holder Ns) that have no BLAST hit at 99% identity for finished data and 95% identity for light-shotgun data were considered uncovered. The percentage of each clone not hit by Celera sequence was calculated by dividing the total length of the uncovered sequence by the sequence length of the clone. The total number of nucleotides that have no coverage in the Celera assembled contigs was calculated by summing the regions of no hits for all the clones that covered Celera contigs by less than 90% (95% for finished clones). This cutoff value was chosen to eliminate the occasionally low quality of sequences in the clone sequence data. The cutoff value of 90% was determined by the amount of no-hit sequences in 16 light-shotgun clones that are fully contained within three Celera contigs. A higher cutoff value (95%) was used for the finished data than for the light-shotgun data, because finished clones have better sequence quality. The total amount of uncovered sequence for each light-shotgun clones was calculated by multiplying the no-hit percentage of the clone by the clone length as determined by sizing on agarose gels (*36*). For those light-shotgun clones with unreported insert sizes, the sequence length, excluding Ns, was used instead. For finished clones, the amount of uncovered sequence was calculated by multiplying the no-hit percent of the clone by the clone's length. We created 7-kbp subcontig blocks and considered each block to be fully present in the draft sequence if it was hit by at least 500 bp of external sequences. We chose these parameters conservatively, based on the fact that at 1× sequence coverage, the chance of failing to sample a 7-kbp region covered by a light-shotgun clone is 1 in 10^6. For the WGS assembly, we identified 1380 blocks that were hit by less than 500 bp of clone sequence and 794 blocks that were completely missed by the clone sequence. The total number of missed blocks is 2174, which represents a total 15.2 Mbp.

34. M. Ashburner *et al.*, *Genetics* **153**, 179 (1999).
35. Seven conflicts were identified in this study, six of which appear to be owing to transposable elements. The remaining represents a 30-kbp insert within a Celera contig that does not match the corresponding clone. This discrepancy is still under investigation.
36. www.sciencemag.org/feature/data/1049666.shl
37. S. Altschul *et al.*, *Nucleic Acids. Res.* **25**, 3389 (1997).
38. R. A. Hoskins, personal communication.
39. In order to align the Celera sequences unambiguously to the external data, all significant HSPs at the parameters given in (*27*) were screened to identify "mutually unique regions" where the clone and contig sequences have a unique, reciprocal match relation.
40. Most negative gaps arise because of inaccuracies in the distances implied by bundles—the bundle implies a small amount of overlap between two contigs because it is actually short, whereas the reality is that there is a small gap at that location. In a very small number of cases, there is an overlap, but it is because the distance estimate is too long by 3 standard deviations, or because there is a small bit of foreign DNA at the tip of a contig because of untrimmed vector or a chimeric read. None of these negative gaps has yet been found to imply incorrect assembly.

41. We wish to thank H. Smith and S. Salzburg for the many collegial exchanges, M. Peterson and his team for keeping the machines humming, R. Thompson and his staff for providing us with an environment conducive to such an intense effort, and A. Glodek, C. Kraft, and A. Deslattes Mays, and their staff for getting the data to us.

REVIEW

Comparative Genomics of the Eukaryotes

Gerald M. Rubin,[1] Mark D. Yandell,[3] Jennifer R. Wortman,[3] George L. Gabor Miklos,[4] Catherine R. Nelson,[2] Iswar K. Hariharan,[5] Mark E. Fortini,[6] Peter W. Li,[3] Rolf Apweiler,[7] Wolfgang Fleischmann,[7] J. Michael Cherry,[8] Steven Henikoff,[9] Marian P. Skupski,[3] Sima Misra,[2] Michael Ashburner,[7] Ewan Birney,[7] Mark S. Boguski,[10] Thomas Brody,[11] Peter Brokstein,[2] Susan E. Celniker,[12] Stephen A. Chervitz,[13] David Coates,[14] Anibal Cravchik,[3] Andrei Gabrielian,[3] Richard F. Galle,[12] William M. Gelbart,[15] Reed A. George,[12] Lawrence S. B. Goldstein,[16] Fangcheng Gong,[3] Ping Guan,[3] Nomi L. Harris,[12] Bruce A. Hay,[17] Roger A. Hoskins,[12] Jiayin Li,[3] Zhenya Li,[3] Richard O. Hynes,[18] S. J. M. Jones,[19] Peter M. Kuehl,[20] Bruno Lemaitre,[21] J. Troy Littleton,[22] Deborah K. Morrison,[23] Chris Mungall,[12] Patrick H. O'Farrell,[24] Oxana K. Pickeral,[10] Chris Shue,[3] Leslie B. Vosshall,[25] Jiong Zhang,[10] Qi Zhao,[3] Xiangqun H. Zheng,[3] Fei Zhong,[3] Wenyan Zhong,[3] Richard Gibbs,[26] J. Craig Venter,[3] Mark D. Adams,[3] Suzanna Lewis[2]

A comparative analysis of the genomes of *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*—and the proteins they are predicted to encode—was undertaken in the context of cellular, developmental, and evolutionary processes. The nonredundant protein sets of flies and worms are similar in size and are only twice that of yeast, but different gene families are expanded in each genome, and the multidomain proteins and signaling pathways of the fly and worm are far more complex than those of yeast. The fly has orthologs to 177 of the 289 human disease genes examined and provides the foundation for rapid analysis of some of the basic processes involved in human disease.

With the full genomic sequence of three major model organisms now available, much of our knowledge about the evolutionary basis of cellular and developmental processes will derive from comparisons between protein domains, intracellular networks, and cell-cell interactions in different phyla. In this paper, we begin a comparison of *D. melanogaster, C. elegans*, and *S. cerevisiae*. We first ask how many distinct protein families each genome encodes, how the genes encoding these protein families are distributed in each genome, and how many genes are shared among flies, worms, yeast, and mammals. Next we describe the composition and organization of protein domains within the proteomes of fly, worm, and yeast and examine the representation in each genome of a subset of genes that have been directly implicated as causative agents of human disease. Then we compare some fundamental cellular and developmental processes: the cell cycle, cell structure, cell adhesion, cell signaling, apoptosis, neuronal signaling, and the immune system. In each case, we present a summary of what we have learned from the sequence of the fly genome and how the components that carry out these processes differ in other organisms. We end by presenting some observations on what we have learned, the obvious questions that remain, and how knowledge of the sequence of the *Drosophila* genome will help us approach new areas of inquiry.

The "Core Proteome"

How many distinct protein families are encoded in the genomes of *D. melanogaster*, *C. elegans*, and *S. cerevisiae* (*1*), and how do these genomes compare with that of a simple prokaryote, *Haemophilus influenzae*? We carried out an "all-against-all" comparison of protein sequences encoded by each genome using algorithms that aim to differentiate paralogs—highly similar proteins that occur in the same genome—from proteins that are uniquely represented (Table 1). Counting each set of paralogs as a unit reveals the "core proteome": the number of distinct protein families in each organism. This operational definition does not include posttranslationally modifed forms of a protein or isoforms arising from alternate splicing.

In *Haemophilus*, there are 1709 protein coding sequences, 1247 of which have no sequence relatives within *Haemophilus* (*2*). There are 178 families that have two or more paralogs, yielding a core proteome of 1425. In yeast, there are 6241 predicted proteins and a core proteome of 4383 proteins. The fly and worm have 13,601 and 18,424 (*3*) predicted protein-coding genes, and their core proteomes consist of 8065 and 9453 proteins, respectively. It is remarkable that *Drosophila*, a complex metazoan, has a core proteome only twice the size of that of yeast. Furthermore, despite the large differences between fly and worm in terms of development and morphology, they use a core proteome of similar size.

Gene Duplications

Much of the genomes of flies and worms consists of duplicated genes; we next asked how these paralogs are arranged. The frequency of local gene duplications and the number of their constituent genes differ widely between fly and worm, although in both genomes most paralogs are dispersed. The fly genome contains half the number of local gene duplications relative to *C. elegans* (*4*), and these gene clusters are distributed randomly along the chromosome arms; in *C. elegans* there is a concentration of gene duplications in the recombinogenic segments of the autosomal arms (*1*). In both organisms, approximately 70% of duplicated gene pairs are on the same strand (306 out of 417 for *D. melanogaster* and 581 out of 826 for *C. elegans*). The largest cluster in the fly contains 17 genes that code for proteins of unknown function; the next largest clusters both consist of glutathione S-transferase genes, each with 10 members. In contrast, 11 of 33 of the largest clusters in *C. elegans* consist of genes coding for seven transmembrane domain receptors, most of which are thought to be involved in chemosensation. Other than these local tandem duplications, genes with similar functional assignment in the Gene Ontology (GO) classification (*5*) do not appear to be clustered in the genome.

We next compared the large duplicated gene families in fly, worm, and yeast without regard to genomic location. All of the known and predicted protein sequences of these three genomes were pooled, and each protein was compared to all others in the pool by means of the program BLASTP. Among the larger protein families that are found in worms and flies but not yeast are several that are associated with multicellular development, including homeobox proteins, cell adhesion molecules, and guanylate cyclases, as well as trypsinlike peptidases and esterases. Among the large families that are present only in flies are proteins involved in the immune response, such as lectins and peptidoglycan recognition proteins, transmembrane proteins of unknown function, and proteins that are probably fly-specific: cuticle proteins, peritrophic membrane proteins, and larval serum proteins.

Gene Similarities

What fraction of the proteins encoded by these three eukaryotes is shared? Comparative analysis of the predicted proteins encoded by these genomes suggests that nearly 30% of the fly genes have putative orthologs in the worm genome. We required that a protein show significant similarity over at least 80% of its length to a sequence in another species to be considered its ortholog (*6*). We know that this results in an underestimate, because the length requirement excludes known orthologs, such as homeodomain proteins, which have little similarity outside the homeodomain. The number of such fly-worm pairs does not decrease much as the similarity scores become more stringent (Table 2A), which strongly suggests that we have indeed identified orthologs, which may share molecular function. Nearly 20% of the fly proteins have a putative ortholog in both worm and yeast; these shared proteins probably perform functions common to all eukaryotic cells.

We also compared the proteins of fly, worm, and yeast to mammalian sequences. Most mammalian sequences are available as short expressed sequence tags (ESTs), so we dispensed with the requirement for similarity over 80% of the length of the proteins. Table 2B presents these data. Half of the fly protein sequences show similarity to mammalian proteins at a cutoff of $E < 10^{-10}$ (where E is expectation value), as compared to only 36% of worm proteins. This difference increases as the criteria become more stringent: 25% versus 15% at $E < 10^{-50}$ and 12% versus 7% at $E < 10^{-100}$. Because many of the comparisons are with short sequences, it is likely that many of these sequence similarities reflect conserved domains within proteins rather than orthology. However, it does suggest that the *Drosophila* proteome is more similar to mammalian proteomes than are those of worm or yeast.

Protein Domains and Families

Proteins are often mosaic, containing two or more different identifiable domains, and domains can occur in different combinations in different proteins. Thus, only a portion of a protein may be conserved among organisms. We therefore performed a comparative analysis of the protein domains composing the predicted proteomes from *D. melanogaster*, *C. elegans*, and *S. cerevisiae* using sequence similarity searches against the SWISS-PROT/TrEMBL nonredundant protein database (*7*), the BLOCKS database (*8*), and the InterPro database (*9*). The 200 most common fly protein families and domains are listed in Table 3, and the 10 most highly represented families in worm and yeast are shown in Table 4. InterPro analyses plus manual data inspection enabled us to assign 7419 fly proteins, 8356 worm proteins, and 3056 yeast proteins to either protein families or domain families. We found 1400 different protein families or domains in all: 1177 in the fly, 1133 in the worm, and 984 in yeast; 744 families or domains were common to all three organisms.

Many protein families exhibit great disparities in abundance, and only the C2H2-type zinc finger proteins and the eukaryotic

[1]Howard Hughes Medical Institute, [2]Department of Molecular and Cell Biology, Berkeley *Drosophila* Genome Project, University of California, Berkeley, CA 94720, USA. [3]Celera Genomics, Rockville, MD, 20850 USA. [4]GenetixXpress, 78 Pacific Road, Palm Beach, Sydney, Australia 2108. [5]Massachusetts General Hospital Cancer Center, Building 149, 13th Street, Charlestown, MA 02129 USA. [6]Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA. [7]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [8]Department of Genetics, Stanford University, Palo Alto, CA 94305, USA. [9]Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. [10]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. [11]Neurogenetics Unit, Laboratory of Neurochemistry, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD 20892, USA. [12]Berkeley *Drosophila* Genome Project, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [13]Neomorphic, 2612 Eighth Street, Berkeley, CA 94710, USA. [14]School of Biology, University of Leeds, Leeds LS2 9JT, UK. [15]Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA. [16]Departments of Cellular and Molecular Medicine and Pharmacology, Howard Hughes Medical Institute, University of California–San Diego, La Jolla, CA 92093, USA. [17]Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA. [18]Howard Hughes Medical Institute, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [19]Genome Sequence Centre, BC Cancer Research Centre, 600 West 10th Avenue, Vancouver, BC, V52 4E6, Canada. [20]Molecular and Cell Biology Program, University of Maryland at Baltimore, Baltimore, MD 21201, USA. [21]Centre de Génétique Moléculaire, CNRS, 91198 Gif-sur-Yvette, France. [22]Center for Learning and Memory, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. [23]Regulation of Cell Growth Laboratory, Division of Basic Sciences, National Cancer Institute–Frederick Cancer Research and Development Center, National Institutes of Health, Frederick, MD 21702, USA. [24]Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143, USA. [25]Center for Neurobiology and Behavior, Columbia University, New York, NY 10032, USA. [26]Baylor College of Medicine Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA.

Table 1. Numbers of distinct gene families versus numbers of predicted genes and their duplicated copies in *H. influenzae*, *S. cerevisiae*, *C. elegans*, and *D. melanogaster*. Row one shows the total number of genes in each species. Row two shows the total number of all genes in each genome that appear to have arisen by gene duplication. Row three is the total number of distinct gene families for each genome. Each proteome was compared to itself using the same parameters as described in (*63*).

	H. influenzae	*S. cerevisiae*	*C. elegans*	*D. melanogaster*
Total no. of predicted genes	1709	6241	18424	13601
No. of genes duplicated	284	1858	8971	5536
Total no. of distinct families	1425	4383	9453	8065

protein kinases are among the top 10 protein families common to all three organisms. There are 352 zinc finger proteins of the C2H2 type in the fly but only 138 in the worm; whether this reflects greater regulatory complexity in the fly is not known. The protein kinases constitute approximately 2% of each proteome. Curation of the genomic data revealed that *Drosophila* has approximately 300 protein kinases and 85 protein phosphatases, around half of which had previously been identified. In contrast, there are approximately 500 kinases and 185 phosphatases in the worm; the difference is largely due to the worm-specific expansion of certain families such as the CK1, FER, and KIN-15 families. There are currently approximately 600 kinases and 130 phosphatases in humans, and it is expected that these figures will rise to 1100 and 300, respectively, when the sequence of the human genome is completed (*10*). Of the proteins uncovered in this analysis, over 70% exhibit sequence similarity outside the kinase or phosphatase domain to proteins in other species. In the kinase group, approximately 75% are serine/threonine kinases, and 25% are tyrosine or dual-specificity kinases. Over 90% of the newly discovered kinases are predicted to phosphorylate serine/threonine residues; this group includes the first atypical protein kinase C isoforms identified in *Drosophila*. In addition, we found counterparts of the mammalian kinases CSK, MLK2, ATM, and Peutz-Jeghers syndrome kinase, and additional members of the *Drosophila* GSK3B, casein kinase I, SNF1-like, and Pak/STE20-like kinase families. In the fly protein phosphatase group, approximately 42% are predicted to be serine/threonine phosphatases; 48% are tyrosine or dual-specificity phosphatases. Among the newly discovered phosphatases, 35% are serine/threonine phosphatases, most of which are related to the protein phosphatase 2C family, and 65% are tyrosine or dual-specificity phosphatases. The fly and worm both contain close relatives to many of the known mammalian lipid kinases and phosphatases; however, no SH2-containing inositol 5′ phosphatase SHIP is apparent. Finally, it has been found that the assembly of kinase signaling complexes in vertebrate cells is aided by the presence of scaffolding and adaptor molecules, many of which contain phosphoprotein binding domains; we found 85 such proteins in the fly, including counterparts to IRS, VAV, SHC, JIP, and MP1.

Table 2A. Similarity of sequences in predicted proteomes of *D. melanogaster, S. cerevisiae*, and *C. elegans*. To be scored as a similarity, each pairwise similarity was required to extend over more than 80% of the length of the query sequence at an E value less than that indicated. For example, in "Fly proteins in Fly-yeast," the column labeled $E < 10^{-10}$ shows the number and percentage of fly proteins that match yeast proteins at this E value or less and for which more than 80% of the length of the fly protein is aligned with the yeast protein. Each set of pairs was analyzed without consideration of the third proteome. The rows labeled "Fly-worm-yeast" report the composition of an independent clustering in which only groups containing a member from all three proteomes were counted. The numbers are slightly higher for the "Fly-worm-yeast" counts than for the "Fly-yeast" or "Worm-yeast" counts because of sequence bridging; that is, not all sequences within a group necessarily have a significant match to all other members of that group. See (*6*) for details.

	$E < 10^{-10}$		$E < 10^{-20}$		$E < 10^{-50}$		$E < 10^{-100}$	
	(*n*)	(%)	(*n*)	(%)	(*n*)	(%)	(*n*)	(%)
Fly proteins in:								
Fly-yeast	2345	16.5	1877	13.2	1036	7.3	433	3.1
Fly-worm	4998	35.2	4212	29.7	2442	17.2	1106	7.8
Fly-worm-yeast	3303	23.3	2428	17.1	1113	7.8	435	3.1
Worm proteins in:								
Worm-yeast	2184	11.8	1768	9.5	933	5.0	374	2.0
Fly-worm	4795	25.8	4004	21.6	2403	12.9	1092	5.9
Fly-worm-yeast	3229	17.4	2439	13.1	1115	6.0	419	2.3
Yeast proteins in:								
Fly-yeast	1856	29.4	1567	24.8	891	14.1	376	6.0
Worm-yeast	1704	27.0	1425	22.6	802	12.7	335	5.3
Fly-worm-yeast	1833	29.1	1525	24.2	831	13.2	352	5.6

Table 2B. A comparison of *D. melanogaster, C. elegans*, and *S. cerevisiae* protein sequences to each other and to mammalian sequences (*64*). This table reports the number and percent of fly, worm, or yeast query sequences with similarities less than the indicated E value cutoffs. For example, in the "Fly vs. Yeast" comparison, 3986 or 28.1% of fly proteins have a similarity with a yeast protein with an E value less than 1×10^{-10}. EST E values are not directly comparable to protein E values, because the resulting alignments are shorter.

	No similarity $E > 10^{-4}$		$E < 10^{-10}$		$E < 10^{-20}$		$E < 10^{-50}$		$E < 10^{-100}$	
	(*n*)	(%)	(*n*)	(%)	(*n*)	(%)	(*n*)	(%)	(*n*)	(%)
Fly vs.										
Yeast	8177	57.6	3986	28.1	2677	18.9	1266	8.9	504	3.6
Worm	5110	36.0	6743	47.5	5180	36.5	2832	19.9	1197	8.4
Mammalian	5833	41.1	7032	49.5	5837	41.1	3580	25.2	1772	12.5
Mammalian ESTs	5386	37.9	7329	51.6	5352	37.7	1775	12.5	110	0.8
Worm vs.										
Yeast	12541	68.0	3582	19.4	2378	12.9	1106	6.0	401	2.2
Fly	8603	46.7	7138	38.8	5428	29.5	2880	15.6	1229	6.7
Mammalian	10152	55.1	6550	35.6	4999	27.1	2782	15.1	1211	6.6
Mammalian ESTs	10354	56.2	6005	32.6	4000	21.7	1170	6.4	68	0.4
Yeast vs.										
Fly	2614	41.9	2564	41.0	1910	30.6	1021	16.4	408	6.5
Worm	2762	44.2	2358	37.8	1730	27.7	882	14.1	348	5.6
Mammalian	3230	51.7	2340	37.5	1802	28.9	992	15.9	429	6.9
Mammalian ESTs	3106	49.7	2319	37.1	1553	24.9	503	8.1	18	0.3

Two remarkable findings emerge from the peptidase data that may reflect different approaches to growth and development in flies, worms, and humans. The pattern and distribution of peptidase types are similar between the fly and the worm: there are approximately 450 peptidases in the fly and 260 in the worm. The difference is due almost entirely to the expansion or contraction of a single class of trypsin-like (S1) peptidases. *C. elegans* has seven of this class and yeast has one, but the fly has 199. Of these, 163 are small proteins of approximately 250 amino acids containing single trypsin domains; very few are mosaic proteins. The remainder have either multiple trypsin-like domains or long stretches of amino acids with no readily identifiable motif, usually at the NH_2-terminus. In humans, trypsin-like peptidases perform diverse functions in digestion, in the complement cascade, and in several other signaling pathways (*11*), and flies may have a similarly wide range of uses for these proteins. The extensively characterized members of this family, which include Snake, Easter, Nudel, and Gastrulation-defective, are all key members of a regulatory cascade that controls dorsoventral patterning in the fly (*12*). In addition, flies have only two members of the M10 class of peptidases, which include the matrix metalloproteases, collagenases, and gelatinases that are essential for tissue remodeling and repair in vertebrates.

The number of identifiable multidomain proteins is similar in the fly and the worm: 2130 and 2261, respectively. Yeast has only 672 (Table 5). Part of this difference is ac-

counted for by proteins with extracellular domains involved in cell-cell and cell-substrate contacts (*13*), such as the immunoglobulin domain–containing proteins, which are more abundant in flies than in worms (153 versus 70) and are nonexistent in yeast. Two other common extracellular domains occur in similar numbers in fly and worm: EGF (110 versus 109, respectively) and fibronectin type III (46 versus 43) but are rare or absent in yeast. Extracellular regions of proteins often contain a variety of repeated domains (*14*), and so these proteins may account for our finding that flies have a larger number of proteins with multiple InterPro domains than either worms or yeast (2107 versus 1747 and 525, respectively) (Table 6). Some multidomain proteins of the fly are particularly heterogeneous: Two low-density lipoprotein receptor–related proteins have 75 InterPro domains each. Another protein of unknown function has 62 InterPro domains; the most heterogeneous worm and yeast proteins [SWISS-PROT/TrEMBL accession numbers (AC), Q04833 and P32768, respectively] have 61 and 18 InterPro domains, respectively. There can be extensive repetition of the same domain within a protein; for example, an immunoglobulin-like domain is repeated 52 times within one protein of unknown function in the fly. The large worm protein UNC-89 contains 48 immunoglobulin-like domains (SWISS-PROT/TrEMBL AC, Q17362). In contrast, the largest number of repeats in yeast, of a C2H2-type zinc finger domain, occurs nine times in the transcription factor TFIIIA (SWISS-PROT/TrEMBL AC, P39933).

The heterotrimeric GTP-binding protein (G protein)–coupled receptors (GPCRs) are a large protein family in flies, worms, and vertebrates whose members are involved in synaptic function, hormonal physiology, and the regulation of morphological movements during gastrulation and germ band extension (*15*). There are predicted to be at least 700 GPCRs in the human genome (*16*) and roughly 1100 GPCRs in *C. elegans* (*17*). We found approximately 160 GPCR genes in the *Drosophila* genome, 57 of which appear to be olfactory receptors. *Drosophila*, *C. elegans,* and vertebrates each have diverse families of odorant receptors that, although recognizable as GPCRs, are unrelated by sequence and therefore apparently evolved independently. The number of odorant receptors in vertebrates ranges from around 100 in zebrafish and catfish to approximately 1000 in the mouse; *C. elegans* also has approximately 1000. In the fly, as in zebrafish and mouse, there is a correlation between the number of odorant receptors and the number of discrete synaptic structures called glomeruli in the olfactory processing centers of the brain (*16*, *18*). In the mouse, each glomerulus is dedicated to receiving axonal input from neurons expressing a particular odorant receptor (*16*). Therefore, the correlation between number of odorant receptors and number of glomeruli may reflect a conservation in the organizational logic of odor recognition in insect and vertebrate brains. Although the fly odorant receptors are extremely diverse, there are a number of subfamilies whose members share 50 to 65% sequence identity. The distribution of odorant receptor genes is different among these organisms as well. Unlike *C. elegans* or vertebrate odorant receptors, which are in large linked arrays, the fly odorant receptor genes are distributed as single genes or in arrays of two or three. Vertebrate receptors are encoded by intronless genes, but both fly and worm receptor genes have multiple introns. These distinctions suggest that in addition to differences in the sequences of the odorant receptors of the different organisms, the processes generating the families of receptors may have differed among the lineages that gave rise to flies, worms, and vertebrates.

The data suggest conservation of hormone receptors between flies and vertebrates; nevertheless, there is a greater diversity of hormone receptors in both *C. elegans* and vertebrates than in *Drosophila*. Insects are subject to complex hormonal regulation, but no apparent homologs of vertebrate neuropeptide and hormone precursors were identified. However, many receptors with sequence similarity to vertebrate receptors for neurokinin, growth hormone secretagogue, leutotropin (follicle-stimulating hormone and luteinizing hormone), thyroid-stimulating hormone, galanin/allatostatin, somatostatin, and vasopressin were identified. Other GPCRs include a seventh *Drosophila* rhodopsin and homologs of adenosine, metabotropic glutamate, γ-aminobutyric acid (GABA), octopamine, serotonin, dopamine, and muscarinic acetylcholine receptors. In addition, there are GPCRs that are unique to *Drosophila*, others with sequence similarity to *C. elegans* and human orphan receptors, and an insect diuretic hormone receptor that is closely related to vertebrate corticotropin-releasing factor receptor. Finally, we found several atypical seven-transmembrane domain receptors, including 10 Methuselah (MTH)–like proteins and four Frizzled (FZ)–like proteins. A mutation in *mth* increases the fly's life-span and its resistance to various stresses (*19*); the FZ-like proteins probably serve as receptors for different members of the Wingless/Wnt family of ligands.

Human Disease Genes

Studies in model organisms have provided important insights into our understanding of genes and pathways that are involved in a variety of human diseases. In order to estimate the extent to which different types of human disease genes are found in flies, worms, and yeast, we compiled a set of 289 genes that are mutated, altered, amplified, or deleted in a diverse set of human diseases and searched for similar genes in *D. melanogaster*, *C. elegans*, and *S. cerevisiae*, as described in the legend to Fig. 1. Of these 289 human genes, 177 (61%) appear to have an ortholog in *Drosophila* (Fig. 1). Only proteins with similar domain structures were considered to be orthologs; this judgment was made by human inspection of the InterPro domain composition of the fly and human proteins. The importance of human inspection, as well as consideration of published information, is underscored by the fact that some sequences with extremely high similarity scores to proteins encoded by fly genes, such as LCK and Myotonic Dystrophy 1, were judged not to be orthologous, but others with relatively low scores, such as p53 and Rb1, were considered to be orthologs. We attempted this additional level of analysis only for the fly proteins, as the lower overall level of similarity of worm and yeast proteins made these subjective judgments even more difficult. Some of the human disease genes that are absent in *Drosophila* reflect clear differences in physiology between the two organisms. For instance, none of the hemoglobins, which are mutated in thalassemias, have orthologs in *Drosophila*. In flies, oxygen is delivered directly to tissues via the tracheal system rather than by circulating erythrocytes. Similarly, several genes required for normal rearrangement of the immunoglobulin genes do not have *Drosophila* orthologs.

Of the cancer genes surveyed, 68% appear to have *Drosophila* orthologs. In addition to previously described proteins, these searches identified clear protein orthologs for menin (MEN; multiple endocrine neoplasia type 1), Peutz-Jeghers disease (STK11), ataxia telangiectasia (ATM), multiple exostosis type 2 (EXT2), a second bCL2 family member, a second retinoblastoma family member, and a p53-like protein. Despite its relatively low sequence similarity to the human genes, the *Drosophila* gene encoding p53 was considered an ortholog because it shows a conserved organization of functional domains, and its DNA binding domain includes many of the same amino acids that appear to be hot spots for mutations in human cancer. Comparison of the fly p53-like protein with the human p53, p63, and p73 proteins suggests that it may represent a progenitor of this entire family. In mammalian cells, levels of p53 protein are tightly regulated in vivo by its interaction with the Mdm2 protein, which in turn binds to p19ARF (*20*). This mode of regulation, which modulates the activity of p53 but probably not of p63 or p73 (*21*), may not apply to the *Drosophila* protein, because we have not been able to identify orthologs of

Table 3. Number of proteins in *D. melanogaster* (F), *C. elegans* (W), and *S. cerevisiae* (Y) containing the 200 most frequently occurring protein domains in *D. melanogaster*. Domain identifiers are from InterPro (*9*), a new database that has begun to integrate the independent databases of localized protein sequence patterns into a single resource. The beta release used includes PROSITE, PRINTS, and PFAM. InterPro considers a signature to be true if its score is above a threshold specified for that signature by the individual database. Results of the InterPro analysis may differ from results obtained based on human curation of protein families, due to the limitations of large-scale automatic classifications. In some instances, different InterPro domains correspond to different features of proteins within the same family; for example, IPR001650 and IPR001410 (26 and 42 in the table). See (*62*) for live links to the InterPro database.

	Acc. No.	F	W	Y	Interpro Domain Name
1.	IPR000694	579	398	40	Proline-rich region
2.	IPR000822	352	138	47	Zinc finger, C2H2 type
3.	IPR000719	249	388	119	Eukaryotic protein kinase
4.	IPR001254	199	13	1	Serine proteases, trypsin family
5.	IPR001314	178	5	0	Chymotrypsin serine protease family (S1)
6.	IPR001680	167	95	90	G-protein beta WD-40 repeats
7.	IPR000504	160	92	55	RNA-binding region RNP-1 (RNA recognition motif)
8.	IPR000495	153	70	0	Immunoglobulins & major histocompatibility complex proteins
9.	IPR000345	145	17	7	Cytochrome c family heme-binding site
10.	IPR000379	140	112	38	Esterase/lipase/thioesterase
11.	IPR002290	138	171	110	Serine/Threonine protein kinases active-site
12.	IPR002048	130	79	16	EF-hand family
13.	IPR001356	113	88	10	Homeobox domain
14.	IPR000561	110	109	0	EGF-like domain
15.	IPR001611	108	48	7	Leucine-rich repeat
16.	IPR001841	105	113	35	Zinc finger, C3HC4 type (RING finger)
17.	IPR002356	100	335	0	G-protein coupled receptors, rhodopsin family
18.	IPR001066	97	54	46	Sugar transporter
19.	IPR001128	94	73	3	Cytochrome P450 enzyme
20.	IPR002110	90	77	19	Ankyrin-repeat
21.	IPR000618	87	0	0	Insect cuticle protein
22.	IPR001245	87	63	0	Tyrosine kinase catalytic domain
23.	IPR001440	82	46	34	TPR repeat
24.	IPR000130	79	19	8	Neutral zinc metallopeptidases, zinc-binding region
25.	IPR002380	78	41	22	Transforming protein P21 RAS
26.	IPR001650	76	66	75	DNA/RNA helicase domain (DEAD/DEAH box)
27.	IPR001617	72	56	32	ABC transporters family
28.	IPR001849	71	67	27	PH domain
29.	IPR001478	69	60	2	PDZ domain (also known as DHR or GLGF)
30.	IPR001488	69	8	5	Myc-type, helix-loop-helix dimerization domain signature
31.	IPR001051	67	61	38	ATP-binding transport protein, 2nd P-loop motif
32.	IPR001993	67	43	35	Mitochondrial energy transfer proteins
33.	IPR000734	66	9	4	Lipase
34.	IPR000210	64	103	1	Btb/ttk domain
35.	IPR000575	63	54	36	ATP/GTP-binding site motif A (P-loop)
36.	IPR001452	63	55	25	Src homology 3 (SH3) domain
37.	IPR001092	61	38	8	Helix-loop-helix DNA-binding domain
38.	IPR002198	61	63	13	Short-chain dehydrogenase/reductase (SDR) superfamily
39.	IPR002106	58	14	17	Aminoacyl-transfer RNA synthetases class-II
40.	IPR001806	51	46	23	Ras family
41.	IPR002347	50	22	1	Glucose/ribitol dehydrogenase family
42.	IPR001410	46	43	48	DEAD/DEAH box helicase
43.	IPR001777	46	43	2	Fibronectin type III domain
44.	IPR000169	43	22	1	Eukaryotic thiol (cysteine) proteases active sites
45.	IPR000521	42	44	6	Glutathione S-transferase
46.	IPR001622	42	91	1	Potassium channel
47.	IPR002557	42	6	0	Chitin binding domain
48.	IPR000051	40	38	21	SAM (and some other nucleotide) binding motif
49.	IPR002172	40	32	0	Low density lipoprotein (LDL)-receptor class A (LDLRA) domain
50.	IPR000063	38	32	12	Thioredoxin family
51.	IPR001623	38	29	22	DnaJ domain
52.	IPR002018	38	44	0	Carboxylesterases type-B
53.	IPR001304	37	165	0	C-type lectin domain
54.	IPR000387	36	83	12	Tyrosine specific protein phosphatase
55.	IPR000215	35	9	0	Serpins
56.	IPR001005	35	16	19	Myb DNA binding domain
57.	IPR001412	35	15	14	Aminoacyl-transfer RNA synthetases class-I
58.	IPR001939	35	27	29	AAA-protein (ATPases associated with various cellular activities)
59.	IPR001965	35	22	16	PHD-finger
60.	IPR000008	34	34	9	Protein kinase C2 domain
61.	IPR000608	34	18	16	Ubiquitin-conjugating enzymes
62.	IPR001781	34	33	4	LIM domain
63.	IPR000980	33	43	1	Src homology 2 (SH2) domain
64.	IPR002213	33	59	0	UDP-glucoronosyl & UDP-glucosyl transferases
65.	IPR000301	32	19	0	Transmembrane 4 family
66.	IPR000934	31	56	21	Serine/threonine specific protein phosphatase family
67.	IPR001251	31	16	6	CRAL/TRIO domain
68.	IPR001881	31	34	0	Calcium-binding EGF-like domain
69.	IPR002173	31	4	2	PfkB family of carbohydrate kinases
70.	IPR000194	30	5	22	ATP synthase alpha & beta subunits
71.	IPR000217	29	22	4	Tubulin family
72.	IPR000873	29	23	11	AMP-binding domain
73.	IPR000073	28	17	16	Alpha/beta hydrolase fold
74.	IPR000152	28	28	0	Aspartic acid & asparagine hydroxylation site
75.	IPR000408	28	6	3	Regulator of chromosome condensation (RCC1)
76.	IPR000834	28	9	1	Zinc carboxypeptidases, carboxypeptidase A metalloprotease (M14) family
77.	IPR001715	28	22	3	Calponin homology (CH) domain
78.	IPR002086	28	13	13	Aldehyde dehydrogenase family
79.	IPR002219	28	36	1	Phorbol esters/diacylglycerol binding domain
80.	IPR000483	27	7	0	Leucine rich repeat C-terminal domain
81.	IPR000886	27	8	11	Endoplasmic reticulum targeting sequence
82.	IPR001175	27	81	0	Neurotransmitter-gated ion-channel
83.	IPR000219	26	17	5	Dbl domain (dbl/cdc24 rhoGRF family)
84.	IPR000626	26	27	9	Ubiquitin domain
85.	IPR000629	26	22	20	ATP-dependent helicase, DEAD-box subfamily
86.	IPR000859	26	55	0	CUB domain
87.	IPR000958	26	21	6	KH domain
88.	IPR001752	26	22	6	Kinesin motor domain
89.	IPR002067	26	11	6	Mitochondrial carrier protein
90.	IPR000205	25	22	10	NAD binding site
91.	IPR000299	25	13	0	Band 4.1 family
92.	IPR000449	25	10	8	Ubiquitin-associated domain
93.	IPR000910	25	15	8	HMG1/2 (high mobility group) box
94.	IPR001054	25	32	1	Guanylate cyclase
95.	IPR001202	25	17	5	WW/rsp5/WWP domain
96.	IPR000595	24	19	2	Cyclic nucleotide-binding domain
97.	IPR000832	24	10	0	G-protein coupled receptors family 2 (secretin-like)
98.	IPR001140	24	30	10	ABC transporter transmembrane region
99.	IPR001214	24	27	6	SET-domain of transcriptional regulators (TRX, EZ, ASH1 etc)
100.	IPR001871	24	18	15	bZIP (Basic-leucine zipper) transcription factor family
101.	IPR002049	23	16	0	Laminin-type EGF-like (LE) domain
102.	IPR002111	23	21	2	Cation channels, 6TM region (transient receptor potential subtype)
103.	IPR000048	22	16	2	IQ calmodulin-binding domain
104.	IPR001353	22	12	14	Multispecific proteases of the proteasome
105.	IPR001810	22	215	11	F-box domain
106.	IPR002223	22	34	0	Pancreatic trypsin inhibitor (Kunitz) family
107.	IPR000718	21	29	0	Neprilysin metalloprotease (M13) family
108.	IPR000964	21	15	3	Sterile-alpha module (SAM) domain
109.	IPR001311	21	13	0	Solute binding protein/glutamate receptor domain
110.	IPR001394	21	24	18	Ubiquitin carboxyl-terminal hydrolases family 2
111.	IPR001594	21	13	6	DHHC-type Zn-finger
112.	IPR001628	21	224	0	C4-type steroid receptor zinc finger
113.	IPR002017	21	19	3	Spectrin repeat
114.	IPR002113	21	6	4	Adenine nucleotide translocator 1
115.	IPR002126	21	15	0	Cadherin domain
116.	IPR000195	20	17	12	RabGAP/TBC domain
117.	IPR000198	20	19	10	RhoGAP domain
118.	IPR000795	20	17	15	GTP-binding elongation factor
119.	IPR001930	20	11	4	Membrane alanyl dipeptidase, family M1
120.	IPR002422	20	14	7	Permeases for amino acids & related compounds, family II
121.	IPR000166	19	33	16	Histone-fold/TFIID-TAF/NF-Y domain
122.	IPR000690	19	8	7	RNA-binding protein C2H2 Zn-finger domain
123.	IPR001766	19	19	4	Fork head domain
124.	IPR002130	19	17	8	Cyclophilin-type peptidyl-prolyl cis-trans isomerase
125.	IPR002293	19	16	25	Permeases for amino acids & related compounds, family I
126.	IPR000175	18	12	0	Sodium:neurotransmitter symporter family
127.	IPR000330	18	20	17	SNF2 & others N-terminal domain

Table 3 (continued).

	Acc. No.	F	W	Y	Interpro Domain Name
128.	IPR000742	18	9	0	EGF-like domain, subtype 2
129.	IPR000961	18	24	10	Protein kinase C terminal domain
130.	IPR001173	18	17	4	Glycosyl transferase, family 2
131.	IPR000242	17	76	3	Tyrosine specific protein phosphatases
132.	IPR000467	17	11	4	D111 domain
133.	IPR000636	17	22	1	Cation channels, 6TM region (non-ligand gated)
134.	IPR000717	17	13	8	Domain in components of the proteasome, COP9-complex & eIF3 (PCI)
135.	IPR000953	17	15	2	Chromo domain
136.	IPR001071	17	0	0	Alpha-tocopherol transport protein
137.	IPR001163	17	11	16	Small nuclear ribonucleoprotein (Sm protein)
138.	IPR001327	17	4	4	FAD-dependent pyridine nucleotide reductase
139.	IPR001395	17	11	6	Aldo/keto reductase family
140.	IPR001734	17	3	1	Sodium:solute symporter family
141.	IPR001757	17	22	17	E1-E2 ATPases phosphorylation site
142.	IPR001791	17	16	0	Laminin-G domain
143.	IPR001873	17	22	0	Amiloride-sensitive sodium channel
144.	IPR001969	17	8	42	Eukaryotic & viral aspartyl proteases active site
145.	IPR000087	16	166	0	Collagen triple helix repeat
146.	IPR000253	16	6	16	Forkhead-associated (FHA) domain
147.	IPR000536	16	88	0	Ligand-binding domain of nuclear hormone receptor
148.	IPR001320	16	10	0	Ligand-gated ion channel
149.	IPR001487	16	13	10	Bromodomain
150.	IPR002027	16	11	24	Amino acid permease
151.	IPR002046	16	1	1	SAR1 GTP-binding protein family
152.	IPR000014	15	8	1	Generalized PAS domain
153.	IPR000172	15	1	0	GMC oxidoreductases
154.	IPR000251	15	12	7	ADP-ribosylation factors family
155.	IPR000569	15	5	5	HECT-domain (Ubiquitin-transferase)
156.	IPR000772	15	12	0	Lectin domain of ricin b-chain, 3 copies
157.	IPR001223	15	34	1	Glycosyl hydrolases family 18
158.	IPR001609	15	20	5	Myosin head (motor domain)
159.	IPR001828	15	19	0	Receptor family ligand binding region
160.	IPR002129	15	7	1	Pyridoxal-dependent decarboxylase family
161.	IPR002465	15	1	0	Growth factor & cytokine receptor family signature 2
162.	IPR000159	14	11	2	Ras-associated (RalGDS/AF-6) domain
163.	IPR000225	14	6	2	Armadillo/plakoglobin ARM repeat
164.	IPR000279	14	10	8	Actin
165.	IPR000566	14	6	0	Lipocalin & cytosolic fatty-acid binding protein
166.	IPR000577	14	3	2	Carbohydrate kinase, FGGY family
167.	IPR000746	14	0	0	Pheromone/general odorant binding protein, PBP/GOBP family
168.	IPR000884	14	27	0	Thrombospondin type I domain
169.	IPR001100	14	5	3	Pyridine nucleotide-disulfide oxidoreductase, class I
170.	IPR001159	14	9	2	Double-stranded RNA binding (DsRBD) domain
171.	IPR001199	14	8	5	Cytochrome B5
172.	IPR001357	14	21	11	BRCT domain
173.	IPR001589	14	8	1	Actinin-type actin-binding domain
174.	IPR001753	14	10	3	Enoyl-CoA hydratase/isomerase
175.	IPR001878	14	24	9	Zn-finger CCHC type
176.	IPR001952	14	0	1	Alkaline phosphatase family
177.	IPR002216	14	17	1	Ion transport protein
178.	IPR002464	14	9	8	DEAH-box subfamily ATP-dependent helicase
179.	IPR000107	13	8	3	SPRY domain
180.	IPR000425	13	8	6	MIP family
181.	IPR000508	13	2	3	Signal peptidase
182.	IPR000727	13	14	15	t-SNARE coiled-coil domain
183.	IPR000901	13	6	7	Carbamoyl-phosphate synthase
184.	IPR001461	13	16	7	Pepsin (A1) aspartic protease family
185.	IPR001506	13	36	0	Astacin (Peptidase family M12A) family
186.	IPR001523	13	11	0	'Paired box' domain
187.	IPR001827	13	2	0	'Homeobox' antennapedia-type protein
188.	IPR001876	13	7	1	Zn-finger in ranbp & others
189.	IPR002423	13	8	9	TCP-1 (Tailless complex polypeptide)/cpn60 chaperonin family
190.	IPR002893	13	8	1	MYND finger
191.	IPR000461	12	4	8	Alpha amylase
192.	IPR000798	12	4	0	Ezrin/radixin/moesin family
193.	IPR001023	12	13	14	Heat shock protein hsp70
194.	IPR001508	12	1	0	NMDA receptor
195.	IPR001683	12	9	15	PX (Bem1/NCF1/PI3K) domain
196.	IPR001917	12	6	4	Aminotransferases class-II
197.	IPR001932	12	9	8	Protein phosphatase 2C
198.	IPR000050	11	8	0	Phosphotyrosine interaction domain (PID)
199.	IPR000182	11	9	9	cetyltransferase (GNAT) family
200.	IPR000243	11	2	7	Proteasome B-type subunit

either Mdm2 or p19ARF in *Drosophila*. Interestingly, likely orthologs of the breast cancer susceptibility genes *BRCA1* and *BRCA2* were not found in *Drosophila*. In most instances, cancer genes that have a *Drosophila* ortholog also have an ortholog in *C. elegans*, although the extent of sequence similarity to the worm gene is lower. In a minority of instances, a *C. elegans* ortholog was clearly absent. Cancer genes with orthologs in *Drosophila* and apparently not in *C. elegans* include *p53* and *neurofibromatosis type 1* (*22*), the two genes implicated in tuberous sclerosis (*TSC1* and *TSC2*) (*23*), and *MEN*. The two TSC gene products are thought to bind to each other and may function in a pathway that is conserved between humans and *Drosophila* but is absent in *C. elegans* and *S. cerevisiae*. However, the limitations of this type of analysis are clearly illustrated by our inability to find a *bCL2* ortholog in *C. elegans* using these search parameters. The *C. elegans ced-9* gene has been shown to function as a *bCL2* homolog, and its protein is 23% identical to the human protein over its entire length (*24*).

Numerous orthologs of neurological genes are also found in the *Drosophila* genome. Some, such as *Notch* (CADASIL syndrome), the *beta amyloid protein precursor-like* gene, and *Presenilin* (Alzheimer's disease), were already known from previous studies in the fly. The genome sequencing effort has uncovered several additional genes that are likely to be orthologs of human neurological genes, such as *tau* (frontotemporal dementia with Parkinsonism), the Best macular dystrophy gene, *neuroserpin* (familial encephalopathy), genes for limb girdle muscular dystrophy types 2A and 2B, the Friedreich ataxia gene, the gene for Miller-Dieker lissencephaly, *parkin* (juvenile Parkinson's disease), and the Tay-Sachs and Stargardt's disease genes. Several genes implicated in expanded polyglutamine repeat diseases, including Huntington's and spinal cerebellar ataxia 2 (*SCA2*), are found in the fruit fly. Most human neurological disease genes surveyed were also detected in *C. elegans*, and some were even found in yeast, although a few examples are apparently present only in *Drosophila*, such as the *Parkin* and *SCA2* orthologs.

Among genes implicated in endocrine diseases, those functioning in the insulin pathway are mostly conserved. In contrast, members of pathways involving growth hormone, mineralocorticoids, thyroid hormone, and the proteins that regulate body mass in vertebrates, such as those encoding leptin, do not appear to have *Drosophila* orthologs. Surprisingly, a protein that shows significant sequence similarity to the luteinizing hormone receptor is present in *Drosophila* (*25*). The physiological ligand for this receptor is not known. A number of genes that have been implicated in human renal disorders have orthologs in *Drosophila*, despite the differences between human kidneys and insect Malpighian tubules. In many instances, these gene products are involved in fluid and electrolyte transport across epithelia. Not surprisingly, most disease genes that function in intracellular metabolic pathways appear to have *Drosophila* orthologs.

Developmental and Cellular Processes

Developmental strategies in various phyla are overtly very different, from the fixed cell lineage of *C. elegans* to the syncytial embryogenic development of the fly, to early embryogenesis in amphibians and mammals. A number of major processes—cell division, cell shape, signaling pathways, cell-cell and

cell-substrate adhesion, and apoptosis—determine the developmental outcomes of these very different embryos. Although there are many more, such as the processes that determine embryonic gradients, cell polarities, and cell movement, here we examine the first five, beginning with cell cycle components, and examine what new insights have been gained from the genomic data that affect our knowledge of the evolution of developmental processes. We then discuss the processes of neuronal signaling and innate immunity.

Cell cycle. Despite conservation of the mechanisms regulating cell cycle progression, many of the functions governing this progression are encoded by gene families whose individual members are not conserved between vertebrates and yeast. For example, the cyclins of *S. cerevisiae* can be divided into a G_1 class (Cln1, Cln2, and Cln3) and an S/G_2 class (Clb1 through Clb6); it is not possible to identify orthologs of individual vertebrate cyclins. Consequently, analysis of the roles of particular vertebrate cell cycle genes benefits from a genetic model in which parallels are more evident. Analysis of the *Drosophila* genome sequence supports and extends previous suggestions of strong parallels between fly and human cell cycle regulators. Orthologs of vertebrate cell cycle cyclins—cyclin A (CycA), CycB, CycB3, CycE, and CycD—have been identified in *Drosophila*, as have orthologs of cyclins that appear to have roles in transcription: CycC, CycH, CycK, and CycT. Apparent orthologs of these cyclins can be also be found in *C. elegans*; however, the level of similarity to the vertebrate members is invariably substantially less. Indeed, BLAST comparisons suggest that vertebrate and *Drosophila* CycA and CycB share more sequence similarity with yeast than with proposed *C. elegans* orthologs. Examination of other cell cycle regulators confirms that quite precise comparisons can be made between vertebrates and flies; parallels with yeast are looser. For example, like vertebrates, *Drosophila* uses several different cyclin-dependent kinases (Cdks) to regulate different aspects of the cell cycle; *S. cerevisiae* and *Schizosaccharomyces pombe* use only one. Cloning efforts and the genome sequence revealed *Drosophila* orthologs of vertebrate Cdk1 (*cdc2*) and Cdk2 (*cdc2c*), as well as a single *Drosophila* Cdk (*Cdk4/6*) with close similarity to both Cdk4 and Cdk6. As in vertebrates, *Drosophila* has two distinct kinases that add inhibitory phosphate to Cdk1, the previously identified Wee, and a recently recognized homolog of Myt1, which was initially identified as a membrane-associated inhibitory kinase in *Xenopus* (*26*). *C. elegans* also has two homologs of these kinases (Wee1.1 and Wee1.3); however, similarity scores do not place these into distinct Wee1 and Myt1 subtypes. Each of these genes appears to be present in a single copy, a factor that simplifies genetic interpretations.

Fig. 1.

The retinoblastoma gene product pRb is a crucial cell cycle regulator in mammals and is thought to modulate S-phase entry via its interactions with the transcriptional regulator E2F and its dimerization partner (DP). This important mode of regulation is not found in yeast, but many components of the Rb pathway have been identified and studied in *Drosophila* (*27*). The sequencing effort uncovered a second *Rb*-related gene in *Drosophila* and confirmed the existence of only two E2F family members and a single DP ortholog. *C. elegans* also has an *Rb*-related gene, isolated in a genetic screen for mutations affecting cell fate decisions (*28*), but it has not been shown to play a direct role in cell cycle regulation. Also evident from the sequence are eight *skp*-like genes and six *cullin*-related genes. The Skp and Cullin proteins function in a complex that mediates the degradation of specific target proteins during crucial cell cycle transitions. Further exploration of the genome sequence should define orthologs to most vertebrate cell cycle genes and lead to genetic tests of their regulation and function.

Cytoskeleton. A large number of proteins link events at the cell surface with cytoskeletal networks and intracellular messengers (*13*). We found approximately 230 genes (approximately 2% of the predicted genes) that encode cytoskeletal structural or motor proteins; these represent most major families found in other invertebrates and vertebrates (*29*). The fraction of the *Drosophila* genome devoted to cytoskeletal functions appears to

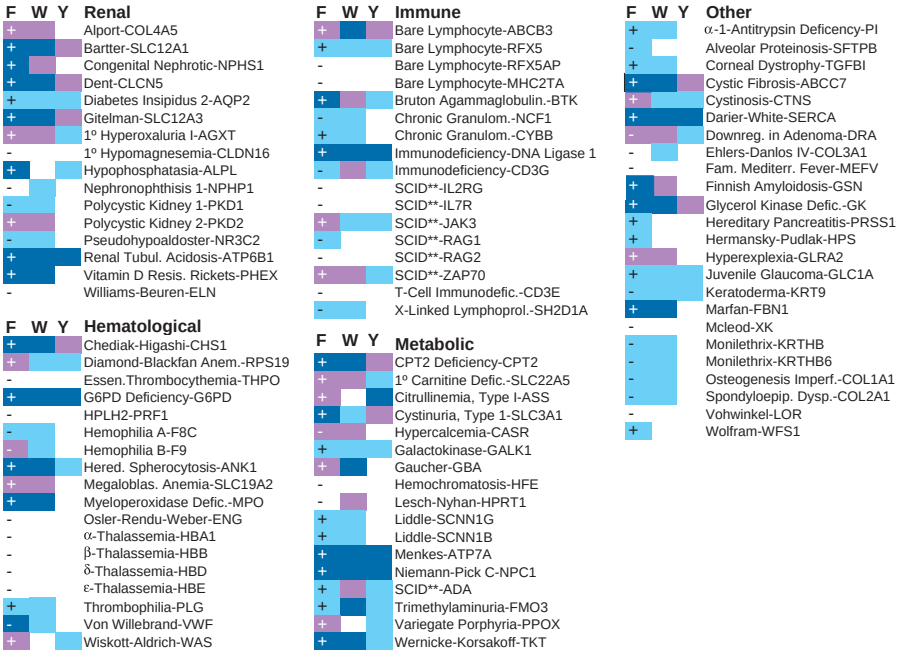

Fig. 1 (continued). Fly (F), worm (W), and yeast (Y) genes showing similarity to human disease genes. This collection of human disease genes was selected to represent a cross section of human pathophysiology and is not comprehensive. The selection criteria require that the gene is actually mutated, altered, amplified, or deleted in a human disease, as opposed to having a function deduced from experiments on model organisms or in cell culture. Due to redundancy in gene and protein sequence databases, a single reference sequence for each gene had to be chosen. Most reference sequences represent the longest mRNA of several alternatives in GenBank. Authoritative sources in the literature and electronic databases [Online Mendelian Inheritance in Man (OMIM)] were also consulted. In all, 289 protein sequences met these criteria. These were used as queries to search a database consisting of the sum total of gene products (38,860) found in the complete genomes of fly, worm, and yeast. 12,953 was used as the effective database size (the z parameter in BLAST). BLASTP searches were conducted as described for full genome searches, except for the z parameter. To control for potential frameshift errors in the *Drosophila* genome sequence, searches against a six-frame translation of the entire genome (using TBLASTN) were also conducted with the disease gene sequences using the z parameter above. Only two cases in which matches to genomic sequence were better than to the predicted protein were found, and these were manually corrected to reflect the better TBLASTN scores in the table. Results are scaled according to various levels of statistical significance, reflecting a level of confidence in either evolutionary homology or functional similarity. White boxes represent BLAST E values $>1 \times 10^{-6}$, indicating no or weak similarity; light blue boxes represent E values in the range of 1×10^{-6} to 1×10^{-40}; purple boxes represent E values in the range of 1×10^{-40} to 1×10^{-100}; and dark blue boxes represent E values $<1 \times 10^{-100}$, indicating the highest degree of sequence conservation. Actual E values can be found in the Web supplement to this figure (*62*), where links to OMIM and GenBank may also be found. A plus sign indicates our best estimate that the corresponding *Drosophila* gene product is the functional equivalent of the human protein, based on degree of sequence similarity, InterPro domain composition, and supporting biological evidence, when available. A minus sign indicates that we were unable to identify a likely functional equivalent of the human protein.

be somewhat smaller than that found in *C. elegans* (5%) (*30*); whether this reflects a true biological difference or a difference in classification criteria remains to be discovered. Of the *Drosophila* cytoskeletal genes, 90 encode proteins belonging to the kinesin, dynein, or myosin motor superfamilies, or accessory or regulatory proteins known to interact with the motor protein subunits. Approximately 80 genes encode actin-binding proteins, including proteins belonging to the spectrin/α-actinin/dystrophin superfamily of membrane cytoskeletal and actin–cross-linking proteins. Twenty genes encode proteins that are likely to bind microtubules, based on their similarity to microtubule-binding proteins found in other organisms. Fourteen genes encode members of the actin superfamily, 12 encode members of the tubulin superfamily, and 5 encode septins. Overall, the representation of predicted cytoskeletal protein types and families is similar to what has been found for *C. elegans*, although *Drosophila* has many more dyneins, probably because *C. elegans* lacks motile cilia and flagella.

Among this collection of cytoskeletal genes are several interesting and in some cases long-sought genes. One gene encodes a protein with striking homology to proteins of the tau/MAP2/MAP4 family that share a characteristic repeated microtubule-binding domain. Two encode new tubulins; one appears most closely related to α-tubulin, and the other appears most closely related to β-tubulin, both with approximately 50% identity. Neither new tubulin has greater similarity to the other, more divergent members of the tubulin superfamily, such as γ-, δ-, or ε-tubulin (*31*). Thus, both *Drosophila* and *C. elegans* appear to lack δ- and ε-tubulin, even though δ-tubulin is highly conserved between *Chlamydomonas* and humans. There are also three new members of the central motor domain family of kinesins that encode nonmotor proteins that regulate microtubule dynamics (*32*). There are clear homologs of the dystrophin complex and of dystrobrevin. Finally, the fly lacks cytoplasmic intermediate filament proteins, other than nuclear lamins, although other invertebrates, including *C. elegans*, appear to have genes encoding these (*33*). *Drosophila* and *C. elegans* both also appear to lack a gene encoding kinectin, the proposed receptor for kinesin and cytoplasmic dynein on vesicles and organelles (*34*). Flies and worms must thus use different proteins to link microtubule motors to vesicles and organelles.

Cell adhesion. Cell-cell adhesion and cell-substrate adhesion molecules have been crucial to the development of multicellular organisms and the evolution of complex forms of embryogenesis (*13*). The transmembrane extracellular matrix-cytoskeleton linkage via integrins is ancient. There are five α and two β integrins in the fly, two α and one β in *C. elegans*, and at least 18 α and eight β in vertebrates. Integrin-associated cytoplasmic proteins (talin, vinculin, α-actinin, paxillin, FAK, p130CAS, and ILK) are encoded by single-copy fly genes, as are tensin and syndecan.

Two genes for type IV collagen subunits and genes for the three subunits of laminin were already known in the fly. Analysis of the genome revealed no more laminin genes and only one more collagen, which is closest to types XV and XVIII of vertebrates. A counterpart of this collagen is found in *C. elegans*, which has on the order of 170 collagens. Most important, it appears that the core components of basement membranes (two type IV collagen subunits, three laminin subunits, entactin/nidogen, and one perlecan), are all present in flies. This constitution of basement membranes was clearly established early in evolution and has been well conserved in metazoans; remarkably, the fly preserves the linked head-to-head organization of vertebrate type-IV collagen genes. In contrast to this conservation, many well-known vertebrate integrin (ECM) ligands are absent from the fly: fibronectin, vitronectin, elastin, von Willebrand factor, osteopontin, and fibrillar collagens are all missing.

The fly has three classic cadherins, two of which are closely linked, but no protocadherins of the type found in vertebrates as clusters with common cytoplasmic domains (*35*). Vertebrates have three such clusters encoding over 50 protocadherins and close to 20 classical

cadherins. The fly has no reelin, an ECM ligand for CNR-type protocadherins in vertebrates (*36*). However, there are other fly proteins with cadherin repeats, including the previously known Fat, Dachsous, and Starry night, and a new very large protein related to Fat. *C. elegans* has 15 genes containing cadherin repeats; the number in humans is now 70 and will undoubtedly rise (*13*).

Cell signaling. Components of known signaling pathways in the fly and worm have largely been uncovered by examinations of developmental systems. It is a tribute to the previous genetic analyses done in these organisms that only a modest number of new components of the known signaling pathways were revealed by analysis of the genomic sequence. The core components defined in flies and worms have been used in modified and expanded forms in vertebrates (*37*). The predominant pathways—transforming growth factor–β (TGF-β), receptor tyrosine kinases, Wingless/Wnt, Notch/lin-12, Toll/IL1, JAK/STAT/cytokine, and Hedgehog (HH) signaling networks—all have largely conserved fly and vertebrate components. The worm, by contrast, does not appear to possess the HH or Toll/IL1 pathways, nor does it have all of the components of the Notch/lin-12 network (*38*). Two new proteins of the TGF-β superfamily were identified, bringing the total to seven; all seven are members of the bone morphogenetic protein (BMP) or β-activin subfamilies. We detected no representatives of the other branches of this superfamily, namely the TGF-β, α-inhibin, and Mullerian inhibiting substance (MIS) subfamilies. Three new members of the Wingless/Wnt family were identified, bringing the total to seven. Each of these proteins has sequence similarity to a different vertebrate Wnt protein; this ancient family clearly underwent much of its expansion before the divergence of the arthropod and chordate lineages. There is only one member of the Notch and HH families, in contrast to the many members of these families in vertebrates.

Apoptosis. The core apoptotic machinery of *Drosophila* shares many features in common with that of mammals. Many apoptosis-inducing signals lead to activation of members of the caspase family of proteases. These proteases function in apoptotic processes as cell death signal transducers and death effectors, and in nonapoptotic processes in flies and mammals (*39*). *Drosophila* contains genes encoding 8 caspases, as compared to 4 in the worm and at least 14 in mammals. Three of the fly caspases contain long NH_2-terminal prodomains of 100 to 200 amino acids that are characteristic of caspases that function as signal transducers. These prodomains are thought to mediate caspase recruitment into signaling complexes in which activation occurs in response to oligomerization. In one pathway described in mammals but not in worms, death signals cause the release of proteins, including cytochrome c and the apoptosis-inducing factor (AIF), from mitochondria (*40*). The human protein Apaf-1, in conjunction with cytochrome c, activates CARD domain–containing caspases (*41*). *Drosophila* has an Apaf-1 counterpart, a CARD domain–containing caspase, and AIF; *Drosophila* also has counterparts to the caspase-activated DNAse CAD/CPAN/DFF40, its inhibitor ICAD/DFF45, and the chromatin condensation factor Acinus (*42*).

Table 4. The 10 InterPro protein domains occurring in the largest number of different proteins in *S. cerevisiae* and *C. elegans*.

Acc. no.	InterPro domain name	No. of proteins
	S. cerevisiae	
IPR000719	Eukaryotic protein kinase	119
IPR001680	G-protein beta WD-40 repeats	90
IPR001650	DNA/RNA helicase domain (DEAD/DEAH box)	75
IPR001138	Fungal transcriptional regulatory protein, N-terminus	60
IPR001042	TYA transposon protein	57
IPR000504	RNA-binding region RNP-1 (RNA recognition motif)	55
IPR001410	DEAD/DEAH box helicase	48
IPR000822	Zinc finger, C2H2 type	47
IPR001066	Sugar transporter	46
IPR001969	Eukaryotic and viral aspartyl proteases active site	42
	C. elegans	
IPR000168	7-Helix G-protein coupled receptor, nematode (probably olfactory) family	545
IPR000694	Proline-rich region	398
IPR000719	Eukaryotic protein kinase	388
IPR002356	G-protein–coupled receptors, rhodopsin family	335
IPR001628	C4-type steroid receptor zinc finger	224
IPR001810	F-box domain	215
IPR000087	Collagen triple helix repeat	166
IPR001304	C-type lectin domain	165
IPR002900	Domain of unknown function	142
IPR000822	Zinc finger, C2H2 type	138

Pro- and anti-apoptotic BCL2 family members regulate apoptosis at multiple points (*43*). *Drosophila* encodes two BCL2 family proteins, though more divergent family members may exist. Fifteen BCL2 family proteins have been identified in mammals and two in the worm. In addition, inhibitor of apoptosis (IAP) family proteins negatively regulate apoptosis (*44*). They are defined by the presence of one or more NH_2-terminal repeats of a BIR domain, a motif that is essential for death inhibition. *Drosophila* has four proteins with this motif, as compared to seven identified thus far in mammals. There are several BIR domain–containing proteins in *C. elegans* and yeast, but none has been implicated in cell death regulation. Reaper (RPR), Wrinkled (W), and Grim are essential *Drosophila* cell death activators (*45*). Orthologs have not been identified in other organisms, but they are likely to exist because RPR, W, and Grim induce apoptosis in vertebrate systems and physically interact with apoptosis regulators that include IAPs and the *Xenopus* protein Scythe (*46*), for which there is a predicted *Drosophila* homolog.

Neuronal signaling. The neuronal signaling systems in flies, worms, and vertebrates reveal extensive conservation of some components, as well as extreme divergence, or the total absence, of others. There is no voltage-activated sodium channel in the worm (*17*); flies and vertebrates generate sodium-dependent action potentials. The fly genome encodes two pore-forming subunits for sodium channels (Para and NaCP60E), and also four voltage-dependent calcium channel α subunits, including one T-type/α1G, one L-type/α1D (Dmca1D), one N-type/α1A (Dmcα1A), and one protein that is more similar to an outlying *C. elegans* protein than to known vertebrate calcium channels. Additional fly calcium channel subunits include one β, one γ 2, and three α 2 subunits.

The worm genome encodes over 80 potassium channel proteins (*17*); the fly genome has only 30. The extent to which these different family sizes contribute to the establishment of unique electrical signatures is unknown. The fly potassium channel family includes five *Shaker*-like genes (*Shaker*, *Shab*, *Shal*, and two *Shaws*); a large conductance calcium-activated channel gene (*slowpoke*); a slack subunit relative; three members of the *eag* family (*eag*, *sei*, and *elk*); one small conductance calcium-regulated channel gene; one KCNQ channel gene; and four cyclic nucleotide–gated channel genes. In addition, there are 50 TWIK members in the worm, but only 11 fly members of the two-pore/TWIK family with four transmembrane domains. There are also three fly members of the inward rectifier/two transmembrane family. Finally, neither the fly nor the worm has discernible relatives of a number of mammalian channel-associated subunits such as minK and miRP1.

There are also major differences postsynaptically. *C. elegans* has approximately 100 members of a family of ligand-gated ion channels (*17*); flies have about 50. The worm has 42 nicotinic acetylcholine receptor subunits and 37 GABA(A)-like receptor subunits; the fly contains only 11 nicotinic receptor subunit genes and 12 GABA(A)/glycine-like receptor subunit genes. In contrast, there are 30 members of the excitatory glutamate receptor family in the fly but only 10 in the worm. These include subtypes of the AMPA, kainate, NMDA, and delta families. In addition, the fly genome contains a large number of PDZ-containing genes, approximately a dozen of which encode proteins that have high sequence similarity to mammalian proteins that interact with specific subsets of ion channels. We also found a number of additional ion channel families, including three voltage-dependent chloride channels, 14 Trp-like channels, 24 amiloride-sensitive/degenerin-like sodium channels, one ryanodine receptor, one IP_3 (inositol 1,4,5-trisphosphate) receptor, eight innexins, and two porins. *C. elegans* is missing a nitric oxide synthase gene, copies of which occur in fly and vertebrate genomes.

A large array of proteins mediates specific aspects of synaptic vesicle trafficking and contributes to the conversion of electrical signals to neurotransmitter release. These components of exocytosis and endocytosis are relatively well conserved with respect to both domain structures and amino acid identities (50 to 90%). The fly has enzymes for the synthesis of the neurotransmitters glutamate, dopamine, serotonin, histamine, GABA, acetylcholine, and octopamine, and a family of conserved transporters is likely to be involved in loading vesicles with these neurotransmitters. The conserved vesicular trafficking proteins, with 50 to 80% amino acid identity, include members of the Munc-18, SCAMP, synaptogyrin, HRS2, tomosyn, cysteine string protein, exocyst (SEC 5, 6, 7, 8, 10, 13, 15, EXO 70, and EXO84), synapsin, rabphilin-3A, RIM, rab-3, CAPS, Mint, Munc-13, NSF, α and γ SNAP, DOC-2B, latrophilin, Veli, CASK, VAP-33, Snapin, SV2, and complexin families. Generally, there is only one homolog in *Drosophila* for every three to four isoforms in mammals. However, there are eight fly synaptotagmin-like genes, making this the largest family of vesicle proteins in *Drosophila* (*47*). However, there is no homolog of synaptophysin, an early candidate for a vesicle fusion pore, which indicates a nonessential role in exocytosis for this particular protein across phyla.

Membrane trafficking also requires interactions between compartment-specific vesicular and target membrane proteins (v-SNAREs and t-SNAREs, respectively), whose subcellular distribution and combinatorial binding patterns are predicted to define organelle identity and targeting specificity (*48*). The completed fly genome allows us to address whether there is any correlation between the increased developmental complexity of multicellular organisms and a larger number of SNAREs than that found in unicellular organisms. In the fly, we find six synaptobrevins, three SNAP-25s, 10 syntaxins, and four additional t-SNAREs (membrin, BET1, UFE1, and GOS28), and the number of SNAREs is similar between yeast (*49*) and *Drosophila*. Thus, basic subcellular compartmentalization and membrane trafficking to and between these various compartments has not changed dramatically in multicellular versus unicellular organisms. Dynamin, clathrin, the clathrin adapter proteins, amphiphysin, synaptojanin, and a number of additional genes that encode proteins with defined endocytotic motifs are all present.

In contrast to the conservation of the synaptic vesicle trafficking machinery, the few identified proteins present at mammalian active zones, namely aczonin, bassoon, and piccolo, do not have relatives in *Drosophila*. There are, however, numerous proteins in the fly with combinations of C2 domains, PDZ domains, zinc fingers, and proline-rich domains, indicating that the precise protein composition of active zones is likely to vary among metazoans. In addition, *Drosophila* contains a neurexin III gene and four neuroligin genes that may be part of a neurexin-neuroligin complex that has been widely proposed to provide a synaptic scaffold for linking pre- and postsynaptic structures in mammals (*50*). Potential agrin and Musk genes are also present, though the overall sequence similarity is low.

Immunity. Multicellular organisms have elaborate systems to defend against microbial pathogens. Only vertebrates have an acquired immune system, but both vertebrates and invertebrates share a more primitive innate immune system. Innate immunity is based on the detection of common microbial molecules such as lipopolysaccharides and peptidoglycans by a class of receptors known as pattern recognition receptors (*51*). We identified a large family of genes encoding homologs of receptors that are involved in microbial recognition in other organisms. These include two new homologs of the *Drosophila* Scavenger Receptors (dSR-CI), nine members of the CD36 family, 11 members of the peptidoglycan recognition protein (PGRP) family, three Gram-negative binding protein (GNBP) homologs, and several lectins (*52*).

The recognition of infection by immunoresponsive tissues induces a battery of defense genes via Toll/nuclear factor kappa B (NF-κB) pathways in both *Drosophila* and mammals (*53*). The Toll receptor was initially discovered as an essential component of

Table 5. Proteins in *D. melanogaster*, *C. elegans*, and *S. cerevisiae* with more than one InterPro domain. These numbers represent the total number of recognizable domains within a single protein, no matter whether they are multiple copies of the same domain or different domains.

InterPro domains per protein	*D. melanogaster* (number of proteins)	*C. elegans* (number of proteins)	*S. cerevisiae* (number of proteins)
2	920	1236	410
3	388	458	121
4	219	182	58
5	163	98	26
6	101	72	17
7	92	53	15
8	58	27	7
9	42	25	4
10	22	18	7
11–15	73	43	6
16–20	18	17	1
21–30	22	22	0
31–50	8	5	0
51–75	4	5	0

Table 6. Proteins in *D. melanogaster*, *C. elegans*, and *S. cerevisiae* with multiple different InterPro domains. Individual InterPro domains are counted only once per protein, regardless of how many times they occur in that protein.

Unique InterPro domains per protein	*D. melanogaster* (number of proteins)	*C. elegans* (number of proteins)	*S. cerevisiae* (number of proteins)
2	1474	1248	402
3	413	335	95
4	156	114	23
5	52	38	4
6	8	9	1
7 or more	4	3	0

the pathway that establishes the dorsoventral axis of the *Drosophila* embryo. Recent genetic studies now reveal that Toll signaling pathways are key mediators of immune responses to fungi and bacteria in both *Drosophila* and mice (*53*). We found seven additional homologs of Toll proteins in *Drosophila*, all of which are more similar to each other than to their mammalian counterparts. Some of these other Toll proteins, like 18-wheeler, will probably mediate innate immune responses. In *Drosophila*, infection by at least some microbes induces a proteolytic cascade that leads to the processing of Spaetzle (SPZ), a cytokine-like protein, which then activates Toll (*53*). We found two proteins related to SPZ with similarities that include most or all of the cysteine residues of SPZ. Given the presence of multiple Toll-like receptors in *Drosophila*, these new SPZ-like proteins may also function in the immune system. With the exception of the two I-κB kinase homologs and the three rel proteins (Dorsal, Dif, and Relish), the *Drosophila* genome appears to contain only single copies of the genes encoding intracellular components of the Toll pathway: Tube, Pelle, and Cactus. How do the different Toll receptors trigger specific immune responses using the same intracellular intermediates? One explanation is that additional signaling components remain unidentified; another explanation is crosstalk with other signaling pathways. In contrast, a Toll ortholog has not been identified in *C. elegans*, although there are some Toll-like receptors. *C. elegans*, in addition, does not possess homologs of NF-κB/dorsal transcriptional activators that function downstream of Toll. Although it is probable that the worm has retained parts of the innate immunity network, there is no clear evidence of an inducible host defense system in the worm.

One of the most potent innate immune responses in insects is the transcriptional induction of genes encoding antimicrobial peptides (*53*). In contrast to Metchnikowin, Drosocin, and Defensin peptides, which are encoded by single genes, the sequence data indicate that, like the previously identifed cecropin clusters, several antimicrobial peptides are encoded by gene families that are larger than previously suspected. Four genes appear to encode antifungal peptide Drosomycin isoforms, and two genes each code for the antibacterial proteins Attacin and Diptericin. These additional genes may generate peptides with slightly different spectra of antimicrobial activity or may simply amplify the antimicrobial response.

Concluding Remarks

What have we learned about the proteins encoded by the three sequenced eukaryotic genomes? Some information emerges readily from the comparison of the fly, worm, and yeast genomes. First, the core proteome sizes of flies and worms are similar and are only twice the size of that of yeast. This is perhaps counterintuitive, because the fly, a multicellular animal with specialized cell types, complex development, and a sophisticated nervous system, looks more than twice as complicated as single-celled yeast. The lesson is that the complexity apparent in the metazoans is not achieved by sheer number of genes (*54*). Second, there has been a proliferation of bigger and more complex proteins in the two metazoans relative to yeast, including, not surprisingly, more proteins with extracellular domains involved in cell-cell and cell-substrate interactions. Finally, the population of multidomain proteins is somewhat larger and more diverse in the fly than in the worm. There is presently no practical way to quantify differences in biological complexity between two organisms, however, so it is not possible to correlate this increased domain expansion and diversity in the fly with differences in development and morphology.

The availability of the annotated sequence of the *Drosophila* genome enhances the fly's usefulness as an experimental organism. By greatly facilitating positional cloning, the genome sequence will increase the efficiency of genetic screens that seek to identify genes underlying many complex processes of cell biology, development, and behavior. Such screens have been the mainstay of *Drosophila* research and have contributed enormously to our knowledge of metazoan biology. The genome sequencing effort has revealed a number of previously unknown counterparts to human genes involved in cancer and neurological disorders; for example, *p53*, *menin*, *tau*, *limb girdle muscular dystrophy type 2B*, *Friedrich ataxia*, and *parkin*. All of these fly genes are present in a single copy in the genome and can be genetically analyzed without uncertainty about redundant copies. More genetic screens are important in order to uncover interacting network members. Orthologs of these network members can then be sought in the human genome to determine if alterations in any of them predispose humans to the disease in question, an experimental paradigm that has already been successfully executed in several cases. Flies can also play an important role in exploring ways to rectify disease phenotypes. For example, at least 10 human neurodegenerative diseases are caused by expansion of polyglutamine repeats (*55*). Human proteins containing expanded polyglutamine repeats have been expressed in flies, resulting in the formation of nuclear inclusions that contain the protein as well as other shared components (*56*), just as in humans. It has been shown that directed expression of the human HSP70 chaperone in the fly can totally suppress neurodegeneration resulting from expression of the human spinocerebellar ataxia type 3 protein (*57*). The power and speed of this in vivo system are unparalleled, and we anticipate the increased use of such "humanized" fly models.

Knowing the complete genomic sequence also allows new experimental approaches to long-standing problems. For example, it makes it possible to study networks of genes rather than individual genes or pathways. Assaying the level of transcription of every gene in the genome makes it at least theoretically possible to monitor the expression of an entire network of genes simultaneously. One problem that is approachable this way is the combinatorial control of gene transcription. The fly genome appears to encode only about 700 transcription factors, and mutations in over 170 have already been isolated and characterized. The techniques are available to measure the changes in expression of every gene in individual cell types as a consequence of loss or overexpression of each transcription factor. We can look for common sequence elements in the promoters of coregulated genes and perform chromatin immunoprecipitation to identify the in vivo binding sites of individual factors. For the first time, we can envision obtaining the data needed to understand the behavior of a complex regulatory network. Of course, collecting these data is a massive task, and developing methods to analyze the data is even more daunting. But it is no longer ludicrous to try.

How big is the core proteome of humans? Vertebrates have many gene families with three or four members: the HOX clusters, calmodulins, Ezrins, Notch receptors, nitric oxide synthases, syndecans, and NF1 transcription factor genes are some examples (*58*). This is evidence for two genome doublings during mammalian evolution, superimposed on which were the amplifications and contractions over evolutionary time that uniquely characterize each lineage (*59*). The human genome, with 80,000 or so genes, is likely to be an amplified version of a very much smaller genome, and its core proteome may not be much larger than that of the fly or worm; that is, the more complex attributes of a human being are achieved using largely the same molecular components. The evolution of additional complex attributes is essentially an organizational one; a matter of novel interactions that derive from the temporal and spatial segregation of fairly similar components.

Finally, approximately 30% of the predicted proteins in every organism bear no similarity to proteins in its own proteome or in the proteomes of other organisms. In other words, sequence similarity comparisons consistently fail to give us information about nearly a third of the components that make every organism uniquely itself. What does this mean with respect to the evolution and function of these proteins? Does each genome contain a subpopulation of very rapidly evolving genes? One-third of randomly chosen cDNA clones do

not cross-hybridize between *D. melanogaster* and *Drosophila virilis* (*60*). Even though these are distantly related species, they are developmentally and morphologically very similar. Crystallographic data will be needed to determine whether these proteins that have diverged in primary sequence have maintained their three-dimensional structures or have diverged so far that new folds and domains have formed.

Our first look at the annotated fly genome provokes these and other questions. Access to the genomic sequence will help us design the experiments needed to answer them. The relative simplicity and manipulability of the fly genome means that we can address some of these biological questions much more readily than in vertebrates. That is, after all, what model organisms are for.

References and Notes

1. M. D. Adams *et al.*, *Science* **287**, 2185 (2000); *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998); A. Goffeau *et al.*, *Science* **274**, 546 (1996).
2. R. D. Fleischman *et al.*, *Science* **269**, 496 (1995).
3. *C. elegans* data were taken from A C. Elegans Database (ACEDB) release WS8.
4. Local gene duplications were determined by searching for *N* similar genes within 2*N* genes on each arm. For example, if three similar genes are found within a region containing six genes, this counts as one cluster of three genes. Genes were judged to be similar if a BLASTP High Scoring Pair (HSP) with a score of 200 or more existed between them. Histone gene clusters were not included. *C. elegans* data were taken from ACEDB release WS8, containing 18,424 genes.
5. More information about GO is available at http://www.geneontology.org/. The Gene Ontology project provides terms for categorizing gene products on the basis of their molecular function, biological role, and cellular location using controlled vocabularies.
6. Initial results came from an NxN BLASTP analysis performed for each fly, worm, and yeast sequence in a combined data set of these completed proteomes. The databases used are as follows: Celera–Berkeley *Drosophila* Genome Project (BDGP), 14,195 predicted protein sequences (1/5/2000); WormPep 18, Sanger Centre, 18,576 protein sequences; and Saccharomyces Genome Database (SGD), 6306 protein sequences (1/7/2000). A version of NCBI-BLAST2 was used with the SEG filter and with the effective search space length (Y option) set to 17,973,263. Pairs were formed between every query sequence with a significant BLASTP to one of the other organisms' sequences. Significance was based on E-value cutoffs and length of match. These pairs were then independently grouped using single linkage clustering (*61*). Finally, the number of proteins from each proteome was counted. The requirement for 80% alignment of sequences makes this method of defining orthology particularly sensitive to errors that arise from incorrect protein prediction. However, the results comparing yeast and worm are essentially identical to those previously reported (*61*), even though the effective database size was different, the data sets have changed (Chervitz: yeast 6217 and worm 19,099; this study: yeast 6306, and worm 18,576), and the version of BLAST used is quite different (Chervitz: WashU BLAST 2.0a19MP; this study: NCBI BLAST 2.08).
7. A. Bairoch and R. Apweiler, *Nucleic Acids Res*. **28**, 45 (2000).
8. J. G. Henikoff, E. A. Greene, S. Pietrokovski, S. Henikoff, *Nucleic Acids Res*. **28**, 228 (2000).
9. InterPro (Integrated resource for protein domains and functional sites) is a collaborative effort of the SWISS-PROT, TrEMBL, PROSITE, PRINTS, Pfam, and ProDom databases to integrate the different pattern databases into a single resource. The database and a detailed description of the project can be found under http://www.ebi.ac.uk/interpro/. PROSITE is described in K. Hofmann, P. Bucher, L. Falquet, A. Bairoch, *Nucleic Acids Res*. **27**, 215 (1999); PFAM is described in A. Bateman *et al.*, *Nucleic Acids Res*.**27**, 260 (1999); and PRINTS is described in T. K. Attwood *et al.*, *Nucleic Acids Res*. **27**, 220 (1999).
10. G. D. Plowman, S. Sudarsanam, J. Bingham, D. Whyte, T. Hunter, *Proc. Natl. Acad. Sci. U.S.A*. **96**, 13603, (1999).
11. J. Barrett, N. D. Rawlings, J. F. Wessner, Eds., *Handbook of Proteolytic Enzymes* (Academic Press, San Diego, CA, 1998).
12. C. L. Smith and R. DeLotto, *Nature* **368**, 548 (1994); K. D. Konrad, T. J. Goralski, A. P. Mahowald, J. L. Marsh, *Proc. Natl. Acad. Sci. U.S.A*. **95**, 6819 (1998); E. K. LeMosy, C. C. Hong, C. Hashimoto, *Trends Cell Biol*. **9**, 102 (1999).
13. R. O. Hynes, *Trends Cell Biol*. **9**, M33 (1999).
14. P. Bork, A. K. Downing, B. Kieffer, I. D. Campbell, *Quart. Rev. Biophys*. **29**, 119 (1996).
15. P. Vernier, B. Cardinaud, O. Valdenaire, H. Philippe, J.-D. Vincent, *Trends Pharmacol. Sci*. **16**, 375, (1995); J. Colas, J. Launay, J. Vonesch, P. Hickel, L. Maroteaux, *Mech. Dev*. **87**, 77 (1999); M. R. Costa, E. T. Wilson, E. Wieschaus, *Cell* **76**, 1075 (1994).
16. P. Mombaerts, *Science* **286**, 707, (1999).
17. C. I. Bargmann, *Science* **282**, 2028 (1998).
18. P. J. Clyne *et al.*, *Neuron* **22**, 327 (1999); L. B. Vosshall, H. Amrein, P. S. Morozov, A. Rzhetsky, R. Axel, *Cell* **96**, 725 (1999); P. P. Laissue *et al.*, *J. Comp. Neurol*. **405**, 543 (1999).
19. Y. J. Lin, L. Seroude, S. Benzer, *Science* **282**, 943 (1998).
20. Y. Zhang, Y. Xiong, W. G. Yarbrough, *Cell* **92**, 725 (1998).
21. S. N. Jones, A. E. Roe, L. A. Donehower, A. Bradley, *Nature* **378**, 206 (1995).
22. I. The *et al.*, *Science* **276**, 791 (1997).
23. N. Ito and G. M. Rubin, *Cell* **96**, 529 (1999).
24. M. O. Hengartner and H. R. Horvitz, *Cell* **76**, 665 (1994).
25. F. Hauser, H. P. Nothacker, C. J. Grimmelikhuijzen, *J. Biol. Chem*. **272**, 1002 (1997).
26. P. R. Mueller, T. R. Coleman, A. Kumagai, W. G. Dunphy, *Science* **270**, 86 (1995).
27. B. D. Dynlacht, A. Brook, M. Dembski, L. Yenush, N. Dyson, *Proc. Natl. Acad. Sci. U.S.A*. **91**, 6359 (1994); W. Du, M. Vidal, J.-E. Xie, N. Dyson, *Genes Dev*. **10**, 1206 (1996); T. Sawado *et al.*, *Biochem. Biophys. Res. Commun*. **251**, 409 (1998).
28. X. Lu and H. R. Horvitz, *Cell* **95**, 981 (1998).
29. T. Kreis and R. Vale, Eds., *Guidebook to the Cytoskeletal and Motor Proteins* (Oxford Univ. Press, Oxford, 1999).
30. P. Chang and T. Stearns, *Nature Cell Biol*. **2**, 30 (2000).
31. S. K. Dutcher and E. C. Trabuco, *Mol. Biol. Cell* **9**, 1293 (1998).
32. A. Desai, S. Verma, T. J. Mitchison, C. E. Walczak, *Cell* **96**, 69 (1999).
33. K. Weber, in (*29*), pp. 291–293.
34. J. Kumar, H. Yu, M. P. Sheetz, *Science* **267**, 1834 (1995).
35. Q. Wu and T. Maniatis, *Cell* **97**, 779 (1999).
36. K. Senzaki, M. Ogawa, T. Yagi, *Cell* **99**, 635 (1999).
37. M. P. Belvin and K. V. Anderson, *Annu. Rev. Cell Dev. Biol*. **12**, 393 (1996); M. Hammerschmidt, A. Brook, A. P. McMahon, *Trends Genet*. **13**, 14 (1997); C. M. Blaumueller and S. Artavanis-Tsakonas, *Perspect. Dev. Neurobiol*. **4**, 325 (1997); T. Hunter, *Philos. Trans. R. Soc. London Ser. B* **353**, 583 (1998); K. M. Cadigan and R. Nusse, *Genes Dev*. **11**, 3286 (1997); J. Capdevila and J. C. Belmonte, *Curr. Opin. Genet. Dev*. **9**, 427 (1999); L. Engstrom, E. Noll, N. Perrimon, *Curr. Top. Dev. Biol*. **35**, 229 (1997); B. E. Stronach and N. Perrimon, *Oncogene* **18**, 6172 (1999); P. W. H. Holland, J. Garcia-Fernandez, N. A. Williams, A. Sidow, *Development* (suppl.) (1994), p. 125.
38. G. Ruvkun and O. Hobert, *Science* **282**, 2033 (1998).
39. W. C. Earnshaw, L. M. Martins, S. H. Kaufmann, *Annu. Rev. Biochem*. **68**, 383 (1999); A. Zeuner, A. Eramo, C. Peschle, R. DeMaria, *Cell Death Diff*. **6**, 1075 (1999).
40. X. Liu, C. N. Kim, J. Yang, R. Jemmerson, X. Wang, *Cell* **86**, 147 (1996); S. A. Susin *et al.*, *Nature* **397**, 441 (1999).
41. P. Li *et al.*, *Cell* **91**, 479 (1997).
42. A. G. Park, *Trends Cell Biol*. **10**, 394 (2000); S. Sahara *et al.*, *Nature* **401**, 168 (1999).
43. A. Gross, J. M. McDonnell, S. J. Korsmeyer, *Genes Dev*.**13**, 1899 (1999).
44. L. K. Miller, *Trends Cell Biol*. **9**, 323 (1999).
45. J. M. Abrams, *Trends Cell Biol*. **9**, 435 (1999).
46. K. Thress, W. Henzel, W. Shillinglaw, S. Kornbluth, *EMBO J*. **17**, 6135 (1998).
47. J. T. Littleton, T. L. Serano, G. M. Rubin, B. Ganetzky, E. R. Chapman, *Nature* **400**, 757 (1999).
48. T. Solner *et al.*, *Nature* **362**, 318 (1993).
49. R. Jahn and T. C. Sudhof, *Annu. Rev. Biochem*. **68**, 863 (1999).
50. K. Ichtchenko *et al.*, *Cell* **81**, 435 (1995).
51. R. Medzhitov and C. A. Janeway Jr., *Cell* **91**, 295 (1997).
52. A. Pearson, *Current Opin. Immunol*. **8**, 20 (1996); N. C. Franc *et al.*, *Immunity* **4**, 431 (1996); D. Kang *et al.*, *Proc. Natl. Acad. Sci. U.S.A*. **95**, 10078 (1998); W. J. Lee *et al.*, *Proc. Natl. Acad. Sci. U.S.A*. **93**, 7888 (1996).
53. J. A. Hoffmann and J.-M. Reichhart, *Trends Cell Biol*. **7**, 309 (1997); K. V. Anderson, *Curr. Opin. Immun*.**12**, 13 (2000).
54. G. L. G. Miklos, *J. Am. Acad. Arts Sci*. **127**, 197 (1998).
55. M. Perutz, *Trends Biochem. Sci* **24**, 58 (1999).
56. J. M. Warrick *et al.*, *Cell* **93**, 939 (1998); G. R. Jackson *et al.*, *Neuron* **21**, 633 (1998).
57. J. M. Warrick *et al.*, *Nature Genet*. **23**, 425 (1999).
58. J. Spring, *FEBS Lett*.**400**, 2 (1997).
59. S. Aparicio, *Trends Genet*. **16**, 54 (2000).
60. K. J. Schmid and D. Tautz, *Proc. Natl. Acad. Sci. USA*. **94**, 9746 (1997).
61. S. A. Chervitz *et al.*, *Science* **282**, 2022 (1998).
62. See www.sciencemag.org/feature/data/1049664.shl for complete protein domain analysis.
63. Paralogous gene families (Table 1) were identified by running BLASTP. A version of NCBI-BLAST2 optimized for the Compaq Alpha architecture was used with the SEG filter and the effective search space length (Y option) set to 17,973,263. Each protein was used as a query against a database of all other proteins of that organism. A clustering algorithm was then used to extract protein families from these BLASTP results. Each protein sequence constitutes a vertex; each HSP between protein sequences is an arc, weighted by the BLAST Expect value. The algorithm identifies protein families by first breaking all arcs with an E value greater than some user-defined value (1×10^{-6} was used for all of the analyses reported here). The resulting graph is then split into subgraphs that contain at least two-thirds of all possible arcs between vertices. The algorithm is "greedy"; that is, it arbitrarily chooses a starting sequence and adds new sequences to the subgraph as long as this criterion is met. An interesting property of this algorithm is that it inherently respects the multidomain nature of proteins: For example, two multidomain proteins may have significant similarity to one another but share only one or a few domains. In such a case, the two proteins will not be clustered if the unshared domains introduce a large number of other arcs.
64. An NxN BLASTP analysis was performed for each fly, worm, and yeast sequence in a combined data set of these completed proteomes. The databases used are as follows: Celera-BDGP, 14,195 predicted protein sequences (1/5/2000); WormPep18, Sanger Centre, 18,424 protein sequences; and SGD, 6246 protein sequences (1/7/2000). BLASTP analysis was also performed against known mammalian proteins (2/1/2000, GenBank nonredundant amino acid, Human, Mouse, and Rat, 75,236 protein sequences), and TBLASTN analysis was performed against a database of mammalian ESTs (2/1/00, GenBank dbEST, Human, Mouse, and Rat). A version of NCBI-BLAST2 optimized for the Compaq Alpha architecture was used with the SEG filter and the effective search space length (Y option) set to 17,973,263.

65. The many participants from academic institutions are grateful for their various sources of support. Participants from the Berkeley *Drosophila* Genome Project are supported by NIH grant P50HG00750 (G.M.R.) and grant P4lHG00739 (W.M.G.).

Comparative Genomics of the Eukaryotes

Gerald M. Rubin, Mark D. Yandell, Jennifer R. Wortman, George L. Gabor , Miklos, Catherine R. Nelson, Iswar K. Hariharan, Mark E. Fortini, Peter W. Li, Rolf Apweiler, Wolfgang Fleischmann, J. Michael Cherry, Steven Henikoff, Marian P. Skupski, Sima Misra, Michael Ashburner, Ewan Birney, Mark S. Boguski, Thomas Brody, Peter Brokstein, Susan E. Celniker, Stephen A. Chervitz, David Coates, Anibal Cravchik, Andrei Gabrielian, Richard F. Galle, William M. Gelbart, Reed A. George, Lawrence S. B. Goldstein, Fangcheng Gong, Ping Guan, Nomi L. Harris, Bruce A. Hay, Roger A. Hoskins, Jiayin Li, Zhenya Li, Richard O. Hynes, S. J. M. Jones, Peter M. Kuehl, Bruno Lemaitre, J. Troy Littleton, Deborah K. Morrison, Chris Mungall, Patrick H. O'Farrell, Oxana K. Pickeral, Chris Shue, Leslie B. Vosshall, Jiong Zhang, Qi Zhao, Xiangqun H. Zheng, Fei Zhong, Wenyan Zhong, Richard Gibbs, J. Craig Venter, Mark D. Adams and Suzanna Lewis

Science **287** (5461), 2204-2215.
DOI: 10.1126/science.287.5461.2204