# A Genome Grid for Finding new Bug Genes

2008.07

Don Gilbert, Indiana University, gilbertd@indiana.edu

## Abstract

Accurate gene prediction and automated annotation is lagging behind needs for the rapidly growing number of new genomes. Prediction tools are increasingly sophisticated and accurate. These draw on the range of available gene evidence and improved modeling of gene structures. Yet they are sensitive to available organism data and expected structures. Detection of novel and diverged genes remains problematic. Even for species clades with a well-characterized model such as *Drosophila*, gene finding is an uncertain task. Next generation genomics technology, such as genome-wide tile expression, finds thousands of genes have been missed by current methods. Combining **Next-Gen genome data** with prediction tools for both new and well-characterized Arthropod genomes uncovers 10% to **50% new species genes** and diverged genes.

We envision **Genome Grid** analysis pipelines that many scientists can use, installed on NSF TeraGrid and other shared cyberinfrastructure, as part of a **Generic Model Organism Database** project (GMOD, gmod.org/Genome_grid). Middleware and science gateway methods to use computational grids for genome analyses are in development as an open source project. Genome community software is now available, at **TeraGrid tg_community/genomes**, and for local use. These tools help genome analyses keep pace with rapid expansion of genomics information.

**Contact**: Don Gilbert, gilbertd@indiana.edu

## Genome Grid Overview

Genome informatics *still* needs to provide bioscientists with:

(1) effective use of rapidly changing, growing complex genome data;
(2) rapid analyses of data from Next-Gen genome technology;
(3) frequent re-analyses encompassing expanded new evidence;
(4) easy access for many scientists to informatics methods

Clusters, grids and clouds of computers now provide infrastructure for these. Available genome processing tools can put these resources at the call of any scientist. The focus of this project is to enable a scientist or group with genome sequence and evidence (related proteins, ESTs, tile array expression, etc.) to analyze these with a range of available tools, without large infrastructure costs.

Computational engineering of this effort is on middleware for parallelizing genome data and collating results on grid systems. Genome informatics is known for its embarrassingly parallel, data-bound paradigms, where genome sequence and annotations can be subdivided as desired for many single-cpu analyses on cluster or grid systems. Parallel results are then re-assembled by genome locations. Large volume data now emerging from next-generation genomics technologies such as genome-wide tile expression arrays and short-read sequencing are incorporated in analyses in a similar data parallelization.

## Genome Grid Software

A basic component of this genome grid framework is genome partitioning and result assembly. It is based on the design and Perl scripts of EVidenceModeler [Haas et al 2008]. Analyses are run in parallel on many genome parts. Results are collated to full genome data sets. Many genome informatics tools work well on genome parts.

Components of this package now available and in use include genome partitioning, grid job submission for standard genome applications, and re-assembly of results in GFF and FastA formats. Applications used include Augustus+, NCBI Blast, EVidenceModeler, Exonerate, and SNAP. Planned enhancements include more genome applications, Ergatis workflow (J Orvis and colleagues), BioMart for genome data access, and PASA EST assembly.

"Instant" results web access is planned via a TeraGrid science gateway. This will provide genome results through BLAST sequence searches, GBrowse maps, and annotation text search/reports. This type of access helps immeasurably in assessing value, checking and correcting errors.

**Table 1.** TeraGrid usage steps.

| *Step* | *Notes* |
|---|---|
| *Preparation* | *One time* |
| 1. TeraGrid account | http://www.teragrid.org/userinfo/ |
| 2. Establish certificates | Grid-security; local workstation certificate |
| 3. Locate bio software | Install applications (tg_community/genomes) |
| *Analyze* | *Repeat Per analysis* |
| 1. Collect, partition data | Copy to shared disk; Partition & randomize |
| 2. Parallel run analyses | Run scripts, check errors, re-run as needed |
| 3. Collate results | Post-process to combine results from nodes |

Basic steps in Table 1 for using TeraGrid for genome analyses are not complicated, but require learning for a new user. Web documentation is sufficient for those with cluster or grid experience. Data selection, preparation, transport to TeraGrid, and return of results can be automated with data grid and workflow tools. A TeraGrid science gateway provides genome projects with ready infrastructure to rapidly produce single or multi-species analyses, with standard bioinformatics tools. Genome Grid applications are available at TeraGrid sites in TG_COMMUNITY/genomes/. Any US scientist can obtain a start-up allocation in two weeks to analyze genomes with these tools.

**Table 2.** Bug genomes analyzed on TeraGrid with genome grid methods.

| *Genomes* | *Notes* |
|---|---|
| *Daphnia* waterflea | Full genome assembly, analysis, annotation and tile array analyses. Plans to repeat for many species of Daphnia. |
| *Drosophila* fruitflies | 12 fruitfly species gene predictions, homology, plus one genome tile array analysis |
| *Acyrthosiphon* pea aphid | Full genome analysis and annotation; plan tile array analyses |
| *Nasonia* jewel wasp | Gene predictions; plan tile array analyses |
| *Ixodes* tick | Comparative arthropod analyses |

Dozens of bug, or Arthropod, genomes have been analyzed with these methods (Table 2), with a variety of interesting outcomes (see below). The most recent genome analyzed, pea aphid, included PASA EST assembly, Arthropod proteome mapping with BLAST and exonerate, gene prediction with Augustus and SNAP, and gene annotation with RefProt and UniProt databases. The first pass analysis took 12 days. First analyses and new data resulted in a refined set of analyses, a "complete" genome annotation, at the cost of approx. one informatician's half-time effort over six weeks. Results are at http://insects.eugenes.org/DroSpeGe/data/aphid/

## Interesting Genome Biology Results

*Daphnia* has lots of tandem genes, and gene finders make many mistakes with them. The same prediction errors occur in *Drosophila* and other genomes, but are less obvious.

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.

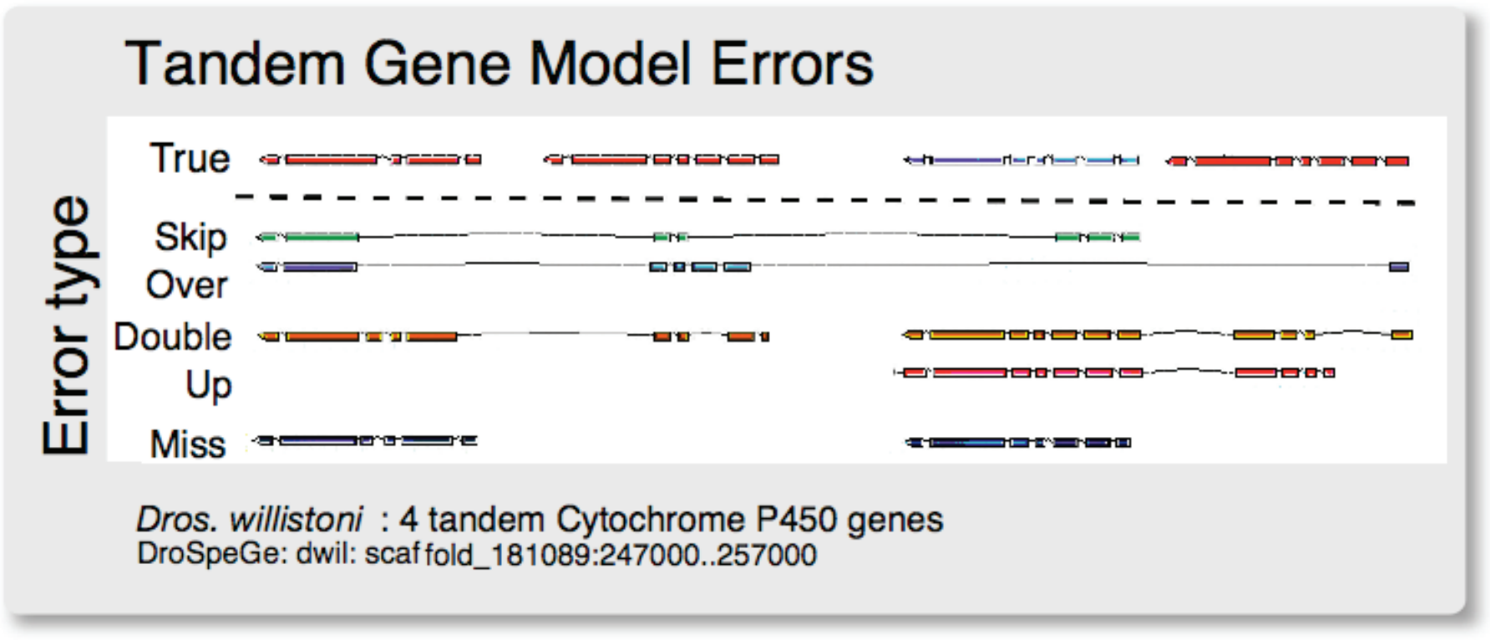
Tandem Gene Model Errors
Error type
True
Skip
Over
Double
Up
Miss
*Dros. willistoni* : 4 tandem Cytochrome P450 genes
DroSpeGe: dwil: scaffold_181089:247000..257000

**Figure 1** Tandem gene prediction errors.

### *Duplicate genes confuse predictors*

Tandem gene prediction error classes are diagrammed in Figure 1, 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 find mistakes*

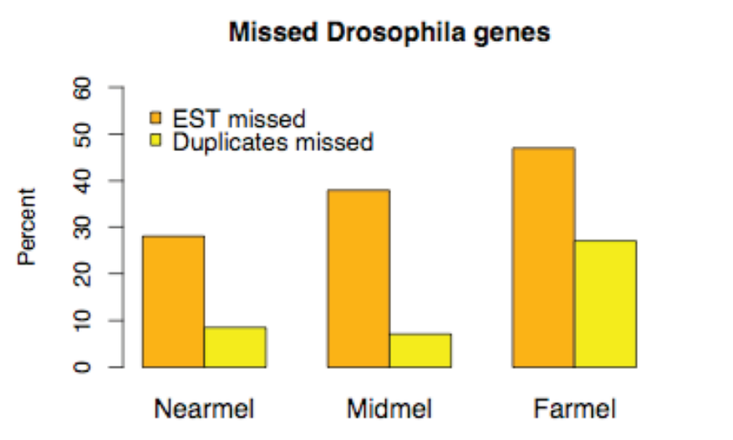

**Figure 2** Novel Drosophila species genes are missed by prediction

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. Figure 2 summarizes species group percentages for ESTs and duplicate genes that are missed by gene predictions. Most misses are those lacking Dmel homology.

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 3 (A). 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.

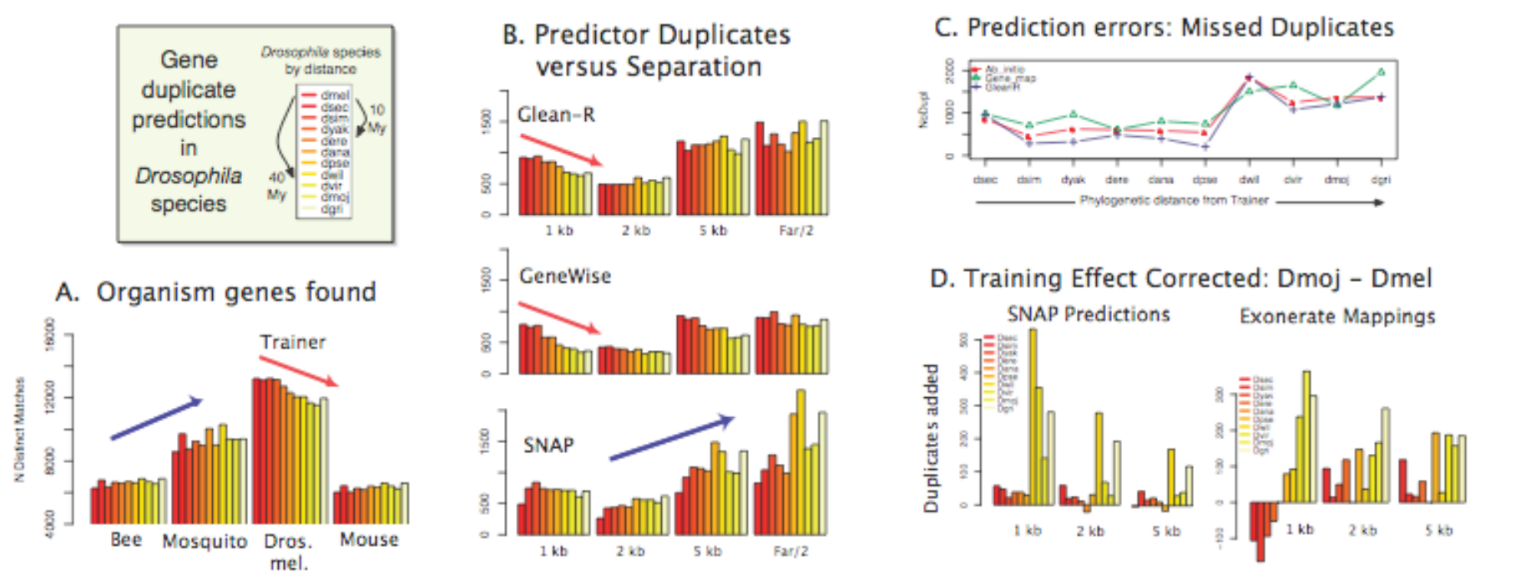

**Figure 3** Gene prediction species clines: biology or computational artifact?

In Figure 3B, two gene predictions show a lower rate of tandem genes predicted for the far-Dmel species. Other predictors show no cline, or a reverse cline comparable to that found for non-Dmel organism gene sets. The dilemma expressed in Figure 3B, of inconsistent predicted clines in duplicates, can be explained in large part by prediction errors, with results shown in Figure 3C. This species-bias error is eliminated by training the predictors with same or near-species gene data, as shown in Figure 3D for two gene calling methods (SNAP, Exonerate). The bar graphs of Figure 3 show gene counts for each of 10 species, arranged phylogenetically in heat colors from near-Dmel (red) to far-Dmel (yellow).

### *Genome tile expression finds novel genes*

Genome grid methods have turned genome tile array expression to gene predictions, for *Daphnia* and *Drosophila*, finding many new genes.

Gene calls made from tile expression experiments find 5,000 to 10,000 new genes above the 30,000 predicted for Daphnia. Figure 4 shows one such new gene and tile evidence. The analysis approach combines gene prediction software (Augustus) with tile transcription evidence, much like EST evidence. A similar amount of total new gene expression for *Dros. melanogaster* was found by Manak *et al* 2006.

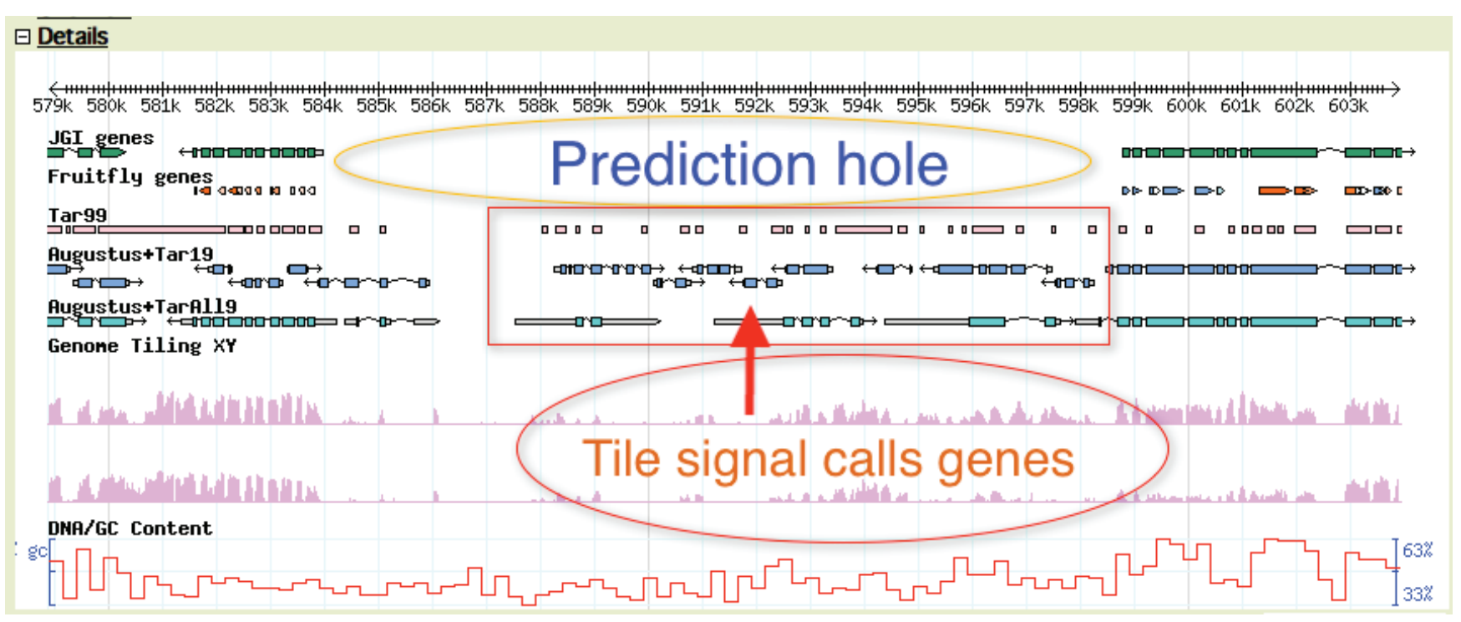

**Figure 4** Tile expression finds missing genes., Daphnia example

What does tile expression uncover? Among new tile expression genes, 10% have protein homology, and 19% have EST support (25% have one of these). This is a beginning to understand novel genes. But why have they been missed by current gene prediction? Using many treatment groups from cell lines and development stages, the modENCODE project seeks detailed answers. A set of *Drosophila melanogaster* gene predictions have been produced using Affymetrix tile expression data sets (modENCODE & Manak et al 2006) that include 33 treatment groups. There is high concordance (83%) between tile transcription fragments and predicted exons.

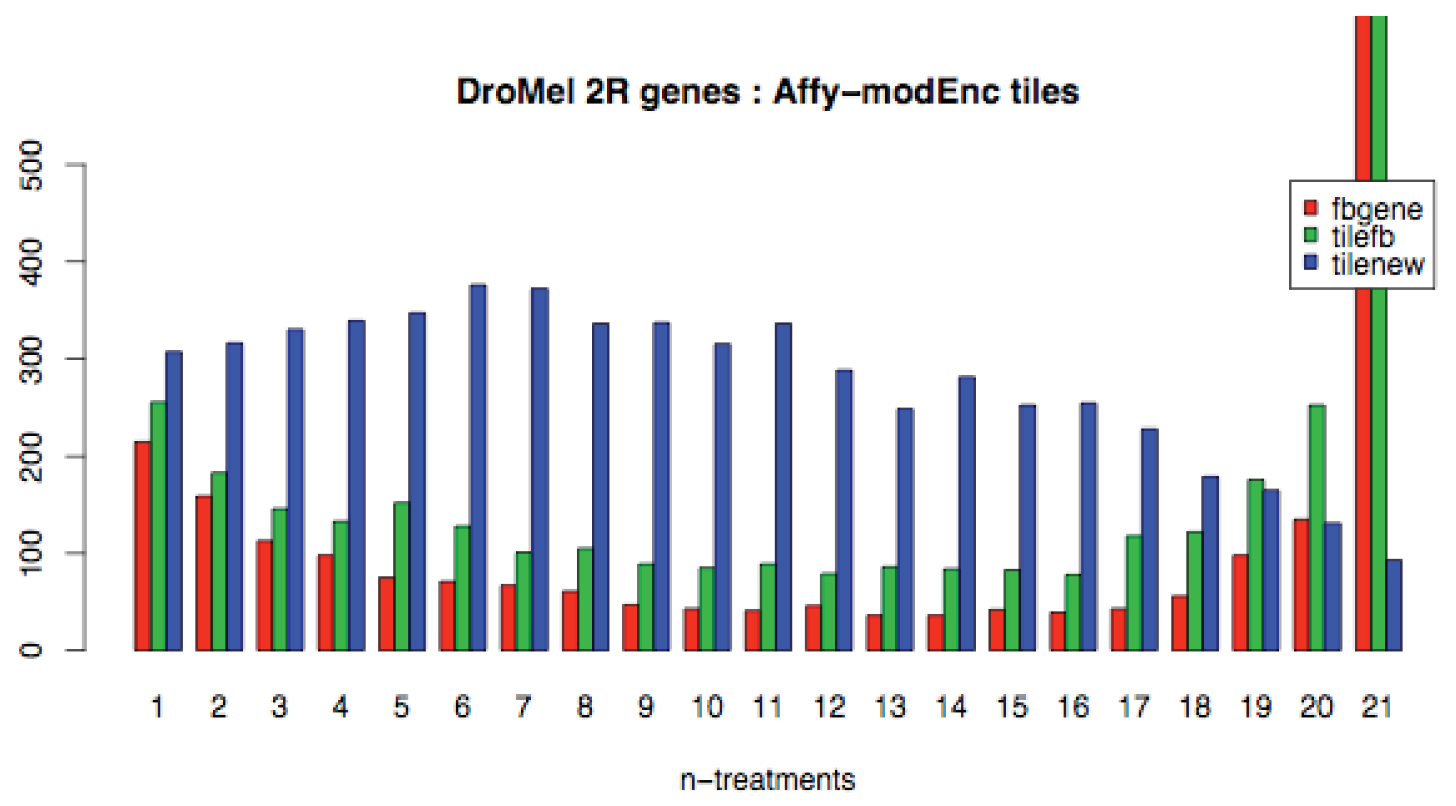

**Figure 5** Tile expression finds cell-line specific genes in *Dros. melanogaster*. New tile-predicted genes (tilenew, blue) are expressed in subsets of treatments (cell-lines), while most known genes (red, green) are expressed in all cell lines.

A key finding is that most known genes are expressed in all conditions (cell lines, development stages) while most newly predicted tile-genes are expressed in a subset of conditions (Figure 5). However, known genes that are expressed only in a subset of conditions are similar to new tile-predicted genes in their expression results. The expression score is lower for genes found only in a subset of conditions, whether known or new.

These tile results, along with tandem duplicate errors, point to better gene evidence and gene finder training to detect all genes. Computational tools that fully incorporate tile expression with gene finding are needed, and are an area of research by several groups.

### Summary of Genome Grid for Finding Genes

1. *Genome Grid Overview*
Clusters, Grids and Clouds: genome data parallelization is key to effective use.

2. *Genome Grid Software*
Partition genomes, run 100 parts, and collate results for several genome analyses.

3. *Interesting Genome Biology*
Gene finders miss tandem genes; genome tile expression finds many more genes.

4. *GMOD and TeraGrid are ready for your genomes.*
Please contact Don Gilbert for collaborations on genome analyses, use of genome grid.

## 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 (Illinois), PSC (Pennsylvania), TACC (Texas) and SDSC (California).

## References

Choi, J.H., Tang, H., Kim, S., Andrews, J., Gilbert, D., Colbourne, J. 2008. A machine learning approach to combined evidence validation of genome assemblies. **Bioinformatics**, 24(6):744-750

Clark, AG; Eisen, MB; Smith, DR; Bergman, CM *et al*, 2007. Evolution of genes and genomes on the Drosophila phylogeny, **Nature**, p. 203, vol. 450, doi:10.1038/nature0634

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

Gilbert, D.G. 2008. Tile array expression finds many new genes. http://insects.eugenes.org/DroSpeGe/data/dmel5/modencode/

Gilbert, D.G. 2008. Genome Grid for genome informatics. http://eugenes.org/gmod/genogrid/

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/

Haas, BJ, SL Salzberg, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology, 9:R7 (doi:10.1186/gb-2008-9-1-r7)

Haas, B.J. and J. R. Wortman, 2007. Eukaryotic Gene Structure Annotation, in preparation.

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