DroSpeGe: Rapid access database for new *Drosophila* species genomes

Donald G. Gilbert

Biology Department, Indiana University, Bloomington, Indiana 47405 USA; email gilbertd@indiana.edu

ABSTRACT

The *Drosophila* species comparative genome database DroSpeGe (http://insects.eugenes.org/-DroSpeGe/) provides genome researchers with rapid, usable access to twelve new and old Drosophila genomes, since its inception in 2004. Scientists can use, with minimal computing expertise, the wealth of new genome information for developing new insights into insect evolution. New genome assemblies provided by several sequencing centers have been annotated with known model organism gene homologies and gene predictions to provided basic comparative data. TeraGrid supplies the shared cyberinfrastructure for the primary computations. This genome database includes homologies to *D. melanogaster* and eight other eukaryote model genomes, and gene predictions from several groups. BLAST searches of the newest assemblies are integrated with genome maps. GBrowse maps provide detailed views of cross-species aligned genomes. BioMart provides for data mining of annotations and sequences. Common chromosome maps identify major synteny among species. Potential gain and loss of genes is suggested by Gene Ontology groupings for genes of the new species. Summaries of essential genome statistics include sizes, genes found and predicted, homology among genomes, phylogenetic trees of species, and comparisons of several gene predictions for sensitivity and specificity in finding new and known genes.

Key words: comparative genomics, *Drosophila* species, genome maps, generic model organism database

INTRODUCTION

Many new genomes are becoming available this decade. Current contents of public genome archives exceed 1 billion sequence traces from over 1,000 organisms (1). This number will increase rapidly as costs drop and scientific uses for comparing many genomes increases (2). Biologists should have rapid access to these new genomes, including basic annotations from well-studied model organisms and predictions to locate potential new genes, to make sense of them. Genome annotation and database management can be streamlined now using generic tools, shared computing resources and common genome database techniques to provide useful access to biologists in weeks instead of several months.

New genome sequencing projects and communities are facing large informatics tasks for incorporating, curating and annotating, and disseminating sequence and annotation data. Effective genome studies need an informatics infrastructure that moves beyond individual organism projects to a cost effective use of common tools. Expertise from existing genome projects should be leveraged into building such tools. The Generic Model Organism Database (GMOD; 3) project has this goal, to fully develop and extend a genome database tool set to the level of quality needed to create and maintain new genome databases. GMOD and related genome database tools now support a portion of the basic tasks for such. Two needs in development for GMOD are the creation of new databases for emerging model organisms, and tools for comparative genome databases that integrate data from many sources.

A common, ongoing task for research that uses genome databases is to compare an organism's genome and proteome with related organisms, and other sequence data sets (ESTs, SNPs, transposable elements). This task requires significant computational infrastructure, one where reusable tools, protocols and resources will be valuable and significantly reduce duplicative infrastructure and maintenance effort. Software tools to fully assembly, analyze and compare these genomes are available to bioscientists. The ability to employ these tools on genome data sets is limited to those with extensive computational resources and engineering talent. Effective use of shared cyberinfrastructure in bioinformatics is a problem today. Cluster and Grid computing in bioinformatics have followed other disciplines in parallelizing applications, but this is costly and limited to a subset of bioinformatics applications. This database enables bioscientists to have usable access to new genomes shortly after sequencing centers make them available, facilitating new science discoveries and understanding of the evolution, comparative biology, and genomics of these model organisms.

GENOME INFORMATICS METHODS

Common Components

DroSpeGe has been built with common GMOD database components and open source software shared with other genome databases. Use of common components facilitates rapid construction and interoperability. The GMOD ARGOS replicable genome database template (www.gmod.org/argos/) provides a tested set of integrated components. The genome access tools of GMOD GBrowse (3), BioMart (4) and BLAST (5) are available for the *Drosophila* species genomes. The GMOD Chado relational database schema (www.gmod.org/chado/) is used for managing an extensible range of genome information. Middleware in Perl and Java are added to bring together BLAST, BioMart, sequence reports, searches and other bioinformatics programs for public access. Another aid to integrating and mining these data is GMOD Lucegene (www.gmod.org/lucegene/), that forms a core component for rapid data retrieval by attributes, GBrowse data retrieval, and databank partitioning for Grid analyses. DroSpeGe operates on several Unix computers; the primary server is a SunFire V20z from Sun Microsystems. Genome maps include *D. melanogaster* DNA and protein homology, homologies to nine eukaryote proteomes, marker gene locations, gene predictions using 15 methods produced by several contributing groups. The assemblies and predicted genes can be BLASTed, with links to genome maps. BioMart provides searches of the full genome annotation sets, allowing selections of genome regions with and without specific features.

New Species Genomes

Twelve *Drosophila* genomes, ten recently sequenced, contain over two billion nucleotides, with sizes ranging from a small 133 Megabases of *D. melanogaster* to over 230 Mb in *D. willistoni* (Table 1). The model organism *D. melanogaster* is approaching its 5th major assembly release, and continues to see significant improvements in genes and genome features. It has a known, located complement of about 14,000 protein genes. One main impetus for undertaking the sequencing of 11 additional related species is to improve via comparative analyses the knowledge of this major research organism. *D. pseudoobscura*, the second related genome, is in its second major release. The additional species are at their first major assembly stage, requiring automated annotation, quality assessment and cross-species comparisons. Four of these new genomes

(Dsim, Dsec, Dyak and Dere) are close relatives of the model in the melanogaster subgroup. The remainder range through five other taxonomic groups with an estimated divergence time of 40 million years, with the cactus breeder *D. mojavensis*, widely distributed *D. virilis*, and Hawaiian picture-wing *D. grimshawi* most distant from Dmel. Assembly sequences of the *Drosophila* species comparative annotation freeze 1 (CAF1) are distributed at http://rana.lbl.gov/drosophila/-caf1.html, as listed in Table 1. These form the primary source data for this database. Over the course of two years, this database has provided rapid access to several assembly releases per species, including annotation, searching and viewing services for each release.

Table 1. Drosophila species genomes, abbreviation, sequencing centers and genome size of CAF1 assemblies used at DroSpeGe.

Abbr.	Species	Size (Mb)	Sequencing Center
Dmel	*Drosophila melanogaster*	133	Berkeley Drosophila Genome Project / Celera
Dsim	*Drosophila simulans*	142	Genome Sequencing Center, Washington University
Dsec	*Drosophila sechellia*	167	Broad Institute
Dyak	*Drosophila yakuba*	160	Genome Sequencing Center, Washington University
Dere	*Drosophila erecta*	153	Agencourt Bioscience Corporation
Dana	*Drosophila ananassae*	231	Agencourt Bioscience Corporation
Dper	*Drosophila persimilis*	188	Broad Institute
Dpse	*Drosophila pseudoobscura*	153	Human Genome Sequencing Center, Baylor College of Medicine
Dwil	*Drosophila willistoni*	237	J. Craig Venter Institute
Dmoj	*Drosophila mojavensis*	194	Agencourt Bioscience Corporation
Dvir	*Drosophila virilis*	206	Agencourt Bioscience Corporation
Dgri	*Drosophila grimshawi*	200	Agencourt Bioscience Corporation

Annotations produced by several groups collaboratively are provided for map viewing and data mining. Protein coding gene predictions viewable at this resource include contributions listed in Table 2.

Table 2. Drosophila species genome annotations (partial list) included at DroSpeGe, contributed at http://rana.lbl.gov/drosophila/wiki/index.php/Annotation_Submission

Contributor	Annotation Description
S. Batzoglou Lab, Stanford	Contrast (6) predictions
M. Brent Lab, Washington Univ., St. Louis	N-SCAN (7) predictions with melanogaster alignments
D. Gilbert Lab, Indiana University	SNAP (8) predictions, model organism gene homologies
M. Eisen Lab, UC Berkeley/LBNL	GeneWise (9), GeneMapper (10), Exonerate (11) annotations
R. Guigó Genome Bioinformatics Lab, Barcelona	Geneid (12) predictions
NCBI, Bethesda	Gnomon (13) predictions
B. Oliver Lab, LCDB, NIDDK, NIH	Gene expression evidence from microarray (14)
L. Pachter Lab, UC Berkeley	GeneMapper (10) annotations
C. Ponting Lab, MRC FGU Oxford	Gene prediction pipeline (15) with Exonerate

TeraGrid Genome Analyses

The TeraGrid project (www.teragrid.org) is part of a shared cyberinfrastructure for sciences, funded primarily by NSF. TeraGrid provides collaborative, cost-effective scientific computing infrastructure much in the same way the GMOD initiative is building common tools for genome databases. The TeraGrid system is particularly suitable for genome assembly, annotation, gene finding and phylogenetic analyses. TeraGrid computers have been employed to analyze the twelve *Drosophila* genomes, providing the major contents of DroSpeGe database. This has enabled rapid analyses without the expense of obtaining and maintaining a local compute cluster. This experience forms a basis for other genome projects to use TeraGrid. Scripts used for this analysis are available at the GMOD repository (http://gmod.cvs.sourceforge.net/gmod/genogrid/). Genome database tools from GMOD project are used to organize the computations for public access. Results include *D. melanogaster* genome homology, homologies to nine eukaryote proteomes, gene predictions, marker gene locations, and *Drosophila* microsatellites. For each of twelve *Drosophila* genomes, a comparison is made to a set of nine proteomes, with 217,000 proteins, drawn from source genome databases, Ensembl and NCBI. The reference proteomes are human, mouse, zebrafish, fruitfly (Dmel), mosquito, bee, worm (*C. elegans*), mustard weed (*A. thaliana*), yeast. Sizes of the new genomes are in the 150 Mb to 250 Megabase range. Protein-genome DNA alignment is done using tBLASTn, with a Grid-aware version of NCBI software. The TeraGrid run for each genome took 12 - 18 hours using 64 processors. Whole genome DNA-DNA alignments were performed for a subset of new genomes. Gene predictions with SNAP (8) have been generated. Over the course of 6 months, with 2 to 3 genome assembly updates each per species, and error corrections, the total TeraGrid 64-cpu usage per genome has been approximately 4 days, excluding queue-waiting times.

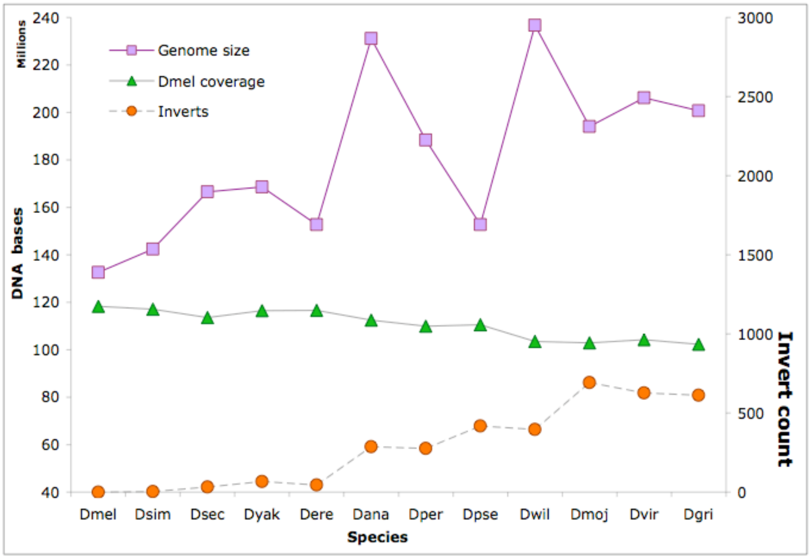


Figure 1. *Drosophila* species assemblies, showing assembly sizes, and coverage of these by *D. melanogaster* genome DNA (top and middle lines, in megabases, left ordinate), and counts of chromosome segments inverted relative to Dmel (bottom line, right ordinate). Species on abscissa are taxonomically ordered with Dgri most distant from Dmel. This is summarized from DroSpeGe/news/genome-summaries/dnacoverage.html.

DATABASE USES

DroSpeGe provides a resource to biologists interested in comparing species differences and similarities, including novel and known genes, genome structure and evolution, gene function associations. Known genes from model organisms are found in the new genomes at expected rates, allowing for variations due to assembly quality. The most divergent species (Dmoj, Dvir, Dgri), with 40 million years divergence from Dmel, match approximately 90% of the model species genes. These known genes provide useful access to the new genomes for many researchers interested in locating a particular gene or gene family. The known gene matches also offer searches and cataloging gene contents by known functions. Figure 1 shows the size and similarity to the Dmel model of these genomes. Genome annotations and analyses produced for this database is available at DroSpeGe/data/ and in bulk form at ftp://eugenes.org/eugenes/-genomes/, including annotations of CAF1 and prior assembly releases. Related genome projects also provide *Drosophila* genome data and complementary services (see Related work).

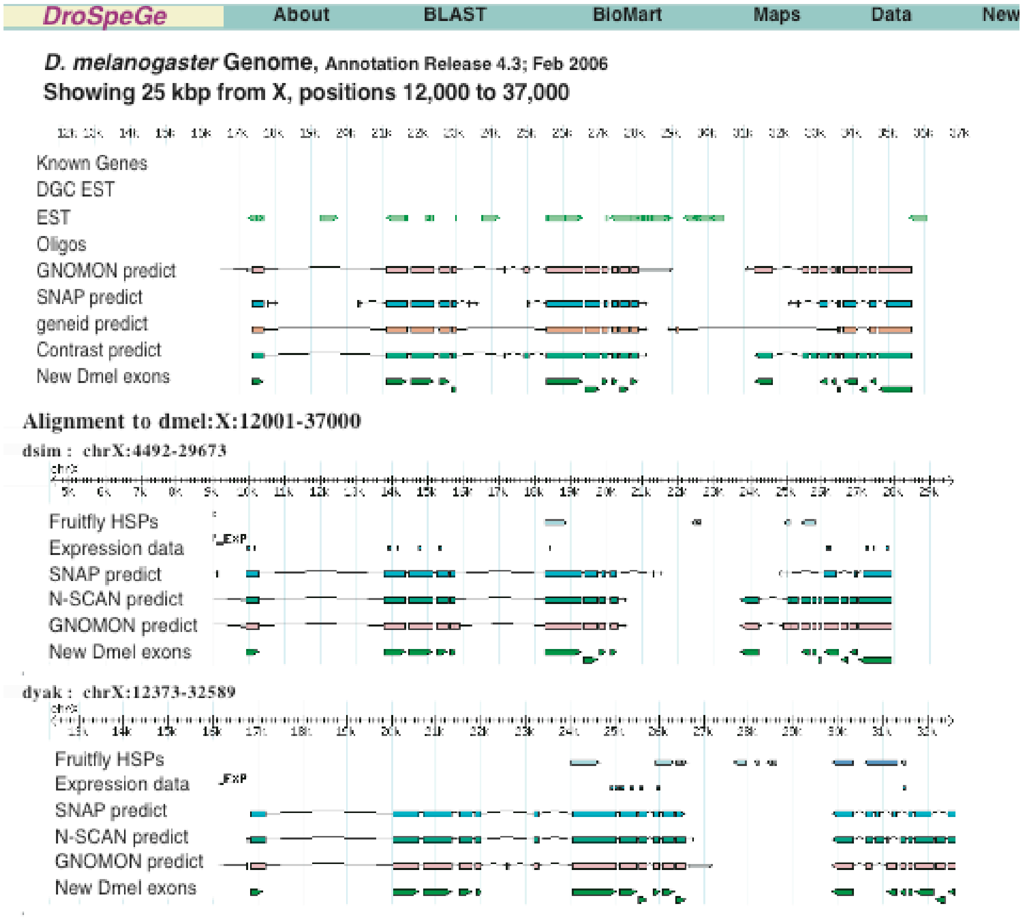
Genome Data Mining

An emerging trend among bioscientists and bioinformaticians is to use data mining of large subsets of genome data, often focused on summary information for a range of common attributes. These data are used in spreadsheets and simple databases or analyses. Genomics web databases often lack methods for effectively mining large subsets of genomes, or are limited in the questions one can pose to the underlying complex data (16). The Ensembl project with its off-shoot BioMart (4), is an example of integrated software and data that bridges the gap in biology data access between bulk files and web portals. A tool for creating BioMart-compliant transaction databases, *gff2biomart,* is a recent addition by the author to GMOD tools collection (http://gmod.cvs.sourceforge.net/gmod/schema/GMODTools/bin/). It has been used for DroSpeGe and other genome data sets. BioMart with annotations of twelve Drosophila genomes has provided numerous bioscientists with a unique data mining access to these new genomes.

With BioMart, one can select genome regions with the available annotations, and exclude others, and download tables or sequences of the selection set. For instance, select the regions with Mosquito gene homologs, but lacking *D. melanogaster* homologs. Or select regions with gene predictions but no known homology. A major reason to undertake the genome sequencing of twelve Drosophila species is to improve genome knowledge of the widely used *D. melanogaster* model organism. A significant application for BioMart has been to identify gene predictions in *D. melanogaster* that do not match known genes. Further phylogenetic analysis of these new gene predictions has identified a subset with cross-species homology and high synonymous substitution rates, validating these as likely new genes and coding exons with phylogenetic evidence. Another application of BioMart has been to compare the qualities of gene predictor methods, identifying predicted exons that coincide with known gene homology, and with gene expression data sets to measure sensitivity at predicting new and known genes.

Genome Maps

Maps of the twelve genomes form the core, with BLAST searches, of discovery tools for bioscientists. Maps including all available annotations from several groups are provided using GBrowse (3). The BLAST result reports include hyperlinks from each alignment match to the respective genome map, as well as to sequence and GFF annotation results. As species comparisons are of much interest, BLAST results also link to a comparative map display of the matches. A recent addition to the genome maps is an aligned comparative map set for any group of the twelve species. As seen in Figure 2, this allows one to view phylogenetic evidence of common gene predictions and features in homologous regions. In this example, genes that are predicted in the model Dmel, but previously not located, are found to be orthologously located across eight species (Dmel through Dmoj). This capability of full comparative annotation maps may be unique to DroSpeGe. Other genome maps offer either a single species map with tracks that summarize homology, or a syntenic view of two species. An overview of all species chromosome maps is provided in the DroSpeGe/maps/ section. These overviews link to detailed genome maps, with known gene homology locations, gene expression evidence, and predictions.



D. melanogaster Genome, Annotation Release 4.3; Feb 2006

Showing 25 kbp from X, positions 12,000 to 37,000

Alignment to dmel:X:12001-37000

dsim : chrX:4492-29673

dyak : chrX:12373-32589

Figure 2. Aligned genomes view of new *D. melanogaster* gene locations on X chromosome, on *D. melanogaster, D. simulans and D. yakuba*, identified with cross-species comparison of coding exons, from DroSpeGe/data/dmel-dspp/newgenes. Several gene predictors match these common coding exons. Additional evidence from EST, protein HSP matches, and gene expression data corroborate the new genes. Genomes with orthologous gene predictions not shown, but viewable at DroSpeGe maps, include Dsec, Dere, Dana, Dpse, and Dmoj.

Common Chromosomes in Drosophila

A series of maps show large scale synteny between genome assembly units (scaffolds or chromosomes), as determined from genome x genome DNA BLAST matches, identified as common Muller elements. These are found in DroSpeGe/maps/muller-elements. Muller's elements are the names A, B, C, D, E, F for six chromosome arms common among Drosophila as coined by Hermann Müller. Chromosome names and centromeric joins differ among the species; Muller elements identify the common units. These synteny maps provide scientists with quick access to common genome regions among the twelve species. The melanogaster group species (Dmel, Dsim, Dsec, Dyak, and Dere) have close matching, with large-scale inversions evident in these maps. Among the more distantly related species, the new genome assembly of *D. mojavensis* has proved most complete, with four Muller elements nearly fully assembled, the autosomes B to E, and the sex chromosome assembled into four major scaffolds.

Gene Variation by Gene Ontology Group

To provide an assessment of possible gene gain and loss among *Drosophila*, gene matches to Gene Ontology categories by species were tabulated, and provided at section DroSpeGe/news/-genome-summaries/gene-GO-function-association. These may indicate species differences in functional categories. Statistically significant deviations are indicated. While low counts, suggestive missing genes, may be due to divergence of genes, extra gene matches more strongly suggest categories where species differ. Among the interesting differences, transport genes (GO:0006810) may show a phylogenetic cline with more in the non-melanogaster group (Dana to Dgri); protein binding genes (GO:00055515) may be more common in the Dmel-Dsim-Dsec siblings; protein biosynthesis (GO:0006412) is higher in the Dpse-Dper sibling species. Individual species peaks such as Dwil for catalytic activity genes (GO:0003824) or signal transduction (GO:0007165) in Dgri, suggest species-specific adaptations. The gene matches are high-scoring segment pair (HSP) groupings, and include various events: gene duplications, alternate splice exons within genes, new genes that appear composed of exons from other genes, as well as computational artifacts. Detailed evidence pages provide links to GBrowse genome map views showing all secondary HSPs. Proteome sources in this analysis are those organism with extensive GO annotations: Dmel fruitfly, mouse, C. elegans worm, and yeast. GO-Slim groupings are used for Biological Process, Molecular Function, Cell Location (125 categories). A table provides the correspondence between MOD gene ID, GO primary ID, and GO-slim groupings. Chris Mungall's GO map2slim software is employed for this, along with current GO gene associations.

Related work

Drosophila species assemblies, analyses and annotations have been coordinated at Michael Eisen's community Wiki (rana.lbl.gov/drosophila/wiki/), in an open way that serves as a model for future genome collaborations. Contributors have here submitted data, genome analyses, summaries and discussion for the benefit of the research community. In conjunction with this, the Eisen lab provides annotations, analyses and GBrowse maps of *Drosophila* species. The FlyBase project (www.flybase.org) has benefited from these community efforts, recently adding a subset of annotations for twelve species to its map and search services. The most comparable effort to DroSpeGe in approach, if not species, is the Fungal Comparative Genomics resource (fungal.genome.duke.edu; 17) that catalogs 56 genomes including the model *S. cerevisiae* and related yeasts and fungi. Fungal Genomics offers Gbrowse maps, BLAST, gene predictions and phylogenetic comparisons. Comprehensive genome resources with *Drosophila* include UCSC Genome Bioinformatics (genome.ucsc.edu), Ensembl (www.ensembl.org), Entrez Genomes (www.ncbi.nih.gov). The later two currently show only *D. melanogaster*. UCSC Genomes has most of the insect genomes, but as of 2006-10 has yet to update to current *Drosophila* assemblies and annotations. UCSC provides a very useful set of comparative homology analyses for each species genome.

The DroSpeGe comparative genome database has provided bioscientists with rapid access to new genomes in a usable way, with new annotations, browsing, search and summary services not available elsewhere. Future plans for this database focus on enhancing genome comparison functions, with improvements to category overviews for gene functions, pathways and orthology evidence. Additional insect genomes and the arthropod *Daphnia pulex* (18) may be integrated to extend the comparative range.

ACKNOWLEDGEMENTS

This project is supported in part by grants from the National Human Genome Research Institute of the National Institutes of Health, and from Sun Microsystems, to D. Gilbert. National Science Foundation TeraGrid access grant provided computing resources. Reviewers who suggested improvements to this paper are thanked for their efforts.

REFERENCES

1. Benson, D. and Wheeler, D. (eds) 2006 Trace Archives at 1 Billion. **NCBI News** 15(1). NIH Publication No. 06-3272 URL: http://www.ncbi.nlm.nih.gov/Web/Newsltr/V15N1/trace.html
2. Siepel, A, G Bejerano, J S. Pedersen, A S. Hinrichs, M Hou, K Rosenbloom, H Clawson, J Spieth, L W. Hillier, S Richards, G M. Weinstock, R K. Wilson, R A. Gibbs, W. J Kent, W Miller, and D Haussler, 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. **Genome Research**, 15, 1034-1050.
3. Stein L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., and Lewis, S. 2002. The generic genome browser: a building block for a model organism system database. **Genome Research** 12: 1599-610. URL: www.gmod.org/gbrowse/
4. Kasprzyk, A, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P Rocca-Serra, T. Cox and E. Birney (2004) EnsMart: A Generic System for Fast and Flexible Access to Biological Data. **Genome Research**, 14(1):160-169. URL: www.biomart.org
5. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research** 25:3389-3402
6. Gross SS, Do CB, Batzoglou S. CONTRAST: de novo gene prediction using a semi-Markov conditional random field. In BCATS 2005 Symposium Proceedings, p. 82, 2005. URL: http://contra.stanford.edu/contrast/
7. Wu J.Q., Shteynberg D., Arumugam M., Gibbs R.A., and Brent, M.R. (2004) Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. Genome Research 14: 665-671. URL: http://ardor.wustl.edu/ with N-SCAN update.
8. Korf, Ian 2004. Gene finding in novel genomes. **BMC Bioinformatics** 2004, 5:59 URL: http://www.biomedcentral.com/1471-2105/5/59
9. Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and Genomewise. Genome Research 14:988-995, doi:10.1101/gr.1865504
10. Chatterji, S. and Pachter, L. (2006) Reference based annotation with GeneMapper, Genome Biology, 7:R29, doi:10.1186/gb-2006-7-4-r29 URL: http://bio.math.berkeley.edu/genemapper/
11. Slater, G. St. C. and Ewan Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 2005, 6:31 doi:10.1186/1471-2105-6-31
12. Parra, G., Blanco, E. and Guigó,R. (2000) Geneid in Drosophila. Genome Research 10(4):511-515. URL: http://genome.imim.es/software/geneid/index.html
13. Souvorov, A., Hlavina, W., Kapustin, Y., Kiryutin, B., Kitts, P., Pruitt, K., Sapojnikov, V. and Ostell, J. (2006) Gnomon annotation of Drosophila species genomes. URL: ftp://ftp.ncbi.nih.gov/genomes/Drosophila_melanogaster/special_requests/CAF1/

14. Sturgill, D., Zhang, Y., Parisi, M., and Oliver, B. (2006) Drosophila species expression arrays, preliminary results. Laboratory of Cellular and Developmental Biology, NIDDK, NIH. URL: http://intramural.niddk.nih.gov/research/nimble/nimblefly.htm
15. Heger, A. and Ponting, C. (2006) Drosophila gene prediction pipeline with Exonerate. URL: http://wwwfgu.anat.ox.ac.uk:8080/flies/documentation.html
16. Stein, L. 2003. Integrating Biological Databases. *Nature Reviews Genetics.* 4: 337-345. doi:10.1038/nrg1065
17. Stajich, J.E. 2006. A comparative genomic investigation of fungal genome evolution. Ph.D. Dissertation. Graduate School of Duke University. 156 pp. URL: http://www.duke.edu/~jes12/thesis/
18. Colbourne, J.K., Singan, V.R., Gilbert, D.G. 2005. wFleaBase: the Daphnia genome database, **BMC Bioinformatics**, 6:45 doi:10.1186/1471-2105-6-45 URL: wfleabase.org