DRAFT

New and old genes in new and old *Drosophila* genomes

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

ABSTRACT

Genome assemblies of twelve *Drosophila* species with 10,000s of segments can be usefully reduced into 7 to 25 major chromosome sections, corresponding to Muller's elements. These include >= 95% of euchromatic genes, except for 3 low-coverage assemblies. Coding gene predictions from eight groups include homology-mapped known genes, *ab initio*, and homology guided predictions. Combining several predictors improves sensitivity and specificity for detecting known and new genes. The predictors vary in kinds of genomic events best predicted; different prediction sets are suggested for investigating new genes versus higher confidence in known genes. New *D. melanogaster* genes and exons are found with phylogenetic comparison, including alternate splicing exons of known genes, and new genes inside and outside known genes. Putative gain and loss of genes viewed by Gene Ontology categories helps to identify functional species differences, with major clusters for the *melanogaster* subgroup, the *obscura* siblings, and individual species effects. Differences among species are seen in reproduction, defense response, and various physiological and developmental processes. Multi-copy gene clusters have been identified through examination of over-abundances. An example is the reproductive *Est-6* tandem genes in *D. melanogaster* that have 5 to 6 tandem copies in *D. mojavensis* and *D. virilis*, with evidence of coding sequence conservation. A question of many unlocated *D. melanogaster* genes is examined further. Environmental stress expressed genes of the eco-genomics model *Daphnia pulex* found in *D. melanogaster* corroborates other evidence for 1000’s of unlocated coding genes in this model *Drosophila.*

Key words: comparative genomics, *Drosophila* species, ecological genomics, gene function, gene ontology, generic model organism database

INTRODUCTION

Twelve *Drosophila* genomes, ten recently sequenced, comprise over two billion nucleotides. The model organism *D. melanogaster* is recently at its 5th major assembly release, and continues to see significant improvements in assembly and genome features. *D. pseudoobscura*, the second sequenced *Drosophila* genome, is at its second major assembly release. One major rationale for sequencing 10 more *Drosophila* genomes is to improve, by comparative analyses, the knowledge of the model research organism. The new genomes are at the comparative annotation freeze 1 stage (CAF1), with automated annotation, quality assessment and cross-species comparisons. Sequence assemblies, annotations and predictions from all groups contributing to the Drosophila species genome community’s collaborative Assembly, Alignment and Annotation Wiki (http://rana.lbl.gov/drosophila/wiki/) have been collected as made available, analyzed and provided to bioscientists via DroSpeGe database (Gilbert 2007) in summarized and searchable formats. The independent genome analyses and summaries provided to the community are reported here. Table 1 summarizes statistics of these 12 *Drosophila* species genomes. *D. sechellia* and *D. persimilis* were sequenced at 4x coverage level, the others at 8x level, except the *D. simulans* assembly is a mosaic from

DRAFT

several different strains at 1x to 4x coverage. These assembly qualities affect results and comparative genomic interpretations.

Table 1. Drosophila species genomes, with sizes, abbreviation, Dmel genes found, and sequencing centers for CAF1 assemblies. Total assembly size (T), euchromatin (E) size, and Dmel (Me) DNA homology sizes, in Megabases, total (Nt) and euchromatic scaffold count (Ne), number of inverts to Dmel (In), and percentages of Dmel genes found in total (%Gt) and euchromatic (%Ge) scaffolds. The major euchromatic scaffolds matched to Dmel euchromatin are shown at DroSpeGe/maps/muller-elements/.

Abbr.	Species	T	Nt	E	Ne	Me	In	%Gt	%Ge	Sequencing Center
Dmel	*Drosophila melanogaster*	169	14	120	6	120	0	100.0	100.0	Berkeley Dros. Genome Prj. & Celera
Dsim	*D. simulans*	143	17	117	7	117	51	99.2	90.4	Genome Seq. Ctr., Washington U.
Dsec	*D. sechellia*	167	14730	100	20	114	33	99.7	85.8	Broad Institute
Dyak	*D. yakuba*	169	20	127	7	116	67	99.5	96.8	Genome Seq. Ctr, Washington U.
Dere	*D. erecta*	153	5124	125	7	116	70	99.4	98.6	Agencourt Bioscience Corp.
Dana	*D. ananassae*	231	13749	120	25	113	258	97.6	93.7	Agencourt Bioscience Corp.
Dper	*D. persimilis*	188	12838	102	25	110	276	95.7	74.2	Broad Institute
Dpse	*D. pseudoobscura*	153	4896	125	12	111	418	96.1	94.9	H.G.S.C., Baylor Coll. Med.
Dwil	*D. willistoni*	237	14927	140	24	103	397	95.4	89.0	J. Craig Venter Institute
Dmoj	*D. mojavensis*	194	6841	150	9	102	691	94.6	92.7	Agencourt Bioscience Corp.
Dvir	*D. virilis*	206	13530	140	15	104	515	94.8	90.3	Agencourt Bioscience Corp.
Dgri	*D. grimshawi*	200	17440	140	12	102	541	94.1	87.0	Agencourt Bioscience Corp.

METHODS

For each of twelve *Drosophila* genomes, a comparison has been made to a set of nine eukaryote 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. Protein homology detection with NCBI tBLASTn, and whole genome DNA BLASTn were performed with shared cyberinfrastructure of the TeraGrid project (www.teragrid.org), as described in Gilbert (2007). Genome database tools from GMOD project are used to organize the computations for public access. Primary results include *D. melanogaster* genome homology, homologies to nine eukaryote proteomes, gene predictions, marker gene locations, and *Drosophila* microsatellite locations. The CAF1 genome set forms the primary source data for these analyses.

The DroSpeGe database (Gilbert 2007; http://insects.eugenes.org/DroSpeGe/; URLs here are abbreviated as DroSpeGe/) provides public access to these new genomes, including sequence and annotation genome data, genome maps, data search, and browsing functions, and several of the analyses and summaries of genomes that are presented in this paper. Full data and results discussed here are available at DroSpeGe. BioMart (Kasprzyk *et al.* 2004), part of DroSpeGe public services, has been used in the analyses reported here to collect annotations of twelve genomes and organize them for analyses. GBrowse (Stein *et al*. 2002), providing detailed genome maps, is used to help validate the analyses reported here.

Total genome assembly sizes ranging from a small 143 Megabases of *D. simulans* to 237 Mb in *D. willistoni* (Table 1). Statistics in Table 1 for total assembly size (T), euchromatin (E) size, and Dmel (Me) DNA homology sizes, in megabases, were taken from faCount and related tools (Kent 2002). Euchromatic scaffold determination and counts were determined from genome BLASTn to Dmel source genome, and collated from scaffolds > 1 MB with significant homology (DroSpeGe/maps/muller-elements/). Number of inverts to Dmel (In), are derived from changes in homology orientation in genome BLASTn results.

DRAFT

Percentages of Dmel euchromatic genes found in total (%Gt) and euchromatic (%Gt) scaffolds are determined from one or more significant (e<=0.001) matches to genomes using tBLASTn.

GENE PREDICTIONS

Gene predictions with SNAP (Korf 2004) have been generated in two ways: using default operation and a training set bootstrapped from each species genome (the DGIL_SNP set), and with protein homology guidance (the DGIL_SNO set). The DGIL_SNP set used Ian Korf's supplied *D. melanogaster* model (HMM) to generate training predictions on each species genome. This training set HMM was then used to generate final predictions. With Dr. Korf's advice, an option to employ evidence from gene homologies in guiding SNAP predictions was used to produce a second set, DGIL_SNO. This set generally has better sensitivity and specificity than SNAP without homology guidance. Software scripts and further details for these prediction methods are at DroSpeGe/data/work/snap-predictions/.

Sequence assemblies, annotations and predictions from other groups were collected from http://rana.lbl.gov/drosophila/wiki/index.php/Annotation_Submission. These are listed in Table 2, and are used in analyses reported here. Protein coding annotations, gene expression evidence, and transposon repeat predictions form the major parts of data analyzed here. A comparison of predictors is made of sensitivity and specificity, using a 1 KB window on genomes. Sensitivity is measured by overlap of predictions with Dmel protein homologies (old genes; DGIL_HSP data), gene expression evidence outside of Dmel genes (new genes; OLIV_EXP data), and phylogenetically conserved exons (DGIL_TEX data). These are all subject to some error determining true biological genes. Specificity is measured by predictions in regions with none of this evidence from homology or expression, also an estimate with error.

Table 2. Genome annotations for these analyses, contributed at the community AAA Wiki

Contributor	IDs	Annotation Description
S. Batzoglou Lab, Stanford	BATZ_CNA, BATZ_CON	Contrast (Gross et al 2005) predictions
M. Brent Lab, Washington Univ., St. Louis	BREN_NSC	N-SCAN (Wu et al 2004) predictions with melanogaster alignments
D. Gilbert Lab, Indiana University	DGIL_SNP, DGIL_SNO, DGIL_HSP, DGIL_TEX	SNAP (Korf 2004) predictions, eukaryote gene homologies, phylogenetic exons
M. Eisen Lab, UC Berkeley/LBNL	EISE_CEX, EISE_CGM, EISE_CGW, EISE_GLN, EISE_GLR	GeneWise (Birney et al 2004), GeneMapper (Chatterji & Pachter 2006), Exonerate (Slater & Birney 2005), Glean (Mackey et al 2006) annotations
R. Guigó Genome Bioinformatics Lab, Barcelona	RGUI_GID, RGUI_GUT	Geneid (Para et al 2000) predictions
NCBI, Bethesda	NCBI_GNO	Gnomon (Souvorov et al 2006) predictions
B. Oliver Lab, LCDB, NIDDK, NIH	OLIV_EXP	Gene expression evidence from microarray (Sturgill et al 2006)
L. Pachter Lab, UC Berkeley	PACH_GMP	GeneMapper (Chatterji & Pachter 2006) annotations
C. Ponting Lab, MRC FGU Oxford	OXFD_GPI,OXFD_GPX	Gene prediction pipeline (Heger & Ponting 2006) with Exonerate

PHYLOGENETIC IDENTIFICATION OF NEW *D. MELANOGASTER* GENES AND EXONS

A subset of four of the predictors (Table 2) contributed new predictions for *D. melanogaster.* In order to determine if new Dmel genes may be identified using phylogenetic comparison to the other species, these predictions were combined with Ka/Ks analyses (Nekrutenko *et al.* 2002) of the new Dmel gene predictions mapped to other species. The steps of this analysis are (a) select Dmel predicted coding exons that are

DRAFT

outside of known genes or transposons, and have consensus of two or more predictors (set 1, ID texon1..texon9000), and (b) select predicted exons inside gene boundaries but outside known exons or transposons (set 2, ID texon10000..12000). The DroSpeGe BioMart database was used for this, created from Drosophila CAF1 assemblies and annotations. The Flybase Dmel release 4.2 gene/feature set was used for known features. Note that some genes added as known in release 4.3 are also found in these results. Coding exon predictions for Dmel are drawn from DGIL_SNO, BATZ_CON, RGUI_GID, and NCBI_GNO (Table 2). In the next step, (c) GFF annotations and sequence were extracted for all potential new Dmel exons. Then (d) BLASTn was used to map new Dmel exons to species genomes, using a 1e-3 probability cut-off, and results were parsed for conserved/changed alignment. Then (e) Ka/Ks ratios were calculated using alignment mismatches (excluding gaps), with exon codon position 1,2 assigned as amino/non-synonymous, and position 3 as synonymous changes. Putative new Dmel exons with at least one other species showing significant alignment and Ka/Ks < 1 were selected for the result set. These results are converted to GFF for all species with exons matching, and overlapping exon predictions in those species were extracted.

SPECIES VARIATION IN BIOLOGICAL PROCESSES AND FUNCTIONS

Methods for exploring species genome variation in biological attributes is based on proteome homology matches from eukaryote model organisms, and their associated Gene Ontology attributes. The gene match data are high scoring segment pair (HSP) groupings that are located in distinct predicted genes in these genomes. From 7,000 to 9,000 predicted genes with known protein matches were identified per species. The matches include various events: gene duplications, alternate splice exons within genes, new genes that appear composed of several exons matching different known genes. The full analysis is available at DroSpeGe database, with detailed evidence pages that provide links to GBrowse genome map views showing all secondary HSPs.

This analysis includes a total of 1,500 GO term categories for genes that differed to some degree among species, and where 5 or more source genes were associated with the GO term (genes each have GO Cell, Function, and Process categories, some with multiple associations per category). An early analysis was done with GO-slim groupings of 125 high-level categories. Further examination at the detail GO-gene association level showed that grouping at higher-level categories cancelled apparent contrasting effects among species in over- and under-representation of functions.

RESULTS

COMMON CHROMOSOMES IN DROSOPHILA

Muller's elements A, B, C, D, E, F are standard names for the six chromosome arms common among *Drosophila*, as coined by Hermann Müller. The species CAF1 data have assembled most of those six chromosome arms in 7 to 25 large scaffolds (columns E, Ne of Table 1). The sex (X, Muller A) chromosome was most resistant to full assembly. The model Dmel has 120 Mb in euchromatic chromosome arms, as do the melanogaster group species with 8x assembly-coverage. Notably the drosophila group, Dmoj, Dvir, Dgri and Dwil, have 20 – 30 Mb more DNA assembled in their common chromosomes. For the high coverage assemblies, all but 2%-5% of Dmel euchromatic genes are found in these

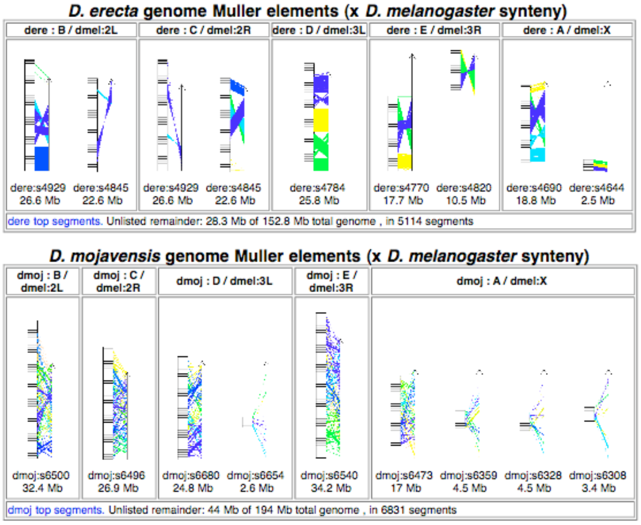

Figure 1. Muller elements of Dere (above) and Dmoj (below), and synteny with Dmel. Scaffold (left) is tied to Dmel element (right) with colored lines of high scoring segment pairs.

DRAFT

major scaffolds. For three lower quality assemblies (Dsim, Dsec, Dper), 10% to 20% Dmel genes remain in smaller scaffolds (columns %Gt - %Ge of Table 1).

A series of maps display the large-scale synteny between species chromosomes, as determined from genome x genome DNA BLAST matches, available at DroSpeGe/maps/muller-elements/. Figure 1 shows Dere and Dmoj chromosome segments, with high scoring syntenic segments to Dmel. Dere has large contiguous synteny spans compared to the many smaller clusters for Dmoj. Note for Dmoj that elements B, C, and E are all larger than Dmel elements. The melanogaster group species (Dmel, Dsim, Dsec, Dyak, and Dere) show close matching, with a small number of large-scale inversions evident in the maps (see also Table 1, column In). Among the distantly related species, the new genome assembly of *D. mojavensis* is most complete, with four Muller elements nearly fully assembled, the autosomes B to E, and the sex chromosome assembled into four major scaffolds. A few large-scale mis-assemblies are apparent for the low-coverage Dsec assembly (public communication from Thom Kaufman). The major Dsec:scaffold_0 has Muller D (3L) in its lower third, with the remainder from Muller E (3R), and Dsec:scaffold_5 contains a small part of E (3R) at its upper end of the mostly B(2L) chromosome segment. These mis-assemblies are visible in Muller and genome detail maps at DroSpeGe (DroSpeGe/maps/muller-elements/dsec-dmel-muller-r5.html, location dsec/scaffold_0:9896444-9951420).

NEW *D. MELANOGASTER* GENES AND EXONS

Over 2000 new D melanogaster genes and exons have been identified using phylogenetic comparison of protein coding region matches. Newly predicted exons in *D. melanogaster*, outside of and inside of known genes, were selected, and BLAST-compared to other species. Dmel predicted exons are classified as valid if they had significant similarity and a high synonymous substitution rate (low Ka/Ks ratio; Nekrutenko *et al.* 2002) in one or more other species. About 1800 new exons were found outside of known gene regions, and 300 inside known gene regions. Table 3 summarizes these statistics. Unclassified and heterochromatic scaffolds contain approximately 500 additional new Dmel exons, but are harder to distinguish from repetitive copies. Classification includes exons of new genes and alternate splice exons of known genes. This is a preliminary search for new Dmel genes; additional ones may be found with predictions from related species. Predictor sensitivity at finding these phylogenetically conserved exons is given in Table 4 as the Enew.P statistic. Detailed results are provided at DroSpeGe/news/newgenes-dmel/, as GFF and sequence data files sorted by texon ID and species, and as web pages with links for viewing locations on the genome maps.

Table 3. Counts of new Dmel exons found on euchromatic scaffolds, average BLAST bitscore, and average Ka/Ks ratio.

Species	N	Bitscore	Ka/Ks
Dmel	1624	--	--
Dsec	2310	414	0.507
Dsim	3353	408	0.572
Dyak	4094	388	0.584
Dere	3154	307	0.560
Dana	1962	87	0.314
Dpse	3159	64	0.172
Dmoj	2494	62	0.173

As example results, two or three new multi-exon gene locations were identified at the base of X chromosome (newly added sequence in *D. melanogaster* genome release 4), comprising texon7201 .. texon7214 that matches known gene *RhoGAP1A*, and texon7216..texon7222 that matches computed gene CG17707. A putative new *Ubx* exon, texon11429, is identified. A gene intronic to *Antp,* texon11244,45 is found also in GenBank (accession BK002361) but not in the Dmel genome release. A new exon, texon11435, of the *Adb-B* gene is located. Two single-exon genes, texon1943, texon1946, are identified between genes *rho* and *stet* on the 3L arm. Two genes, texon1660,61 and texon1664 are identified above *dpr* on the 2R arm. These texons are all similarly located on several species, and show 3rd position codon changes among species in excess of position 1,2 codons.

DRAFT

GENE PREDICTORS COMPARED

Several groups have provided gene predictions for this comparative annotation freeze set. Prediction methods include homology mapping of Dmel known genes, pure *ab initio* prediction, and homology guided *ab initio* predictions, referenced in Table 2. Additional methods for combining prediction sets to a median or consensus of predictors were used. Stajich (2006; Appendix A) provides a useful summary of current gene prediction methods. An estimate of sensitivity for these predictions, at finding new and old genes, along with specificity at calling genes only where there is evidence, is provided here. Homology mapping methods show highest specificity (*Pv*) as they find essentially only known genes, yet miss some apparent homologs compared to *ab initio* methods (lower *Eold*). *Ab intio* with homology guidance methods offer the best sensitivity at finding new and known genes. Using the Glean combination of all predictors produces generally higher values over all statistics. Visual checking of hundreds of the predictions suggests that each predictor makes errors that the others do not, at a given locus.

Table 4. Gene predictors are compared for sensitivity and specificity, averaged over all species, at the exon level. Sensitivities for old (*Eold*) genes are determined by overlap of prediction exons with homology matches. Sensitivities for new genes (*Enew.O* and *Enew.P*) are determined from overlap with species gene expression exons (O) and phylogenetic assessed Dmel exons (P). Specificity, *Pv*, is 1 – (predicted exons with none of the above evidence / total predicted exons). *Ntr* is number of predicted transcripts (distinct genes for some predictors). Predictor classes (*Ab initio*, *Ab initio* with homology guidance, Homology mapping, Combiner) are discussed in the text. The random predictor is a randomly sampled null hypothesis for predictions.

	Ab initio				*Ab initio* + Homology				
Statistic	DGIL_SNP	RGUI_GID	BATZ_CNA	A ave.	BREN_NSC	DGIL_SNO	NCBI_GNO	AH ave.	Random
Eold	0.914	0.912	0.906	0.911	0.938	0.940	0.938	0.939	0.501
Enew.O	0.796	0.627	0.550	0.658	0.608	0.801	0.541	0.650	0.501
Enew.P	0.693	0.535	0.420	0.549	0.504	0.712	0.377	0.531	0.154
Pv	0.564	0.789	0.880	0.744	0.828	0.576	0.842	0.749	0.296
Ntr	34223	22418	15606	24082	18492	33278	19719	23830	0

	Homology mapping						Combiner		
Statistic	EISE_CEX	EISE_CGM	EISE_CGW	OXFD_GPI	PACH_GMP	H ave.	EISE_GLN	EISE_GLR	C ave.
Eold	0.901	0.908	0.891	0.844	0.879	0.885	0.968	0.951	0.960
Enew.O	0.177	0.175	0.173	0.156	0.202	0.177	0.663	0.607	0.635
Enew.P	0.111	0.070	0.073	0.075	0.062	0.078	0.525	0.442	0.484
Pv	0.945	0.958	0.969	0.949	0.948	0.954	0.827	0.860	0.844
Ntr	34429	18440	18817	12281	12237	19241	25501	19219	22360

Predictor Key: DGIL_SNP: SNAP predictor, no homology; RGUI_GID: geneid predictor; BATZ_CNA: CONTRAST predictor; BREN_NSC: N-SCAN predictor using homology evidence; DGIL_SNO: SNAP predictor using homology evidence; NCBI_GNO: GNOMON predictor using homology evidence; EISE_GLN: Glean prediction combiner; EISE_GLR: Glean recombined with higher weight to gene mappers; UMD_JIG : Jigsaw prediction combiner; EISE_CEX: Exonerate (Dmel gene mapping); EISE_CGM: GeneMapper (Dmel mapping); EISE_CGW: GeneWise (Dmel mapping); OXFD_GPI: Oxford gene pipeline (exonerate with Dmel mapping); PACH_GMP: GeneMapper (Dmel mapping); A ave, AH ave, H ave, and C ave are average values for the predictor classes.

The predictions have been made with different goals, affecting this comparison of sensitivities. For example, SNAP predictions were aimed at finding all potential coding exons, several gene mappers were directed at locating known Dmel genes. The Oxford pipeline (OXFD_GPI) calls have been partitioned into

DRAFT

classes of putative pseudogene, gene fragment, non-coding and putative coding gene, with just the putative coding gene set provided for community analyses. Protein HSP matches used here to measure known gene sensitivity will include a different set of HSP exon matches than the Oxford group classification of true versus pseudogene exons (Heger and Ponting, 2007).

The SNAP plus homology predictor DGIL_SNO has a low 60% specificity, compared to 80% to 90% for other *ab initio* methods, but it finds around 30% more biological signals (identified with gene expression and phylogenetic Ka/Ks evidence) than other predictors. In this analysis, SNAP appears to be over-predicting by about 50%, but these are gene-like predictions, and among these are true positives that other *ab initio* and homology predictors are missing.

When one wants to find new genes of a species, SNAP is a good choice. One needs further experimental verification to weed out the false positives. By starting with a conservative gene prediction set, real biological signals will be lost to future work. Using a combiner of several predictor methods can yield a better result. A useful gene prediction collection would include a conservative set for counting genes and high confidence, and an exploratory set, of lower confidence but more likely to contain the full biological gene set.

SPECIES VARIATION IN BIOLOGICAL PROCESSES AND FUNCTIONS

Potential gain and loss of gene functions and biological processes among species genomes has been assessed, in an exploratory analysis to indicate where species may differ genetically in biological attributes. Note that ‘lost genes' may be due to divergence in genes rather than loss, and 'gained genes' should be interpreted with caution also.   Proteomes from *D. melanogaster*, *C. elegans* and *M. musculus* provided source genes with extensive, experimentally determined Gene Ontology annotations. Dmel protein source is discussed here; the other model organisms yielded fewer, but some interesting contrasts. Full results are provided at DroSpeGe/news/genome-summaries/gene-GO-function-association/

Table 5 identifies statistically significant Chi-square deviations in over or under representation of genes in several GO categories. Phylogenetic species clusters are evident with some of these, as well as single species differences. Correspondence analyses identify three major contrast clusters: the contrast between melanogaster subgroup and others, with *D. ananassae* intermediate, a contrast between the obscura group siblings and others, and thirdly, individual species effects.

These contrasts are evident in Table 5: the non-melanogaster species are overrepresented in genes for defense response to bacterium, certain reproductive processes, and glycolipid biosynthesis. The melanogaster subgroup is overrepresented in mating behavior, puparial adhesion, regulation of amino acid metabolism, and a large category of unknown processes, apparently species-specific. The obscura species are overrepresented in morphogenesis of an epithelium, protein modification, and RNA interference, among others.   Among single-species contrasts, *D. willistoni* is overrepresented in steroid metabolism, and *D. grimshawi* is overrepresented in mitosis related genes.   Biological implications of these statistical gains and losses can be learned through details of the genes involved. One such story is discussed below.

DRAFT

Table 5. Potential gain and loss of genes for biological processes in *Drosophila* species. This exploratory analysis identifies GO categories with significant deviations from expected number of genes, as measured with *D. melanogaster* proteins located in part or whole, sometimes at several distinct gene locations. GO ID, term and categories flank the species gain (++) or loss (--) indicators. *Ng* lists the average gene count.

GO_ID	dmel	dsim	dsec	dyak	dere	dana	dper	dpse	dwil	dmoj	dvir	dgri	Ng	Term and Grouping
GO:0000004	+++++	+++	+++++	++++	++++	.	----	---	----	--	---	-----	627	biological_process unknown, P:
GO:0007224	--	.	.	-	-	.	++	.	+	+	.	.	36	smoothened signaling pathway, P:cellular process
GO:0002009	.	-	.	-	.	.	+++	++	-	.	.	.	35	morphogenesis of an epithelium, P:development
GO:0000281	.	.	-	.	.	.	.	++	.	.	.	+++	4	cytokinesis after mitosis, P:physiological process
GO:0006118	-	-	--	-	.	+	.	.	++	.	+	++	302	electron transport, P:physiological process
GO:0006334	.	-	++	--	++	.	+	--	.	.	.	.	44	nucleosome assembly, P:physiological process
GO:0006464	-	-	--	-	-	.	+++	++	.	+	.	+	170	protein modification, P:physiological process
GO:0006508	.	.	--	.	.	.	.	.	+	.	.	++	767	proteolysis, P:physiological process
GO:0006511	-	-	--	.	-	.	+++	+	.	+	+	.	82	ubiquitin-dependent protein catabolism, P:physiological process
GO:0007076	-	.	.	.	.	.	++	++	.	.	.	.	18	mitotic chromosome condensation, P:physiological process
GO:0007594	++	+	+	++	.	.	.	.	.	-	-	--	5	puparial adhesion, P:physiological process
GO:0008202	.	-	-	.	.	++	-	.	++++	.	.	.	108	steroid metabolism, P:physiological process
GO:0008363	-	-	-	++	--	++	+	++	--	-	.	++	5	larval cuticle biosynthesis (sensu Insecta), P:physiological process
GO:0009247	-	--	-	.	-	.	.	.	.	++++	++	++	6	glycolipid biosynthesis, P:physiological process
GO:0015977	-	.	-	-	-	.	+	+	.	++	++	+	7	carbon utilization by fixation of carbon dioxide, P:physiological proc
GO:0016567	--	-	-	-	.	.	+++	+	.	+	.	.	141	protein ubiquitination, P:physiological process
GO:0018401	-	.	.	-	.	.	+++	++	.	.	-	.	19	peptidyl-proline hydroxylation to 4-hydroxy-L-proline, P:physiological
GO:0030261	-	-	.	-	.	-	++	++	+	.	.	.	18	chromosome condensation, P:physiological process
GO:0030422	.	.	.	.	.	-	+++	+++	.	.	.	.	6	RNA interference, production of siRNA, P:physiological process
GO:0030423	.	.	.	.	.	-	++	++	.	.	.	.	8	RNA interference, targeting of mRNA for destruction, P:physiological p
GO:0035087	-	.	.	.	.	-	++	+++	.	-	-	.	6	RNA interference, siRNA loading onto RISC, P:physiological process
GO:0051227	-	.	++	.	-	-	-	-	.	.	.	++++	5	mitotic spindle assembly, P:physiological process
GO:0006521	+	++	++	++	+	-	-	.	-	-	--	--	12	regulation of amino acid metabolism, P:regulation of biological proces
GO:0030728	.	.	-	.	.	.	.	.	-	++	++	.	6	ovulation, P:reproduction
GO:0007618	.	-	-	.	.	.	.	.	.	++	++	.	5	mating, P:response to stimulus
GO:0042742	--	--	---	-	--	.	++	++	+++	.	++	++	46	defense response to bacterium, P:response to stimulus
GO:0045297	++	+++	+++	++	+	.	.	.	---	--	--	---	5	post-mating behavior, P:response to stimulus
GO:0046692	.	.	.	-	.	.	++	+	--	+	++	-	9	sperm competition, P:response to stimulus

DRAFT

DISCUSSION

Biological implications of the statistical gene gains and losses in functions can be found in details of the genes involved. As an example, the reproductive gene excess (GO:0030728,GO:0007618) in Dmoj and Dvir includes the Dmel β-esterase tandem genes *Est-6* (CG6917) and *Est-P* (CG17148; recently renamed *ψ-Est-6* in Balakirev and Ayala, 2004).

There are 5 to 6 copies of these genes in a tight cluster in Dmoj and Dvir, each similar in size with 2 coding exons (dmoj/scaffold_6540:24423577..24448576 and dvir/scaffold_12855:9338335..9363334). All are strongly homologous to each other, and to only the two genes of Dmel. Gene expression evidence matches all 6 genes in Dmoj. Five *ab initio* gene predictors consistently identify all their exons. Five homology predictors however failed to adequately detect this cluster, calling half the genes and partial exons. Eleven gene predictions of NCBI Gnomon, best for this cluster, from Dvir and Dmoj were aligned and coding/noncoding distances (Ka/Ks) calculated, with scores from 0.74 to 0.44 for the most distant, consistent with conservation of the 11 coding sequences within and between species. Dpse and Dper have three copies (dpse/XR_group6:2503342..2528341; dper/scaffold_9:798097..823096). Dwil has only one (dwil/scaffold_180949:5111643..5121642), while Dgri has a full esterase gene and a partial match, suggestive of a degenerate pseudogene (dgri/scaffold_14830:1806937..1816936). Dana has 4 genes in this cluster (dana/scaffold_13337:1444999..1469998). Dere, Dyak, Dsec, Dsim and Dmel are similar in having the known tandem pair, with no other matching fragments in their genomes.

Balakirev and Ayala (2004) describe this β-esterase gene cluster as relatively ancient, with two closely linked genes in melanogaster group Drosophila, but three in *D. pseudoobscura*. The *Est-6* genes have a well-known allozyme dimorphism among populations of different species. These authors examine population variation among the gene cluster, and find clear indications of selection for polymorphism. They conclude this is an interacting complex, or intergene, with two or more genes required to perform the final function. By coincidence, this author studied the reproductive physiology and behavioral effects of *Est-6* many years ago (Richmond et al, 1980; Gilbert & Richmond 1982). Despite this prior work, I would have missed spotting this interesting gene expansion/contraction in the forest of 200,000 species-genes, without examining variation at higher functional levels.

The success of *ab initio* predictors over homology mappers on this particular gene cluster indicates, along with statistics on detecting old and new genes (Table 4), that one should use care in choosing a gene prediction set; different predictors are useful for different analyses.

Phylogenetic comparisons of synonymous/non-synonymous base changes for new predicted coding exons will aid in a fuller characterization of the *D. melanogaster* genome. Further experimental evidence is needed to validate these phylogenetically conserved exons. Yandell *et al.* (2005) provide strong evidence that the ~14,000 known protein genes of Dmel are a complete biological complement of this genome. The phylogenetic evidence reported here for new coding exons adds to other studies (Stolc *et al.* 2004; Halligan and Keightley 2006; Hild *et al.* 2003; Siepel *et al.* 2005) that indicate phylogenetically conserved and selected genome "dark matter" remains to be fully characterized, including unknown protein genes, non-coding genes, and other functional elements.

Gene expression under stressful conditions is an expected natural selection on an organism's gene complement. Genome-wide stress expression studies are common for the models Mouse, Yeast, and *Arabidopsis* (Google Scholar and Entrez PubMed rank these highest), but research on environmental and stress effects on genomes of *Drosophila* and *C. elegans* models have an order of magnitude fewer publications. An alternative explanation for Yandell *et al.* (2005) failing to find evidence for further Dmel genes may be based on their reliance on expression in a single unstressed, inbred lab population.
To further understand whether *Drosophila melanogaster* may have many unlocated protein genes, comparison to the *Daphnia pulex* genome (http://wFleabase.org/prerelease/; Colbourne *et al.* 2005) was

DRAFT

undertaken. This aquatic arthropod has been a model for studies of ecological and ecotoxicological factors for centuries. Recent *Daphnia* whole genome studies include examination of genome-wide expression under known environmental stresses. Some 19,000 *Daphnia* computed genes have been classified by expression in normal and stressed environments, based on 160,000 ESTs derived from 35 environment treatments. The aquatic conditions include abiotic stresses (toxic metals, hypoxia, acid, UV radiation), and biotic stresses (bacterial infection, predation, starvation).

Of 2511 new Dmel conserved exons, 780 matched with tBLASTx to *Daphnia* genes, with 25% more stress expressed than normal genes (p<0.005, measured by count and homology strength). In a complementary examination, *Daphnia* predicted genes were matched with tBLASTx to the *Drosophila melanogaster* genome outside of known exon and transposon regions. Of these, 1100 match to the uncharacterized *Drosophila* euchromatin, in proportions equal to their expression treatment in *Daphnia* (40% stress, 60% normal). There is an indication from the *Daphnia* genes that do not match any known *Drosophila* gene, that stress genes are more frequent in the *Drosophila* uncharacterized regions (60% stress, 40% normal, χ^2 p<1e-10).

This evidence suggests that *D. melanogaster's* genome contains 1000's of protein genes yet to be fully identified, and that expression under environmental stresses is one clue to their experimental validation. Defense response to bacteria is one *Drosophila* species gene difference found (Table 5), understandable for these insects that live on widely variable yeast and bacterial microflora of decomposing plants. The *Daphnia* comparison suggests that genomic studies of such biotic stresses may uncover many more *Drosophila* genes. Ecological and environmental studies, in concert with comparative and phylogenetic studies, are a useful and necessary addition to full knowledge of organism genomes.

ACKNOWLEDGEMENTS

This project has been supported in part by grants from the NHGRI of the National Institutes of Health, and from Sun Microsystems, to D. Gilbert. National Science Foundation TeraGrid access grant provided computing resources. I thank Christopher Mason, Kevin White, and Andreas Heger for comments on aspects of this paper. Several others, including those referenced in Table 2, have contributed information and insights via public discussions on the Multiple Fruitflies mail and AAA web venues hosted by the Michael Eisen laboratory.

REFERENCES

Balakirev, E and Ayala, F. 2004. The beta-esterase Gene Cluster of *Drosophila melanogaster*: is psi-Est-6 a Pseudogene, a Functional Gene, or both? *Genetica*, **121**:1573-6857 doi:10.1023/B:GENE.0000040391.27307.51

Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and Genomewise. *Genome Res.* **14**:988-995, doi:10.1101/gr.1865504

Chatterji, S. and Pachter, L. 2006. Reference based annotation with GeneMapper, *Genome Biol.*, **7**:R29, doi:10.1186/gb-2006-7-4-r29 URL: http://bio.math.berkeley.edu/genemapper/

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

Gilbert, D. G., and R. C. Richmond. 1982. Studies of esterase 6 in *Drosophila melanogaster* XII. Evidence for temperature selection of Est-6 and Adh alleles. *Genetica* **58**: 109- 119.

DRAFT

Gilbert, D.G. 2007. DroSpeGe: rapid access database for new *Drosophila* species genomes. *Nucleic Acids Res.* **35**(Database issue): D480-D485; doi:10.1093/nar/gkl997

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/

Halligan, D. L. and Keightley, P.D., 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* **16**:875-884. doi:10.1101/gr.5022906

Heger A. and Ponting C. 2007. Evolutionary rate analyses of orthologues and paralogues from twelve *Drosophila* Genomes. *Genome Res.*, in review. URL: http://wwwfgu.anat.ox.ac.uk:8080/flies/documentation.html

Hild, M., Beckmann, B., Haas, S. A., Koch, B., Solovyev, V., Busold, C., Fellenberg, K., Boutros, M., Vingron, M., Sauer, F., Hoheisel, J.D., and Paro, R., 2003. An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol.* **5**, R3. http://genomebiology.com/2003/5/1/R3

Kasprzyk, A, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P Rocca-Serra, T. Cox and Birney E. 2004. EnsMart: A Generic System for Fast and Flexible Access to Biological Data. *Genome Res.* **14**(1):160-169. URL: www.biomart.org

Kent, W.J. 2002. BLAT -- The BLAST-Like Alignment Tool. *Genome Res.* **12**(4): 656-664 doi:10.1101/gr.229202; URL: http://www.soe.ucsc.edu/~kent/

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**:59 doi:10.1186/1471-2105-5-59

Mackey A.J., Pereira F.C.N., and Roos, D.S. 2006. GLEAN: improved eukaryotic gene prediction by statistical consensus of gene evidence. Manuscript in draft.

Nekrutenko A., Makova K. D. and Li, W-H. 2002. The KA/KS Ratio Test for Assessing the Protein-Coding Potential of Genomic Regions: An Empirical and Simulation Study. Genome Res. 2002 12: 198-202; doi:10.1101/gr.200901

Parra, G., Blanco, E. and Guigó,R. 2000. Geneid in Drosophila. *Genome Res.* **10**(4):511-515. URL: http://genome.imim.es/software/geneid/index.html

Richmond, R.C., D.G. Gilbert, K.B. Sheehan, M.H. Gromko & F.M. Butterworth, 1980. Esterase 6 and reproduction in *Drosophila melanogaster*. *Science* **207**: 1483–1485.

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 Res.* **15**, 1034-1050.

Slater, G. St. C. and Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**:31 doi:10.1186/1471-2105-6-31

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/

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/

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/

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M. F., Rifkin, S. A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P. E., Bussemaker, H. J., White, K. P., 2004. A Gene Expression Map for the Euchromatic Genome of *Drosophila melanogaster*. *Science* **306**: 655-660; doi:10.1126/science.1101312

DRAFT

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

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 Res.* **14**: 665-671. URL: http://ardor.wustl.edu/ with N-SCAN update.

Yandell, M., Bailey, A. M., Misra, S., Shu, S., Wiel, C., Evans-Holm, M., Celniker, S.E., and Rubin, G.M. 2005. A computational and experimental approach to validating annotations and gene predictions in the *Drosophila melanogaster* genome. *Proc. Natl. Acad. Sci.*, **102**(5):1566–1571 doi:10.1073/pnas.0409421102