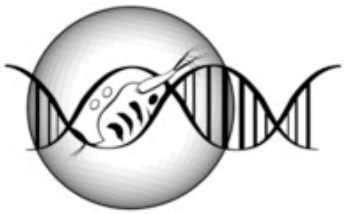
TeraGrid for Genome Analyses

Indy Bioinfo, May 2006

Don Gilbert, gilbertd@indiana.edu



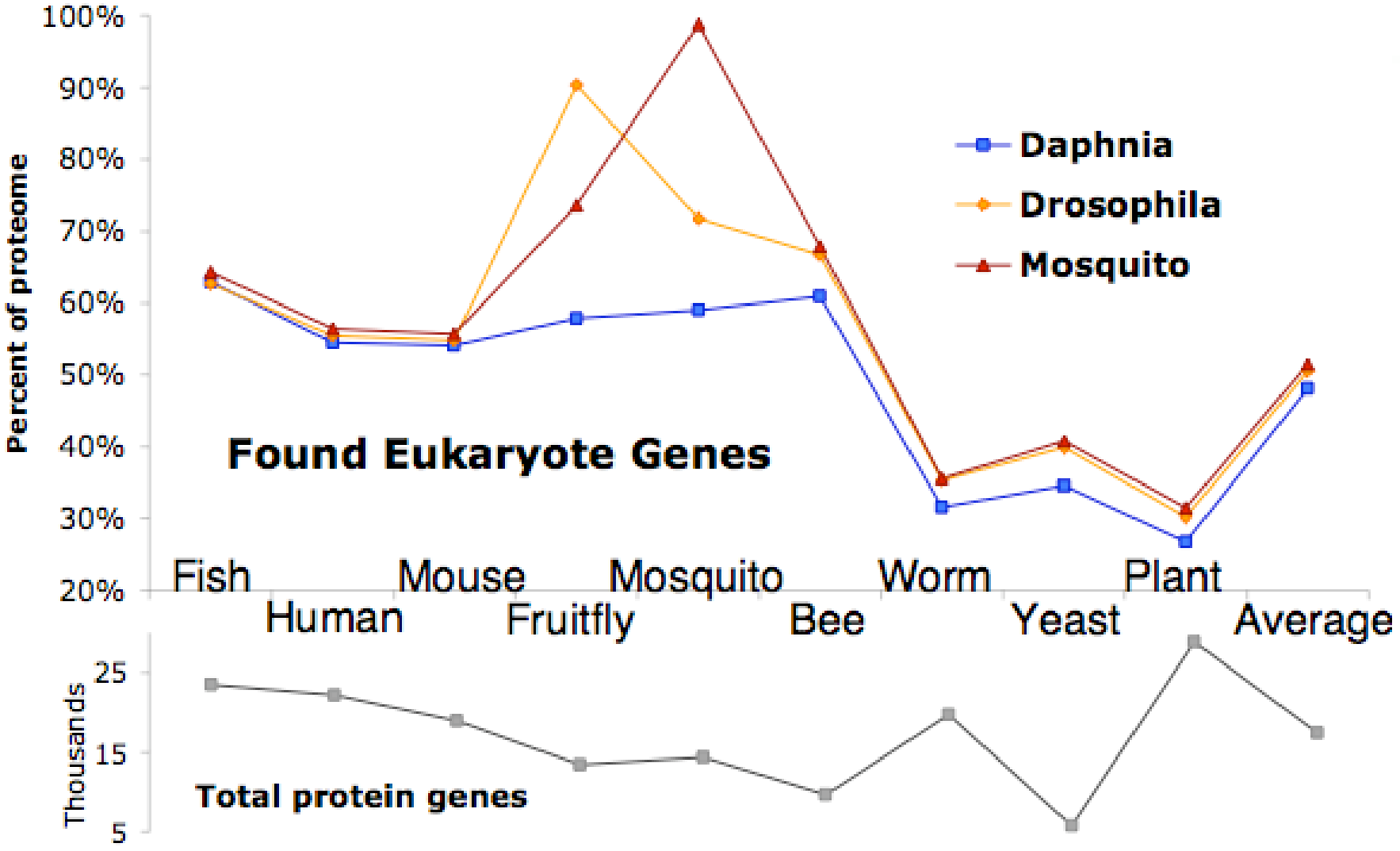
Summary

- **PROBLEM** in bioinformatics: enabling use of large biology data analyses on shared cyberinfrastructure.
- **SOLUTION**: Parallelize data access rather than applications for effective Grid use of existing and new biology analyses.
- **RESULTS**: New insect and crustacean genomes have been analyzed on TeraGrid to assess data grid methods in genome informatics. Rapid Grid analyses have facilitated rapid biology discoveries in these genomes.

New Fly, wFlea genomes

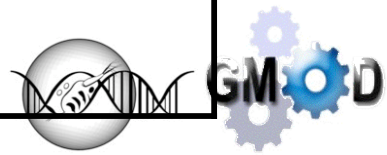
- **Biologists Need rapid access**: to new genomes for *Daphnia pulex* and twelve *Drosophila*
- **Find the Genes**: Compare to 9 proteomes: fly, worm, mouse, yeast, human, …
- **Generic Model Organism Database** (GMOD) tools organize TeraGrid results for public :
  - genome maps (GBrowse), web BLAST, data mining (BioMart), genome summaries
  - wfleabase.org (Daphnia), insects.euGenes.org (Drosophila)

Proteome Annotations



TeraGrid usage steps

Step	***Notes***
Preparation	*One time*
1. Obtain TeraGrid account	Via web http://www.teragrid.org/userinfo/
2. Establish certificates	Grid-security entries; test proxy; local workstation certificate
3. Locate biology software	Find and compile parallel applications
Processing	*Per analysis*
4. Locate and prepare data	partition, shred & randomize
5. Transfer data to TeraGrid	FTP, secure-shell, other
6. Configure and run analysis	Globus run scripts, attention to errors, queuing
7. Return and collate results	Post-process to combine results from nodes; e.g. to-GFF for map view of genome blast.

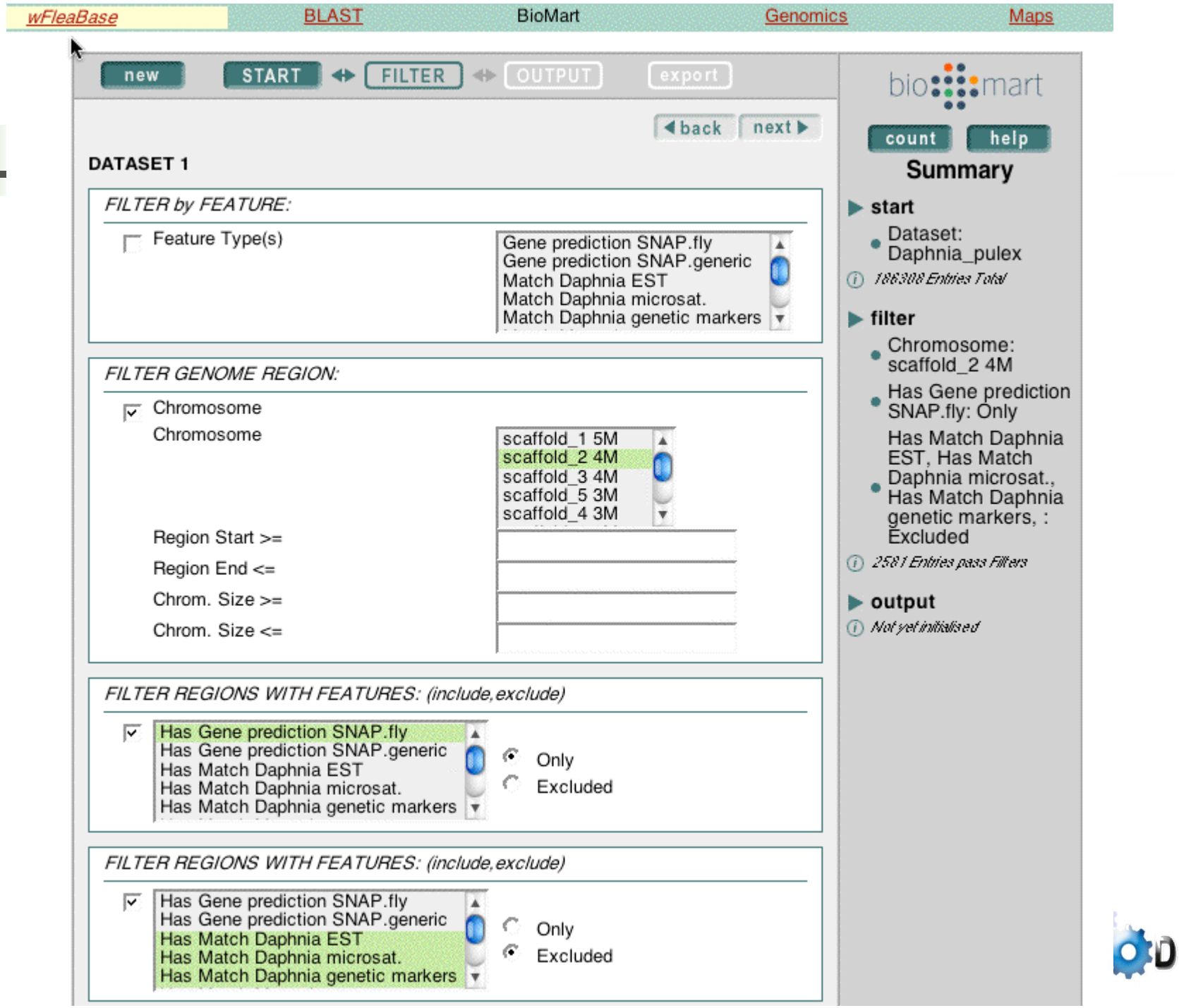


Data grid methods

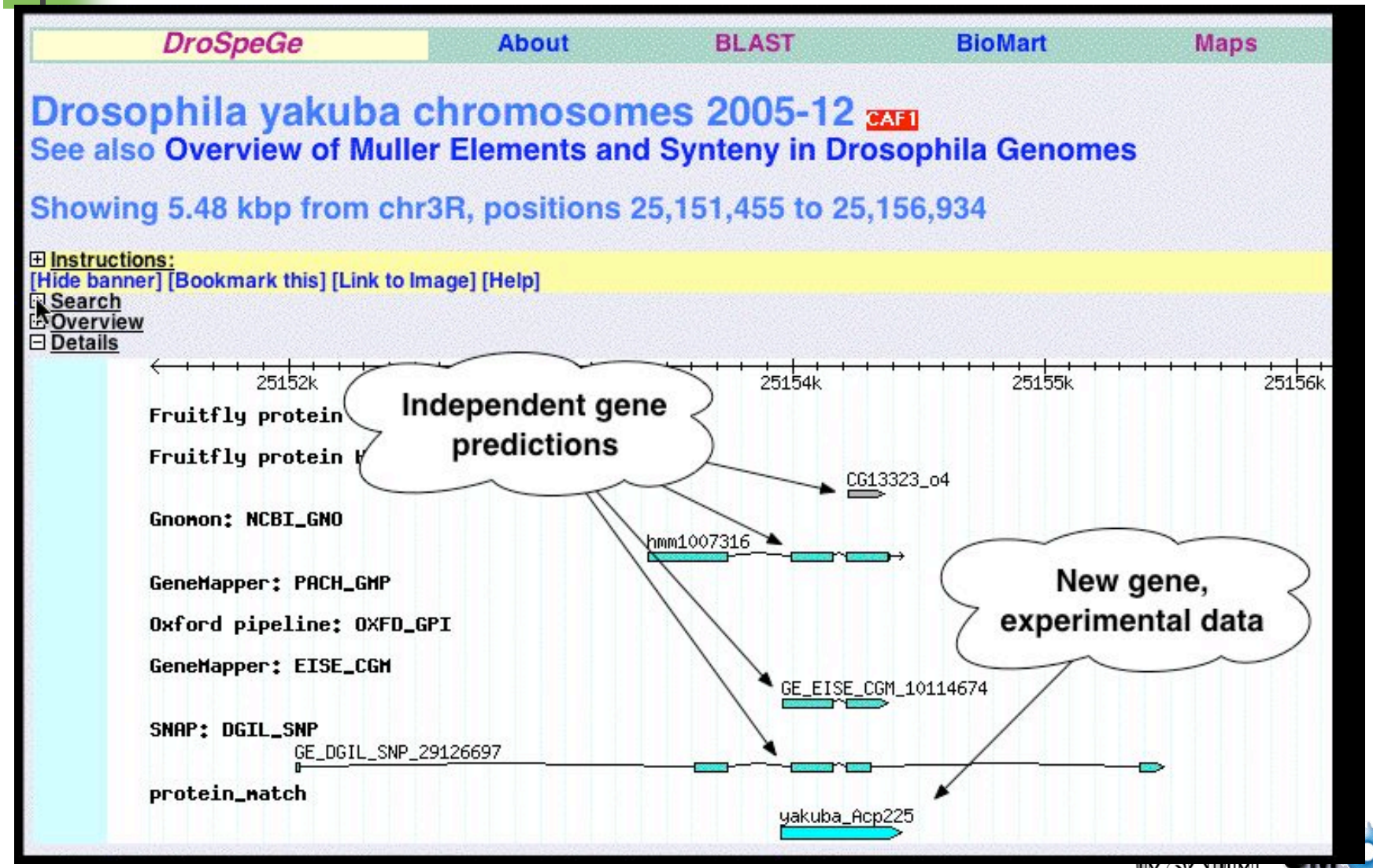
1. @virtualdata= biodirectory("find protein coding sequences for *Drosophila* species"),
2. @realdata= biodirectory("get locators for @virtualdata split *n* ways"), for *n* compute nodes
3. for i (1.. *n*) { copy(realdata[i], gridcpu[i]); results[i]= runapp(gridcpu[i]) }
4. result_table = collate(@results);

These steps will work for gene finders, homology comparison, multiple alignment tools, and phylogenetic comparison.

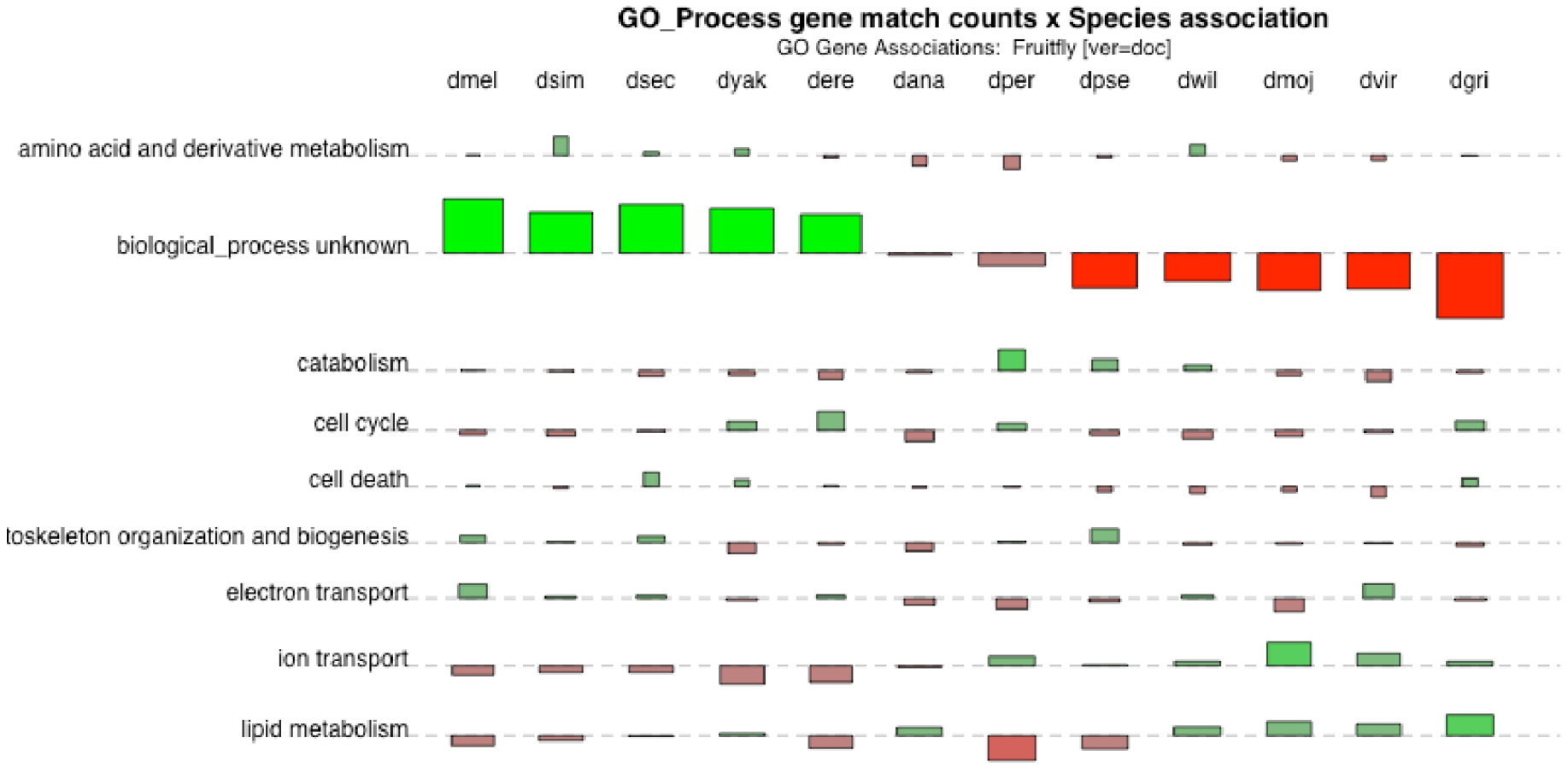
BioMart Filter



New gene evidence



Possible gene gain/loss



Thanks to these folks

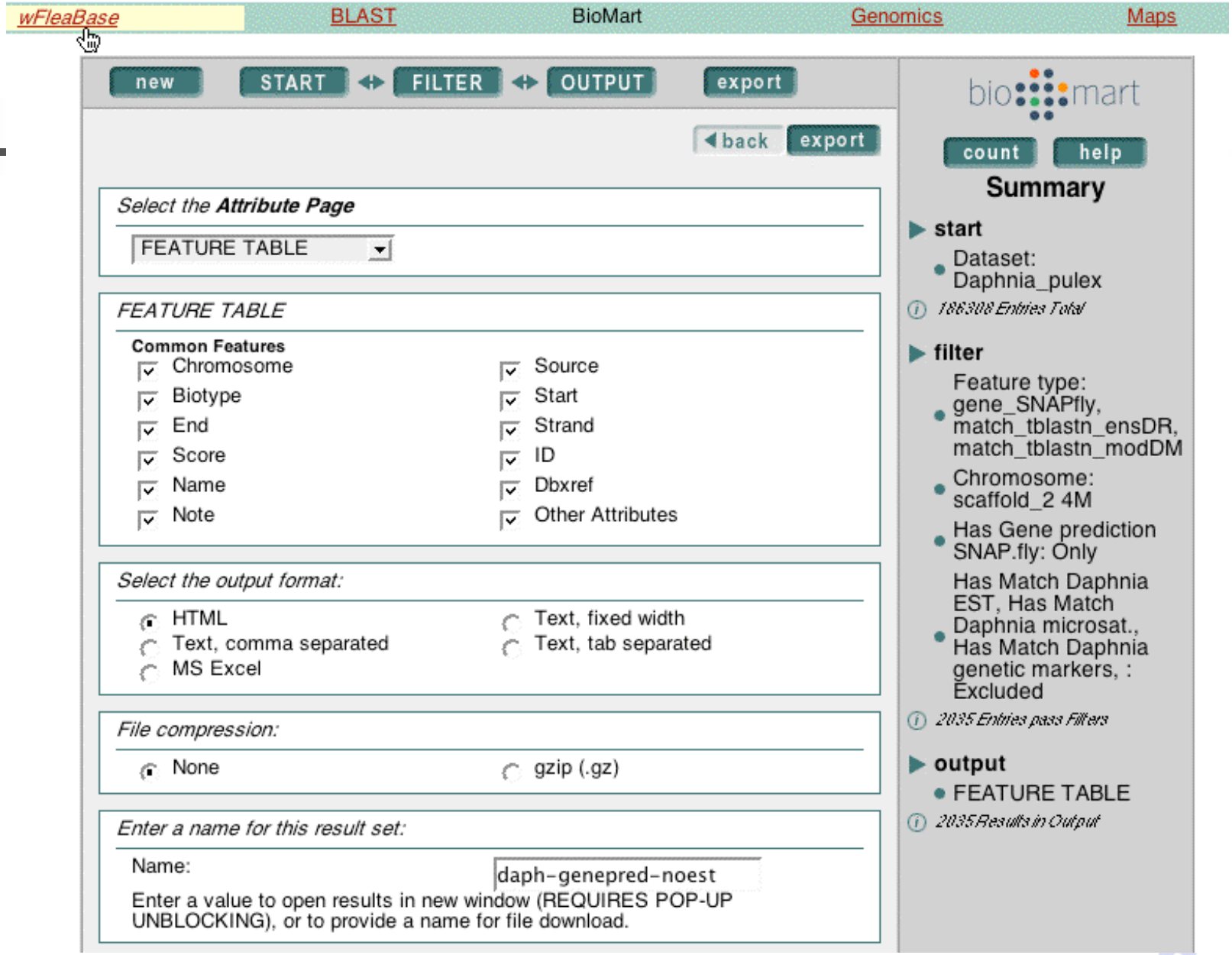
- IU and national TeraGrid group for the CPUs
- NIH for Fruitfly genomes; JGI and DGC for Daphnia genome
- GMOD project developers for the tools

GMOD

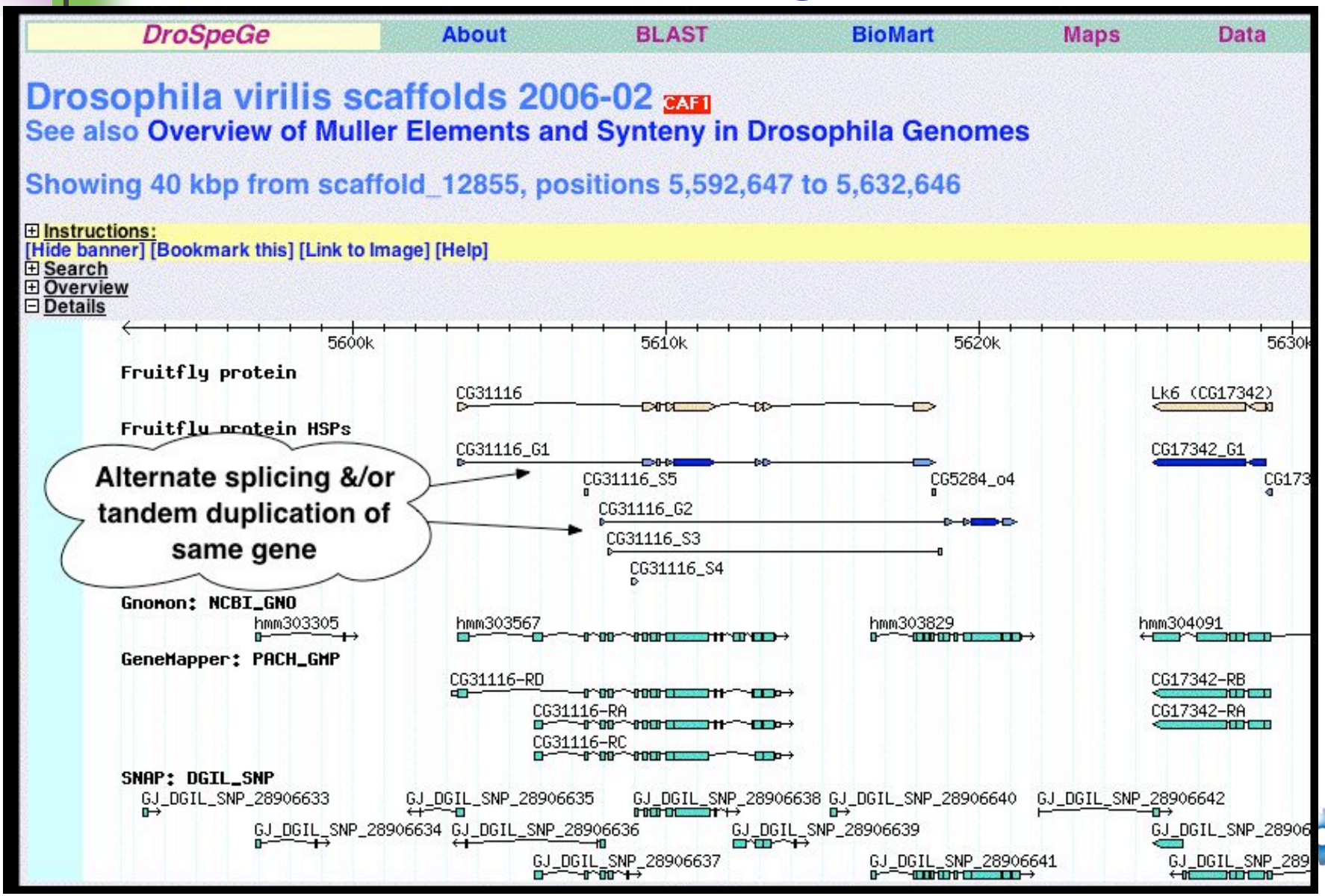
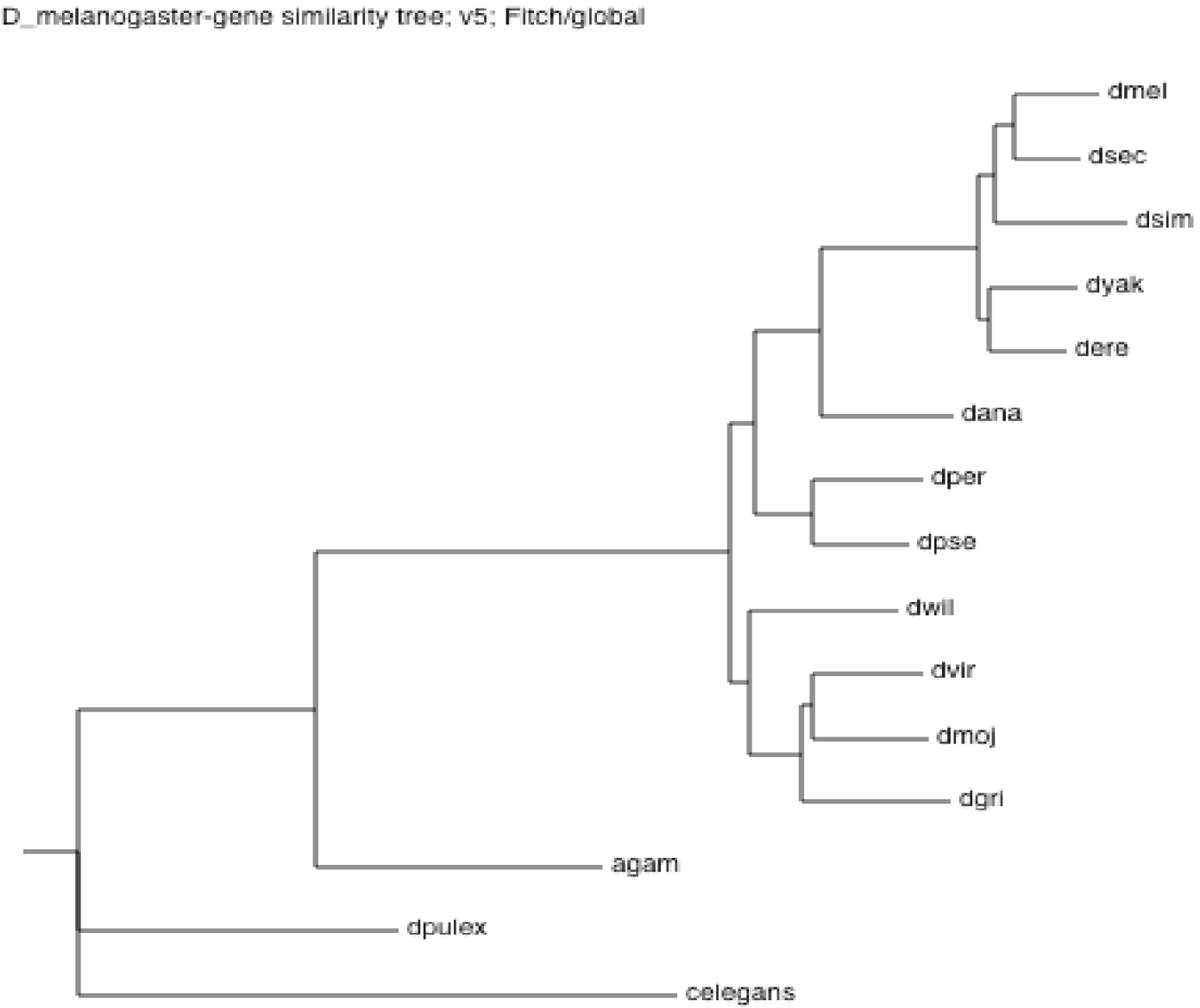
Genome Annotations

- ***Gene Homology***
  - Nine well-annotated proteomes: Yeast, Worm, Mosquito, Fruitfly, Bee, Zebrafish, Mouse, Human, Arabidopsis
  - BLAST the 13+ genomes at TeraGrid.org
- ***Gene Predictions***
  - SNAP - good ab-initio predictor, best finding new Dros. Reproductive genes.
- Collate to Gene Finding Format for map views, BioMart, sharing

BioMart Output



Alternate splicing evidence



Phylogeny from Gene Sim.