods in Molecular Biology, 2009. **558**: p. 235-249.

S18. Salamov, A.A. and V.V. Solovyev, *Ab initio gene finding in Drosophila genomic DNA.* Genome Research, 2000. **10**(4): p. 516-522.

S19. Birney, E. and R. Durbin, *Using GeneWise in the Drosophila annotation experiment.* Genome Research, 2000. **10**(4): p. 547-548.

S20. Korf, I., *Gene finding in novel genomes.* Bmc Bioinformatics, 2004. **5**: p. -.

S21. Haas, B.J., et al., *Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.* Nucleic Acids Research, 2003. **31**(19): p. 5654-5666.

S22. Souvorov, A., T. Tatusova, and D.J. Lipman, *Genome annotation with Gnomon—A multi-step combined gene prediction tool.* ISMB, 2004. **2004**: p. 125.

S23. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Research, 1997. **25**(17): p. 3389-3402.

S24. Burset, M. and R. Guigo, *Evaluation of gene structure prediction programs.* Genomics, 1996. **34**(3): p. 353-367.

S25. Colbourne, J.K., et al., *Sampling Daphnia's expressed genes: preservation, expansion and invention of crustacean genes with reference to insect genomes.* Bmc Genomics, 2007. **8**: p. -.

S26. Tang, Z.J., et al., *ESTPiper - a web-based analysis pipeline for expressed sequence tags.* Bmc Genomics, 2009. **10**: p. -.
S27. Kent, W.J., *BLAT - The BLAST-like alignment tool.* Genome Research, 2002. **12**(4): p. 656-664.
S28. Frohlich, T., et al., *LC-MS/MS-based proteome profiling in Daphnia pulex and Daphnia longicephala: the Daphnia pulex genome database as a key for high throughput proteomics in Daphnia.* Bmc Genomics, 2009. **10**: p. -.
S29. Olmstead, A.W. and G.A. Leblanc, *Juvenoid hormone methyl farnesoate is a sex determinant in the crustacean Daphnia magna.* Journal of Experimental Zoology, 2002. **293**(7): p. 736-739.
S30. Lopez, J. and J.K. Colbourne, *Dual-Labeled Expression Analysis Protocol for NimbleGen Microarrays: Laboratory handbook for Environmental Genomics courses at Mount Desert Island Biological Laboratory and at Indiana University*, in *CGB Technical Report*. 2010: doi:10.2506/cgbtr-201001.
S31. Tollrian, R., *Neckteeth Formation in Daphnia-Pulex as an Example of Continuous Phenotypic Plasticity - Morphological Effects of Chaoborus Kairomone Concentration and Their Quantification.* Journal of Plankton Research, 1993. **15**(11): p. 1309-1318.
S32. Shaw, J.R., et al., *Gene response profiles for Daphnia pulex exposed to the environmental stressor cadmium reveals novel crustacean metallothioneins.* Bmc Genomics, 2007. **8**: p. -.
S33. Kilham, S.S., et al., *COMBO: a defined freshwater culture medium for algae and zooplankton.* Hydrobiologia, 1998. **377**: p. 147-159.
S34. Folt, C.L., et al., *Synergism and antagonism among multiple stressors.* Limnology and Oceanography, 1999. **44**(3): p. 864-877.
S35. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-193.
S36. Kampa, D., et al., *Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.* Genome Research, 2004. **14**(3): p. 331-342.
S37. Gentleman, R., et al. *The R Project for Statistical Computing*. 2009; Available from: http://www.r-project.org/.
S38. Smyth, G.K., *Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**.
S39. Colbourne, J.K. and M. Pfrender. *Daphnia: the companion papers for the genome sequence*. 2009; Available from: http://www.biomedcentral.com/series/Daphnia.
S40. Kendziorski, C.M., et al., *On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles.* Statistics in Medicine, 2003. **22**(24): p. 3899-3914.
S41. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.
S42. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Research, 2004. **32**: p. D258-D261.
S43. Koonin, E.V., et al., *A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.* Genome Biology, 2004. **5**(2): p. -.
S44. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Research, 2004. **32**: p. D277-D280.
S45. Dehal, P.S. and J.L. Boore, *A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database.* Bmc Bioinformatics, 2006. **7**: p. -.
S46. Laslett, D. and B. Canback, *ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences.* Nucleic Acids Research, 2004. **32**(1): p. 11-16.
S47. Lowe, T.M. and S.R. Eddy, *tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence.* Nucleic Acids Research, 1997. **25**(5): p. 955-964.
S48. Gerlach, D., et al., *miROrtho: computational survey of microRNA genes.* Nucleic Acids Research, 2009. **37**: p. D111-D117.

S49. Rho, M., et al., *De novo identification of LTR retrotransposons in eukaryotic genomes.* Bmc Genomics, 2007. **8**: p. -.
S50. Rho, M.N. and H.X. Tang, *MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes.* Nucleic Acids Research, 2009. **37**(21): p. -.
S51. Price, A.L., N.C. Jones, and P.A. Pevzner, *De novo identification of repeat families in large genomes.* Bioinformatics, 2005. **21**: p. I351-I358.
S52. Feschotte, C., et al., *Exploring Repetitive DNA Landscapes Using REPCLASS, a Tool That Automates the Classification of Transposable Elements in Eukaryotic Genomes.* Genome Biology and Evolution, 2009: p. 205-220.
S53. Gilbert, D. *Daphnia Gene Structure*. 2007; Available from: http://wfleabase.org/genome-summaries/gene-structure/.
S54. Gilbert, D. *euGenes/ Arthropod genomes* 2008; Available from: http://arthropods.eugenes.org/arthropods/.
S55. Gilbert, D. *Online Supplemental Material to "The Ecoresponsive Genome of Daphnia pulex"*. 2010; Available from: http://wfleabase.org/release1/current_release/supplement/
S56. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* Bmc Bioinformatics, 2003. **4**: p. -.
S57. Wolf, Y.I., L. Carmel, and E.V. Koonin, *Unifying measures of gene function and evolution.* Proceedings of the Royal Society B-Biological Sciences, 2006. **273**(1593): p. 1507-1515.
S58. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Research, 2004. **32**(5): p. 1792-1797.
S59. Rogozin, I.B., et al., *Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution.* Current Biology, 2003. **13**(17): p. 1512-1517.
S60. Carmel, L., et al., *Patterns of intron gain and conservation in eukaryotic genes.* Bmc Evolutionary Biology, 2007. **7**: p. -.
S61. Kriventseva, E.V., et al., *OrthoDB: the hierarchical catalog of eukaryotic orthologs.* Nucleic Acids Research, 2008. **36**: p. D271-D275.
S62. Richards, S., et al., *The genome of the model beetle and pest Tribolium castaneum.* Nature, 2008. **452**(7190): p. 949-955.
S63. Werren, J.H., et al., *Functional and evolutionary insights from the genomes of three parasitoid Nasonia species (vol 327, pg 343, 2010).* Science, 2010. **327**(5973): p. 1577-1577.
S64. Elsik, C.G., et al., *The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution.* Science, 2009. **324**(5926): p. 522-528.
S65. Kirkness, E.F., et al., *Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(27): p. 12168-12173.
S66. Rivera, A.S., et al., *Gene duplication and the origins of morphological complexity in pancrustacean eyes, a genomic approach.* BMC Evolutionary Biology, 2010. **10**: p. 123.
S67. Seetharam, A., Y. Bai, and G.W. Stuart, *A survey of well conserved families of C2H2 zinc-finger genes in Daphnia.* Bmc Genomics, 2010. **11**: p. -.
S68. Thomson, S.A., et al., *Annotation, phylogenetics, and expression of the nuclear receptors in Daphnia pulex.* Bmc Genomics, 2009. **10**: p. -.
S69. Wilson, K.H.S., *The genome sequence of the protostome Daphnia pulex encodes respective orthologues of a neurotrophin, a Trk and a p75NTR: Evolution of neurotrophin signaling components and related proteins in the bilateria.* Bmc Evolutionary Biology, 2009. **9**: p. -.
S70. McTaggart, S.J., et al., *The components of the Daphnia pulex immune system as revealed by complete genome sequencing.* Bmc Genomics, 2009. **10**: p. -.
S71. Penalva-Arana, D.C., M. Lynch, and H.M. Robertson, *The chemoreceptor genes of the waterflea Daphnia pulex: many Grs but no Ors.* Bmc Evolutionary Biology, 2009. **9**: p. -.

S72. Schurko, A.M., J.M. Logsdon, and B.D. Eads, *Meiosis genes in Daphnia pulex and the role of parthenogenesis in genome evolution.* Bmc Evolutionary Biology, 2009. **9**: p. -.
S73. Sturm, A., P. Cunningham, and M. Dean, *The ABC transporter gene family of Daphnia pulex.* Bmc Genomics, 2009. **10**: p. -.
S74. Baldwin, W.S., P.B. Marko, and D.R. Nelson, *The cytochrome P450 (CYP) gene superfamily in Daphnia pulex.* Bmc Genomics, 2009. **10**: p. -.
S75. Matsui, T., et al., *Expression profiles of urbilaterian genes uniquely shared between honey bee and vertebrates.* Bmc Genomics, 2009. **10**: p. -.
S76. Lemay, D.G., et al., *The bovine lactation genome: insights into the evolution of mammalian milk.* Genome Biology, 2009. **10**(4): p. -.
S77. Waterhouse, R.M., et al., *Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes.* Science, 2007. **316**(5832): p. 1738-1743.
S78. Wyder, S., et al., *Quantification of ortholog losses in insects and vertebrates.* Genome Biology, 2007. **8**(11): p. -.
S79. Gilbert, D. *OrthoMCL clustering among 14 arthropod proteomes (ARP2)*. 2009; Available from: http://arthropods.eugenes.org/arthropods/orthologs/.
S80. Li, L., C.J. Stoeckert, and D.S. Roos, *OrthoMCL: Identification of ortholog groups for eukaryotic genomes.* Genome Research, 2003. **13**(9): p. 2178-2189.
S81. Chen, F., et al., *Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes.* Plos One, 2007. **2**(4): p. -.
S82. van Dongen, S. *MCL - a cluster algorithm for graphs*. 2000; Available from: http://www.micans.org/mcl/index.html?sec_software.
S83. Wilson, D., et al., *The SUPERFAMILY database in 2007: families and functions.* Nucleic Acids Research, 2007. **35**: p. D308-D313.
S84. Castresana, J., *Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.* Molecular Biology and Evolution, 2000. **17**(4): p. 540-552.
S85. Sakarya, O., K.S. Kosik, and T.H. Oakley, *Reconstructing ancestral genome content based on symmetrical best alignments and Dollo parsimony.* Bioinformatics, 2008. **24**(5): p. 606-612.
S86. Conant, G.C. and A. Wagner, *GenomeHistory: a software tool and its application to fully sequenced genomes.* Nucleic Acids Research, 2002. **30**(15): p. 3378-3386.
S87. Hubbard, T.J.P., et al., *Ensembl 2009.* Nucleic Acids Research, 2009. **37**: p. D690-D697.
S88. Larkin, M.A., et al., *Clustal W and clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-2948.
S89. Yang, Z.H. and R. Nielsen, *Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.* Molecular Biology and Evolution, 2000. **17**(1): p. 32-43.
S90. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151-1155.
S91. Lynch, M., *The Origins of Genome Architecture*. 2007, Sunderland, MA: Sinauer Assocs., Inc.
S92. Gilbert, D. *Daphnia tandem genes: Rationale for analysis using Tandy*. 2007; Available from: http://eugenes.org/gmod/tandy/.
S93. Tweedie, S., et al., *FlyBase: enhancing Drosophila Gene Ontology annotations.* Nucleic Acids Research, 2009. **37**: p. D555-D559.
S94. Kashiyama, K., et al., *Molecular Characterization of Visual Pigments in Branchiopoda and the Evolution of Opsins in Arthropoda.* Molecular Biology and Evolution, 2009. **26**(2): p. 299-311.
S95. Oakley, T.H. and D.R. Huber, *Differential expression of duplicated opsin genes in two eye types of ostracod crustaceans.* Journal of Molecular Evolution, 2004. **59**(2): p. 239-249.
S96. Stamatakis, A., *RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.* Bioinformatics, 2006. **22**(21): p. 2688-90.
S97. Plachetzki, D.C., et al., *The origins of novel protein interactions during animal opsin evolution.* PLoS ONE, 2007. **2**(10): p. e1054.

S98. Apweiler, R., et al., *The Universal Protein Resource (UniProt) in 2010.* Nucleic Acids Research, 2010. **38**: p. D142-D148.
S99. Sawyer, S., *Statistical Tests for Detecting Gene Conversion.* Molecular Biology and Evolution, 1989. **6**(5): p. 526-538.
S100. Yang, Z.H., *PAML 4: Phylogenetic analysis by maximum likelihood.* Molecular Biology and Evolution, 2007. **24**(8): p. 1586-1591.
S101. Kimura, S., et al., *Heterogeneity and differential expression under hypoxia of two-domain hemoglobin chains in the water flea, Daphnia magna.* Journal of Biological Chemistry, 1999. **274**(15): p. 10649-10653.
S102. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice.* Nucleic Acids Research, 1994. **22**(22): p. 4673-4680.
S103. Ronquist, F. and J.P. Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models.* Bioinformatics, 2003. **19**(12): p. 1572-1574.
S104. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(25): p. 14863-14868.
S105. Gu, Z.L., et al., *Duplicate genes increase gene expression diversity within and between species.* Nature Genetics, 2004. **36**(6): p. 577-579.
S106. Jensen, L.J., et al., *eggNOG: automated construction and annotation of orthologous groups of genes.* Nucleic Acids Research, 2008. **36**: p. D250-D254.
S107. Saebo, P.E., et al., *PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology.* Nucleic Acids Research, 2005. **33**: p. W535-W539.
S108. Letunic, I., et al., *iPath: interactive exploration of biochemical pathways and networks.* Trends in Biochemical Sciences, 2008. **33**(3): p. 101-103.
S109. Felsenstein, J., *PHYLIP (Phylogeny Inference Package), version 3.57c*. 1995: Department of Genetics, University of Washington, Seattle.
S110. Jones, D.T., W.R. Taylor, and J.M. Thornton, *The Rapid Generation of Mutation Data Matrices from Protein Sequences.* Computer Applications in the Biosciences, 1992. **8**(3): p. 275-282.
S111. Hashimoto, K., et al., *Comprehensive analysis of glycosyltransferases in eukaryotic genomes for structural and functional characterization of glycans.* Carbohydrate Research, 2009. **344**(7): p. 881-887.
S112. Weir, B.S. and C.C. Cockerham, *Estimating F-Statistics for the Analysis of Population-Structure.* Evolution, 1984. **38**(6): p. 1358-1370.
S113. Lynch, M. and B. Walsh, *Genetics and Analysis of Quantitative Traits*. 1999, Sunderland, MA: Sinauser Assoc. 980.
S114. Beaton, M.J. and P.D.N. Hebert, *Variation in Chromosome-Numbers of Daphnia (Crustacea, Cladocera).* Hereditas, 1994. **120**(3): p. 275-279.
S115. Zaffagnini, F., *Reproduction in Daphnia*, in *Daphnia*, R.H. Peters and R. de Bernardi, Editors. 1987, Memorie dell'Istituto Italiano di Idrobiologia. p. 245-284.
S116. Routtu, J., et al., *The first-generation Daphnia magna linkage map.* BMC Genomics, 2010. **11**: p. 508.
S117. Robertson, H.M. and K.H.J. Gordon, *Canonical TTAGG-repeat telomeres and telomerase in the honey bee, Apis mellifera.* Genome Research, 2006. **16**(11): p. 1345-1351.
S118. Fujiwara, H., et al., *Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, Bombyx mori.* Chromosome Research, 2005. **13**(5): p. 455-467.
S119. George, J.A., et al., *Genomic organization of the Drosophila telomere retrotransposable elements.* Genome Research, 2006. **16**(10): p. 1231-1240.
S120. Osanai, M., et al., *Identification and characterization of the telomerase reverse transcriptase of Bombyx mori (silkworm) and Tribolium castaneum (flour beetle).* Gene, 2006. **376**(2): p. 281-289.
S121. Altschul, S.F., et al., *Basic Local Alignment Search Tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.

S122. Adamowicz, S.J., et al., *The scale of divergence: A phylogenetic appraisal of intercontinental allopatric speciation in a passively dispersed freshwater zooplankton genus.* Molecular Phylogenetics and Evolution, 2009. **50**(3): p. 423-436.
S123. Haag, C.R., et al., *Nucleotide Polymorphism and Within-Gene Recombination in Daphnia magna and D. pulex, Two Cyclical Parthenogens.* Genetics, 2009. **182**(1): p. 313-323.
S124. Ronshaugen, M., et al., *The Drosophila microRNA iab-4 causes a dominant homeotic transformation of halteres to wings.* Genes & Development, 2005. **19**(24): p. 2947-2952.
S125. Tyler, D.M., et al., *Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci.* Genes & Development, 2008. **22**(1): p. 26-36.
S126. Stark, A., et al., *A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands.* Genes & Development, 2008. **22**(1): p. 8-13.
S127. Bender, W., *MicroRNAs in the Drosophila bithorax complex.* Genes & Development, 2008. **22**(1): p. 14-19.
S128. Shiga, Y., et al., *Transcriptional readthrough of Hox genes Ubx and Antp and their divergent post-transcriptional control during crustacean evolution.* Evolution & Development, 2006. **8**(5): p. 407-414.
S129. Penton, E.H., B.W. Sullender, and T.J. Crease, *Pokey, a new DNA transposon in Daphnia (Cladocera : Crustacea).* Journal of Molecular Evolution, 2002. **55**(6): p. 664-673.
S130. Arendt, D., *Evolution of eyes and photoreceptor cell types.* International Journal of Developmental Biology, 2003. **47**(7-8): p. 563-571.
S131. Terakita, A., *The opsins.* Genome Biology, 2005. **6**(3): p. -.
S132. Provencio, I., et al., *Melanopsin: An opsin in melanophores, brain, and eye.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(1): p. 340-345.
S133. Arendt, D., et al., *Ciliary photoreceptors with a vertebrate-type opsin in an invertebrate brain.* Science, 2004. **306**(5697): p. 869-871.
S134. Velarde, R.A., et al., *Pteropsin: A vertebrate-like non-visual opsin expressed in the honey bee brain.* Insect Biochemistry and Molecular Biology, 2005. **35**(12): p. 1367-1377.
S135. Bellingham, J., D.J. Wells, and R.G. Foster, *In silico characterisation and chromosomal localisation of human RRH (peropsin) implications for opsin evolution.* Bmc Genomics, 2003. **4**: p. -.
S136. Hill, C.A., et al., *G protein coupled receptors in Anopheles gambiae.* Science, 2002. **298**(5591): p. 176-178.
S137. Sun, H., et al., *Peropsin, a novel visual pigment-like protein located in the apical microvilli of the retinal pigment epithelium.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(18): p. 9893-9898.
S138. Tarttelin, E.E., et al., *Neuropsin (Opn5): a novel opsin identified in mammalian neural tissue.* Febs Letters, 2003. **554**(3): p. 410-416.
S139. Pandey, S., et al., *Cytoplasmic Retinal Localization of an Evolutionary Homolog of the Visual Pigments.* Experimental Eye Research, 1994. **58**(5): p. 605-613.
S140. Kojima, D., et al., *Novel G(o)-mediated phototransduction cascade in scallop visual cells.* Journal of Biological Chemistry, 1997. **272**(37): p. 22979-22982.
S141. Raible, F., et al., *Opsins and clusters of sensory G-protein-coupled receptors in the sea urchin genome.* Developmental Biology, 2006. **300**(1): p. 461-475.
S142. Oakley, T.H., *On homology of arthropod compound eyes.* Integrative and Comparative Biology, 2003. **43**(4): p. 522-530.
S143. Smith, K.C. and E.R. Macagno, *Uv Photoreceptors in the Compound Eye of Daphnia-Magna (Crustacea, Branchiopoda) - a 4th Spectral Class in Single Ommatidia.* Journal of Comparative Physiology a-Sensory Neural and Behavioral Physiology, 1990. **166**(5): p. 597-606.
S144. Schehr, R.S., *Spectral sensitivities of anatomically identified photoreceptors in the compound eye of Daphnia magna.* 1984, Columbia University.

S145. Cronin, T.W., N.J. Marshall, and R.L. Caldwell, *Spectral tuning and the visual ecology of mantis shrimps.* Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 2000. **355**(1401): p. 1263-1267.
S146. Porter, M.L., et al., *Molecular characterization of crustacean visual pigments and the evolution of pancrustacean opsins.* Molecular Biology and Evolution, 2007. **24**(1): p. 253-268.
S147. Felsenstein, J., *Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading.* Systematic Zoology, 1978. **27**(4): p. 401-410.
S148. Pond, S.L.K., S.D.W. Frost, and S.V. Muse, *HyPhy: hypothesis testing using phylogenies.* Bioinformatics, 2005. **21**(5): p. 676-679.
S149. Rubin, E.B., et al., *Molecular and phylogenetic analyses reveal mammalian-like clockwork in the honey bee (Apis mellifera) and shed new light on the molecular evolution of the circadian clock.* Genome Research, 2006. **16**(11): p. 1352-1365.
S150. Sakamoto, K., et al., *Two opsins from the compound eye of the crab Hemigrapsus sanguineus.* Journal of Experimental Biology, 1996. **199**(2): p. 441-450.
S151. Hariyama, T., et al., *Primary Structure of Crayfish Visual Pigment Deduced from Cdna.* Febs Letters, 1993. **315**(3): p. 287-292.
S152. Smith, W.C., et al., *Opsins from the Lateral Eyes and Ocelli of the Horseshoe-Crab, Limulus-Polyphemus.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(13): p. 6150-6154.
S153. Colbourne, J.K., V.R. Singan, and D.G. Gilbert, *wFleaBase: The Daphnia genome database.* Bmc Bioinformatics, 2005. **6**: p. -.
S154. US_Department_of_Energy. *JGI Daphnia Genome Portal*. 2007; Available from: http://www.jgi.doe.gov/Daphnia/.
S155. Thomas, J.H., *Analysis of homologous gene clusters in Caenorhabditis elegans reveals striking regional cluster domains.* Genetics, 2006. **172**(1): p. 127-143.
S156. Plachetzki, D.C., B.M. Degnan, and T.H. Oakley, *The origins of novel protein interactions during animal opsin evolution.* Plos One, 2007. **2**(10): p. e1054.
S157. Velarde, R.A., et al., *Pteropsin: a vertebrate-like non-visual opsin expressed in the honey bee brain.* Insect Biochem Mol Biol, 2005. **35**(12): p. 1367-77.
S158. Pond, S.L., S.D. Frost, and S.V. Muse, *HyPhy: hypothesis testing using phylogenies.* Bioinformatics, 2005. **21**(5): p. 676-9.
S159. Felsenstein, J., *Cases in which parsimony or compatibility methods will be positively misleading.* Systematic Zoology, 1978. **27**: p. 401-410.
S160. Chen, K., D. Durand, and M. Farach-Colton, *NOTUNG: A program for dating gene duplications and optimizing gene family trees.* Journal of Computational Biology, 2000. **7**(3-4): p. 429-447.
S161. Springer, B.A., et al., *Mechanisms of Ligand Recognition in Myoglobin.* Chemical Reviews, 1994. **94**(3): p. 699-714.
S162. Yang, J., et al., *The Structure of Ascaris Hemoglobin Domain-I at 2.2 Angstrom Resolution - Molecular-Features of Oxygen Avidity.* Proceedings of the National Academy of Sciences of the United States of America, 1995. **92**(10): p. 4224-4228.
S163. Carver, T.E., et al., *A Novel Site-Directed Mutant of Myoglobin with an Unusually High O-2 Affinity and Low Autooxidation Rate.* Journal of Biological Chemistry, 1992. **267**(20): p. 14443-14450.
S164. Gilbert, D. *PASA database for Daphnia pulex*. 2008; Available from: http://wfleabase.org/genome/Daphnia_pulex/current/pasa/.
S165. Ye, Y.Z. and A. Godzik, *Comparative analysis of protein domain organization.* Genome Research, 2004. **14**(3): p. 343-353.
S166. Colbourne, J.K. *The Daphnia Genomics Consortium*. 2003; Available from: http://daphnia.cgb.indiana.edu.
S167. Choi, J.H. *Daphnia pulex scaffold dotplot*. 2007; Available from: http://cancer.informatics.indiana.edu/cgi-bin/jeochoi/daphnia/tandemduplicategene/index.cgi.

S168. Colbourne, J.K. and M. Pfrender. *Daphnia NIH Model Organisms for Biomedical Research*. 2009; Available from: http://www.nih.gov/science/models/daphnia/.
S169. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Research, 2010. **38**: p. D5-D16.
S170. Colbourne, J.K., et al., *Five hundred and twenty-eight microsatellite markers for ecological genomic investigations using Daphnia.* Molecular Ecology Notes, 2004. **4**(3): p. 485-490.
S171. Puigbò, P., Y.I. Wolf, and E.V. Koonin, *Search for a 'Tree of Life' in the thicket of the phylogenetic forest.* Journal of Biology, 2009. **8**(6): p. 59.
S172. Rho, M., et al., *LTR retroelements in the genome of Daphnia pulex.* BMC Genomics, 2010. **11**: p. 425.
S173. Schaack, S., et al., *DNA transposons and the role of recombination in mutation accumulation in Daphnia pulex.* Genome Biology, 2010. **11**(4): p. -.
S174. Kaminker, J.S., et al., *The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective.* Genome Biology, 2002. **3**: p. research0084
S175. Quesneville, H., et al., *Combined evidence annotation of transposable elements in genome sequences.* Plos Computational Biology, 2005. **1**(2): p. 166-175.
S176. Nene, V., et al., *Genome sequence of Aedes aegypti, a major arbovirus vector.* Science, 2007. **316**(5832): p. 1718-1723.
S177. Holt, R.A., et al., *The genome sequence of the malaria mosquito Anopheles gambiae.* Science, 2002. **298**(5591): p. 129-+.
S178. Weinstock, G.M., et al., *Insights into social insects from the genome of the honeybee Apis mellifera.* Nature, 2006. **443**(7114): p. 931-949.
S179. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-562.
S180. Mayer, C., F. Leese, and R. Tollrian, *Genome-wide analysis of tandem repeats in Daphnia pulex - a comparative approach.* BMC Genomics, 2010. **11**: p. 277.
S181. Huerta-Cepas, J., et al., *PhylomeDB: a database for genome-wide collections of gene phylogenies.* Nucleic Acids Research, 2008. **36**: p. D491-D496.
S182. Zvokelj, M., S. Zupan, and I. Prebil, *Multivariate and multiscale monitoring of large-size low-speed bearings using Ensemble Empirical Mode Decomposition method combined with Principal Component Analysis.* Mechanical Systems and Signal Processing, 2010. **24**(4): p. 1049-1067.
S183. Woollard, A. (June 25, 2005) *Gene duplications and genetic redundancy in C. elegans*. DOI: doi/10.1895/wormbook.1.2.1.