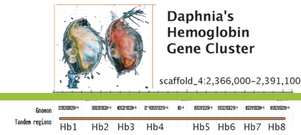
Unlocated Arthropod genes, and ways to find them



2008

Don Gilbert, Indiana University

Abstract

The *Daphnia pulex* genome is rich in tandem duplicate genes, some 20% of its 30,000+ genes. However some gene predictors have missed or incorrectly located almost half of these. Estimates from genome-wide tile expression suggest an additional 5,000 genes have been missed. Gene prediction for new genomes such as this first crustacean is still an uncertain task. Even in clades with a well-characterized model such as *Drosophila*, gene finding remains an uncertain task. Prediction tools are increasingly sophisticated and accurate. Today's methods draw on the range of available gene evidence and improved modeling of gene structures. Yet they are sensitive to available gene data and expected structures. They find well-known genes, but fail at accurate detection of novel and diverged genes. Measures from gene duplication and genome-wide tile expression can more accurately locate those genes missed by other methods. Computational methods are being developed to turn these signals to accurate gene models. Application of these methods to arthropod genomes, including *Daphnia* and *Drosophila*, uncovers some 10% to 25% additional species specific and diverged genes. This work includes development of new automated genome analysis pipelines on NSF TeraGrid shared cyberinfrastructure, as part of the Generic Model Organism Database project.

Contact: Don Gilbert, gilbertd@indiana.edu

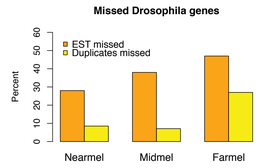
Environmental stresses find novel genes

Most genes are expressed in unusual environs, and rather specific

Homology isn't enough to find *Daphnia* genes. Novel genes are overabundant among stress treatments. This Table 1 lists the percent of genes with EST from stress treatments and normal controls. Protein homology is found for less than half the genes. Novel genes with ESTs expressed only in stress environs are the most abundant group. Data is from 15,400 Daphnia genes with EST expression under several treatments of inorganic and biotic stresses, such as toxic metals, hypoxia, and predation (Daphnia Genome Consortium).

Table 1 *Daphnia* novel genes show up under stresses.

Stress Treatment	Homology? Yes	No	Novel genes
Biotic	7%	9%	1400 / 2580
Inorganic	19%	24%	3700 / 6600
Normal	23%	17%	2630 / 6230

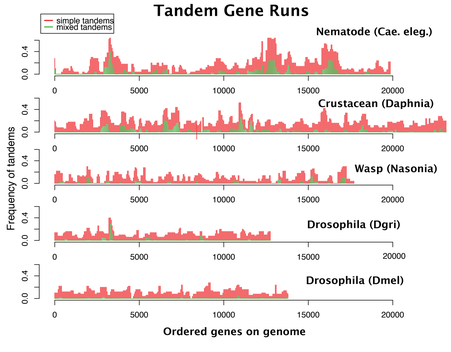


A similar effect holds for *Drosophila* species. Expressed genes are poorly found as homology with *D. melanogaster* declines. Novel genes are poorly predicted, as protein homology and prediction trained with Dmel will miss these. This figure summarizes species group percentages for ESTs , and duplicate genes, that are missed by gene predictions. Most misses are those lacking Dmel homology. EST data and gene duplicates for 9 Drosophila species with >10,000 ESTs in dbEST are use. These are matched to species gene predictions (GleanR). Groups are Near-mel with Dsim, Dyak, Dere, Mid-mel with Dana, Dpse, and Far-mel with Dwil, Dvir, Dmoj, Dgri.

Duplicate genes are common

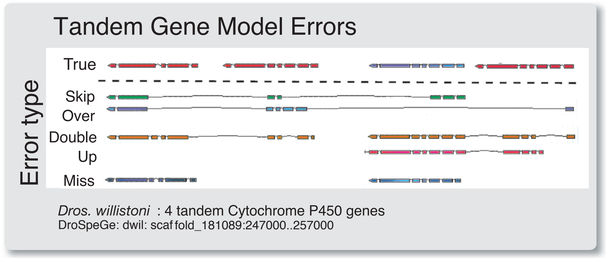
Duplicate genes are common, a computational problem, also an aid to finding genes

Duplicate genes are frequent, and very near (1Kb) tandem duplicates are especially common in *Daphnia*, exceeding the duplicate rich *Cae. elegans*. One aspect of genome biology that is difficult to model is a cluster of nearby duplicate genes. Nearby near-identical exons can confuse computational methods that use alignment, including BLAST, GeneWise and similar gene mappers that align a protein to find genes. Ab initio predictors also can fail to distinguish exons belonging to nearby genes. The initial set of Daphnia gene predictions had many errors finding these, with 5,000 predicted genes spanning two or more distinct matches to the same protein.



Tandem gene runs in several bug genomes (Figure 2) shows hot spots, a few multi-gene duplicate clusters, and wide-spread simple tandem duplicates. The evidence here and elsewhere indicates these are not segmental duplications, but mainly simple duplicated genes.

Duplicate genes confuse gene predictors

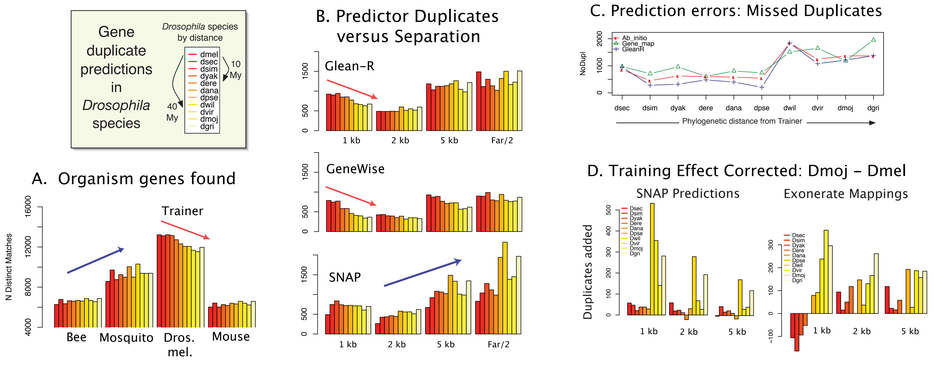


Tandem gene prediction error classes are diagrammed in Figure 3, below a cluster of four nearby, near-identical genes. These error classes are termed skip over, double up, and miss. **Skip over** is a gene model from a subset of exons in a collection of two or more tandem genes. A skip over model protein may be substantially identical to the two or more genes it contains. **Double up** includes all or most exons of two or more genes are predicted, but are joined as one. A double up protein model differs from the true model, and can often be computationally detected as having duplicate motifs. **Missed duplicate** occurs where some of the duplicate genes in a cluster have missed all exon predictions. This example is taken from a case in the *Dros willistoni* genome, where no single predictor correctly called all four Cytochrome P450 genes. However, among 13 predictors were cases of a true model for each gene.

Duplicate genes help correct errors

Same or near species training reduces prediction errors and phylogenetic artifacts

These methods of gene duplicate detection have been applied to predictions for 12 *Drosophila* species genomes. It is one way to independently check predictions without reliance on comparison to the reference species (*Dros. melanogaster*). These tests use only same-species gene duplications. Gene homology content of the twelve Drosophila from perspective of Dmel, two other insects and mouse genes are shown in Figure 5A. This bar graph shows different clines, one for Dmel matching best the near-Dmel group, while the other informant species match the far-Dmel group best.



In Figure 5B, gene predictions by GeneWise and Glean-R show a lower rate of tandem genes predicted for the far-Dmel species. In contrast, *ab initio* predictors (e.g SNAP) show a smaller or no cline, or a reverse cline consistent comparable to that found for non-Dmel organism gene sets. The dilemma expressed in Figure 5B, of inconsistent clines in duplicates among gene predictors, can be explained in large part by prediction errors, with results shown in Figure 5C. This species-bias error is eliminated by training the predictors with same or near-species gene data, as shown in Figure 5D for two gene calling methods (SNAP, Exonerate).

The bar graphs of Figures 5 A-B show gene counts for each of 10 species, arranged phylogenetically in heat colors from near-Dmel (red) to far-Dmel (yellow). Species Dsec, Dsim, Dyak, Dere are nearest Dmel, with Dana, Dpse intermediate. Species Dwil, Dvir, Dmoj and Dgri are the far-Dmel group.

Drosophila Odorant genes and the spurious cline

Prediction errors in Odorant-Binding proteins (Obp) of *Drosophila* species were assessed using results of Vieira *et al*. (2007), and further analysis and curation. GleanR is the combined prediction set. Vieira added genes found using Psi-Blast trained on Dmel and Dpse genes. This work added genes found with Psi-Blast trained on each species group Obp genes. Errors are the sum of these added genes. Computational errors (missed Obp genes) are significantly more common with phylogenetic group distance from *Dros. melanogaster* (averaged groups; ANOVA $p < 0.02$ 5). Gene counts (GleanR and Curated) are not statistically different.

The phylogenetically near Dmel group includes dsec, dsim, dyak, dere, Midmel group includes dana, dper, dpse, and Farmel includes dwil, dmoj, dvir, dgri. This work found 2 probable Obp genes that were mis-predicted due to assembly gaps (gap; counted as error), and 3 pseudogenic fragments (psi; not counted as error). Pseudogenic genes found by Vieira are not listed.

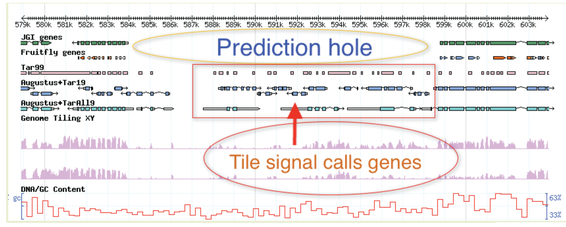
Table 2. Odor gene duplicates, prediction errors in *Drosophila*

Group Ave.	Predict	Curated	Errors
Near-Dmel	53.5	55	1.5
Mid-Dmel	47	51	4 *
Far-Dmel	49.8	55.3	5.5 *

Species	Predicted GleanR	Total Curated	Prediction Errors
dsec	54	54	0
dsim	52	55	3
dyak	56	58	2
dere	52	53	1
dana	48	53	5
dper	46	50	4
dpse	47	50	3
dwil	55	65	10
dmoj	45	49	4
dvir	43	47	4
dgri	56	60	4

Genome tile expession finds novel genes

Gene calls from Daphnia tile expression experiments finds 26% coding sequence bases over all the genome, compared to 17% from gene predictions. This adds 5,000 to 10,000 new genes to 30,000 predicted for Daphnia. This is similar to the result with *Dros. melanogaster* by Manak *et al* 2006, from 18% reference CDS/genome, a higher 24% is found with tile expression.



What does tile expression uncover? Among novel tile expression genes, 10% have homology, 19% have EST support (25% have EST or protein support). This technology is as yet not fully developed. Computational tools need to mature to fully incorporate tile expression with gene and feature annotation. The current results are qualified, with lower quality gene models and uncertainty in the magnitude of errors (both misses and spurious predictions).

Table 3. *Daphnia's* tile-found novel genes with homology

Species group	Count
Insects	311
Nasonia + Apis	175
Tribolium	82
Aedes + Drosophila	50
Aquatic (Zebrafish, Sea urchin, anemone)	215
Bacteria	238
Transposon genes	132
Other	577

Count	Novel Protein type
277	Hypothetical protein
49	Transposase
25	Orf2-encoded protein
19	transcriptional regulator
11	Peptidase
9	abc transporter related
7	major facilitator superfamily mfs_1
6	ankyrin repeat protein
4	heavy metal translocating p-type atpase
4	inner-membrane translocator
4	tonb-dependent siderophore receptor
4	two component transcriptional regulator

Summary of ways to find unlocated Arthropod genes

Most genes are expressed in unusual environs, and rather specific

Use many environmental, developmental and tissue conditions to see range of genes via expression. Understand the limits of gene homology.

Duplicate genes are common, a computational problem, also an aid to finding genes

Examine duplicate genes carefully. Tools that distinguish these can be used to find paralogs missed by traditional methods.

Same or near species training reduces prediction errors and phylogenetic artifacts

Use same-species and near-species data as much as possible in preparing automated annotations. Be aware of and control for informant species-distance as a source of bias.

Genome tile expression finds genes that predictors miss

As an alternative to EST studies, it has values and drawbacks. Computational methods need to improve to use this data well.

Acknowledgements

These analyses were performed with support to Don Gilbert from the National Science Foundation (DBI-0640462) and the National Institutes of Health, including TeraGrid award (TG-MCB060059) for computing resources provided by Indiana University, NCSA, and SDSC. *Daphnia pulex* sequencing and portions of the analyses were performed at the DOE Joint Genome Institute under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program, and in collaboration with the Daphnia Genomics Consortium (DGC) http://daphnia.cgb.indiana.edu.

References

Gilbert, D.G. 2007. New and old genes in new and old Drosophila genome. In preparation . http://insects.eugenes.org/DroSpeGe/about/analysis-doc/

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

Gilbert, D.G. 2008. Tandem genes lost + found. In preparation . http://insects.eugenes.org/DroSpeGe/about/analysis-doc/

Gorr, T.A., J. D. Cahn, H. Yamagata, and H. F Bunn, 2004. Hypoxia-induced Synthesis of Hemoglobin in the Crustacean Daphnia magna. J. Biol. Chemistry, 279(34):36038–36047,

Haas, B.J., Delcher, A.L., Mount, S.M., et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. NAR, 31, 5654-5666. http://pasa.sourceforge.net/

Manak, JR et al., 2006. Biological function of unannotated transcription during the early develop of Drosophila melanogaster. Nature Genetics, 38(10): 1151-1158 doi:10.1038/ng1875

Souvorov, A., et al. (2006) Gnomon annotation of Drosophila species genomes. URL: ftp://ftp.ncbi.nih.gov/genomes/Drosophila_melanogaster/special_requests/CAF1/

Vieira, F.G., Sanchez-Gracia, A., Rozas, J. 2007. Comparative genomic analysis of the Odorant-I Protein family in 12 Drosophila genomes: Purifying selection and birth-and-death evolution Genc Biology 2007, 8:R235 doi:10.1186/gb-2007-8-11-r235