Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 3731–3736, March 1998
Evolution

Data transferability from model organisms to human beings: Insights from the functional genomics of the *flightless* region of *Drosophila*

R. MALESZKA*, H. G. DE COUET†, AND GEORGE L. GABOR MIKLOS‡§

*Research School of Biological Sciences, Australian National University, Canberra ACT 2600, Australia; †Department of Zoology, University of Hawaii at Manoa, 2538 The Mall, Honolulu, HI 96822; and ‡Neurosciences Institute, 10640 John Jay Hopkins Drive, San Diego, CA 92121

Edited by Samuel Karlin, Stanford University, Stanford, CA, and approved February 2, 1998 (received for review November 18, 1997

ABSTRACT **At what biological levels are data from single-celled organisms akin to a Rosetta stone for multicellular ones? To examine this question, we characterized a saturation-mutagenized 67-kb region of the *Drosophila* genome by gene deletions, transgenic rescues, phenotypic dissections, genomic and cDNA sequencing, bio-informatic analysis, reverse transcription–PCR studies, and evolutionary comparisons. Data analysis using cDNA/genomic DNA alignments and bioinformatic algorithms revealed 12 different predicted proteins, most of which are absent from bacterial databases, half of which are absent from *Saccharomyces cerevisiae*, and nearly all of which have relatives in *Caenorhabditis elegans* and *Homo sapiens*. Gene order is not evolutionarily conserved; the closest relatives of these genes are scattered throughout the yeast, nematode, and human genomes. Most gene expression is pleiotropic, and deletion studies reveal that a morphological phenotype is seldom observed when these genes are removed from the genome. These data pinpoint some general bottlenecks in functional genomics, and they reveal the acute emerging difficulties with data transferability above the levels of genes and proteins, especially with complex human phenotypes. At these higher levels the Rosetta stone analogy has almost no applicability. However, newer transgenic technologies in *Drosophila* and *Mus*, combined with coherency pattern analyses of gene networks, and synthetic neural modeling, offer insights into organismal function. We conclude that industrially scaled robogenomics in model organisms will have great impact if it can be realistically linked to epigenetic analyses of human variation and to phenotypic analyses of human diseases in different genetic backgrounds.**

Functional genomics is a widely used descriptor covering almost as many areas of research as it has interpretations. For metazoans, it encapsulates everything from the level of gene expression through morphogenesis to organismal phenotype. A major unknown in this huge field is the extent to which the *processes* giving rise to any metazoan phenotype are transferable from one organism to another. To examine this issue, we utilized *Drosophila* as the experimental organism and *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens* as the major comparators. We first characterized a substantial part of the fly X chromosome bracketed by the genes *amnesiac* and *suppressor of forked* (Fig. 1), obtained an overview of the mutational properties and phenotypes resulting from perturbations of this 2-megabase region (1), then focused on a 67-kb subregion that has been the ongoing focus of our laboratories (2–9) and that is anchored by the *flightless* locus. We now present the genomic and cDNA sequences for this region, the phenotypic analyses at the cellular and organismal levels after genetic, transgenic, and deficiency perturbations, and the evolutionary genomics. The data give an indication of the extent to which phenotypic predictions from model organisms to human beings are currently realistically possible.

MATERIALS AND METHODS

The 66,669-bp sequence of the *flightless* region was generated by dividing the region into 18 fragments, subcloning each into pEMBL, pGEM, and Bluescript vectors, and shotgun cloning fragments into bacteriophage M13 mp10. DNA sequencing was carried out as previously described (7, 8), with the genomic and cDNA sequences being determined on both strands. Database searches were carried out at the National Center for Biotechnology Information by employing the BLAST network service. CLUSTAL W and MACAW were used in sequence alignments.

DNA and RNA extractions, blotting and hybridization, screening, and cloning of cDNAs from λ gt10 and gt11 *Drosophila* libraries were performed as previously described (2, 4–8). mRNA preparations from dissected tissues of a *white*[1] stock were performed by using TRIzol and the conditions recommended by the supplier (GIBCO/BRL). Reverse transcription (RT)-PCR analyses used oligo$(dT)_{12–18}$ primer, GIBCO/BRL reverse transcriptase, and Boehringer Mannheim reagents. Amplification was for 30 cycles; an initial denaturation for 2 min at 94°C, cycle denaturation for 30 sec at 94°C, cycle annealing for 30 sec at 55°C, and cycle extension for 6 min at 72°C. Chromosomal breakpoints were mapped as previously described (6), and 8 parts of the fly genomic region (denoted T1 to T8 in Fig. 1) were cloned into the vector pW8 and used in constructing transgenic organisms (2–4, 8).

RESULTS AND DISCUSSION

Structural Genomics. We cloned the genomic DNA from the region between the breakpoints of deficiencies D1 and D8 (Fig. 1) and used it to exhaustively screen cDNA libraries from different stages of development. These screens yielded 138 cDNAs, which constituted the outputs of 12 loci. We sequenced the longest cDNA and appropriate variants from each locus, sequenced the 67 kb of genomic DNA, and carried out RT-PCR analyses and Northern blotting on the transcription units. These data reveal that 75% of the genomic DNA is converted into 12 transcription units, denoted *tweety, flightless, dodo*, *penguin, small optic lobes, innocent bystander, waclaw*, *bobby sox, sluggish*, *Helicase, misato*, and *la costa* (Figs. 1 and 2). Two transcription units, *dodo* and *penguin*, overlap in their 3′ untranslated regions.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

This paper was submitted directly (Track II) to the *Proceedings* office.
Abbreviation: RT, reverse transcription.
Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. AF017777).
§To whom reprint requests should be addressed. e-mail: miklos@nsi.edu.

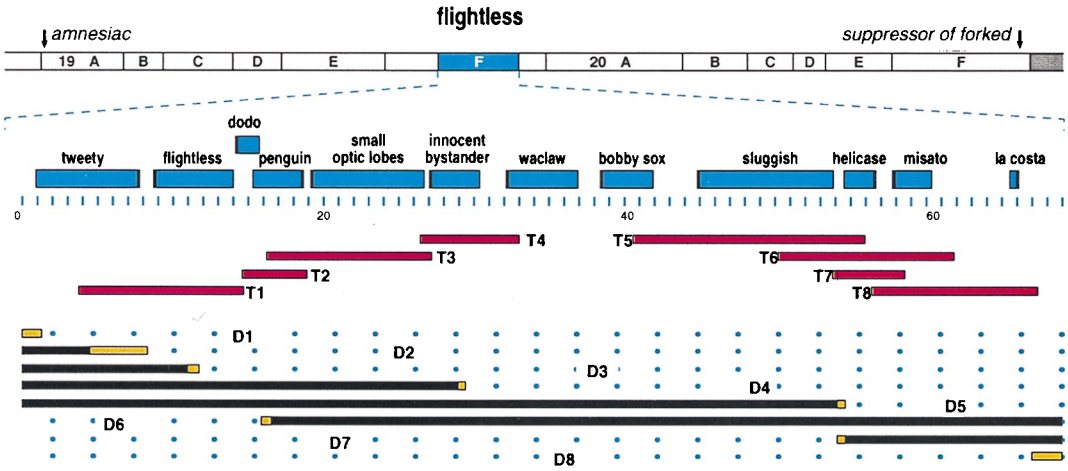

FIG. 1. Cytogenetic and molecular characteristics of the approximately 2-megabase region at the base of the X chromosome of *Drosophila melanogaster*. Polytene subdivisions 19A through 19F and 20A through 20F are as shown. The 67-kb *flightless* region (in blue) is expanded to illustrate the 12 primary transcription units. The genomic transforming fragments used to rescue the mutant phenotypes are T1 through T8. The deficiencies used to uncover the mutant phenotypes are designated D1 through D8 (dotted lines represent the deficiency, and the approximate breakpoints are shown in orange). Deficiencies D1 through D8 correspond to deficiencies 17–257, GA104, GE263, 2/19B, JC77, 16–129, HM44, and Q539 (1–8).

Gene expression was examined by using Northern blotting and RT-PCR analysis. The former reveals that 6 of the 12 transcription units (*flightless, penguin, innocent bystander, Helicase, misato,* and *la costa*) yield a single band in embryos, larvae, pupae, and adults, whereas the remaining 6 yield two or more bands, some as a result of alternative splicing, others because of the expression of relatives elsewhere in the genome (data not shown). The RT-PCR data indicate that most genes are expressed in nearly every tissue we have examined; larval salivary glands, larval fat bodies, larval brain, imaginal disks, pupal brain, adult thorax, and adult ovaries. Two examples of this extensive molecular pleiotropy are shown from the *Helicase* and *misato* loci (Fig. 3). We find that most genes are maternally expressed, all are expressed at some stage in the developing or adult nervous system, and some are differentially expressed during the aging of the adult brain (lanes 8–14 in Fig. 3).

Bio-informatics. When state-of-the-art algorithms pioneered by Burge and Karlin (13) are applied to our genomic

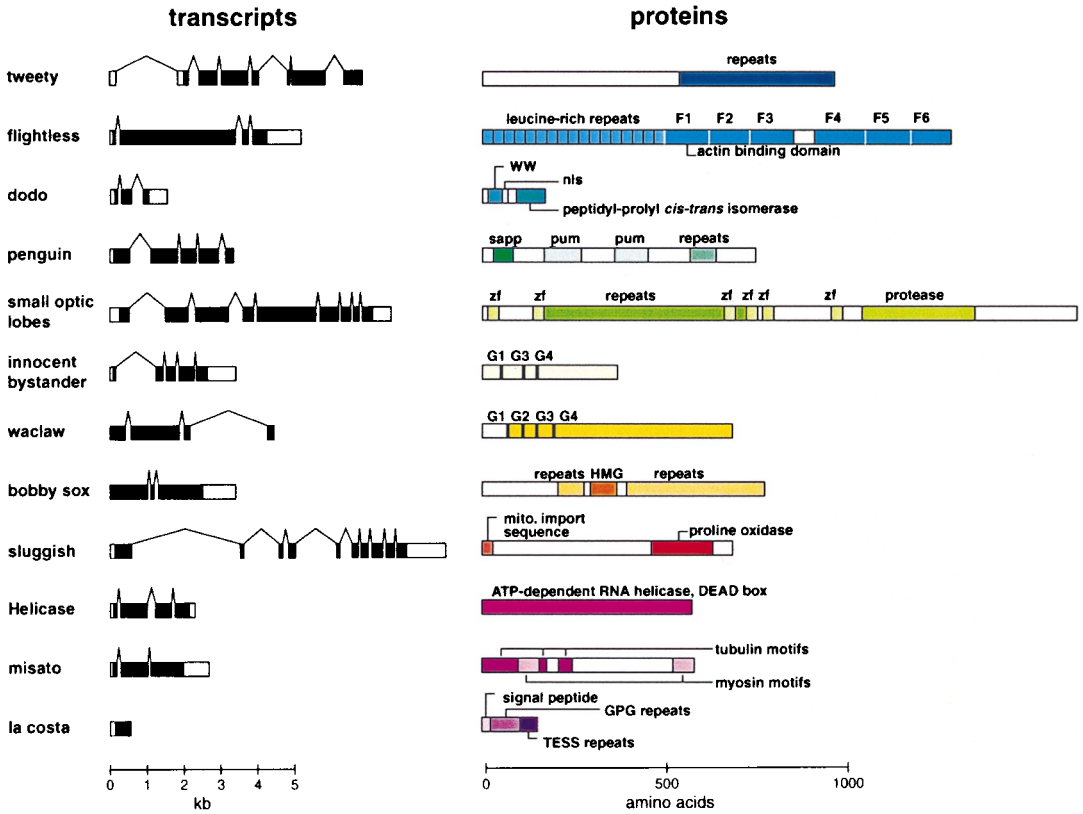

FIG. 2. The intron–exon structures of the 12 transcription units, and the structures of the 12 predicted proteins. The domains, motifs, and repetitive regions are as shown, and their relatives in different phyla are traceable by accession numbers, references, or references in the text. They are as follows: *tweety*; *flightless* (6); *dodo* (7, 9); *penguin* (the repeats, P46061; the pum motif, X62589; the sperm-activating precursor 1, D38490); *small optic lobes* (2, 10); *innocent bystander* (11); *bobby sox* (12); *sluggish* (4); *Helicase* (3, 8); *misato* (8); and *la costa* (TESS motif, yeast glucoamylase, PO8640).

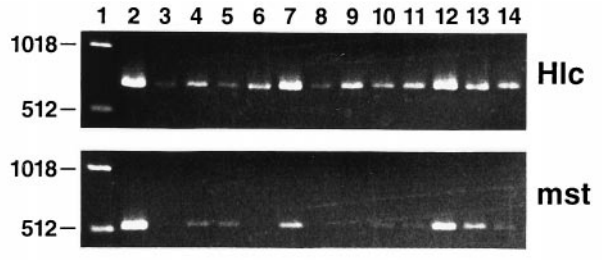

FIG. 3. RT-PCR analyses of the *Helicase* (*Upper*) and *misato* (*Lower*) loci. The molecular weight markers in lane 1 are the 1018-bp and the 517/506-bp doublet of the Boehringer 1-kb ladder. The sources of the mRNA in the remaining lanes are as follows: 2, adult ovaries; 3, larval salivary glands; 4, larval fat bodies; 5, imaginal disks; 6, larval third-instar brains; 7, late pupal brains; 8 and 9, adult female and adult male brains respectively, less than 1 hr after emergence; 10 and 11, adult female and adult male brains at 5 days after emergence; 12 and 13, adult female and adult male brains at 20 days after emergence; and 14, adult thoraces.

sequence data, half of the proteins or their isoforms are predicted exactly. The remaining half of the predictions are close to those from our isolated cDNA sequences, and some of the predictions have been helpful in finding alternatively spliced variants that are not represented in cDNA libraries, but which can now be looked for by using RT-PCR methods. These differences between the actual cDNA data and the bioinformatic predictions are largely due to the presence of micro-exons, alternative splice sites, as well as "in-frame" introns that occur in a number of the transcription units. The ATG codon of the *innocent bystander* transcription unit, for example, lies within a 6-amino acid micro-exon that is over a kilobase removed from the adjacent exon. It is presently impossible to distinguish between this ATG and many others by using bio-informatic approaches. Furthermore, one donor/acceptor splice site combination in *small optic lobes* is gC. . . ag, instead of the usual gt. . . ag. In addition, the cDNA data reveal that the *small optic lobes*, *sluggish*, and *innocent bystander* genes all have some introns that are "in-frame," and hence incorrectly predicted proteins are the outcome if only genomic sequence data are used. Finally, at least three transcripts, *tweety, small optic lobes*, and *sluggish*, are alternatively spliced, each yielding two different proteins, and their actual splice sites can be readily located only by analyzing cDNAs.

Deletion and Transgenic Studies. Molecular methods revealed 12 transcription units between the breakpoints of D1 and D8, whereas saturation genetic methods uncovered only 5 loci. We transgenically rescued all 5 available mutant phenotypes: *flightless* by T1, *small optic lobes* by T3, *sluggish* by T5, *Helicase* by T6 and T7, and *misato* by T6 and T8 (2–8) (Fig. 1). Of these 5 loci, *flightless, Helicase,* and *misato* cause organismal lethality when mutated, whereas *small optic lobes* and *sluggish* produce viable individuals with mutant phenotypes. Is the saturation mutagenesis incomplete, or do redundant genes and/or degenerate networks buffer the phenotypes? We examined this issue by carefully constructing flies that carried certain combinations of deficiencies together with particular transgenes, and asked whether these deficiencies uncovered more phenotypes than were found by using the mutagenesis approach. As we shall show, they did not.

We used an approximately 90-kb deficiency, D4 (Fig. 1), which simultaneously removes 7 genes (*innocent bystander, waclaw, bobby sox, sluggish, Helicase*, *misato*, and *la costa*), and which is also lethal for the *flightless* locus. When wild-type transgenic copies of the three essential loci, *flightless*, *Helicase*, and *misato,* are placed into the genome of these deficiency males, most of these triply transgenic males emerge as morphologically normal, but behaviorally severely debilitated, adults, even though they are still deficient for *innocent bystander, waclaw, bobby sox, sluggish,* and *la costa* (data not shown). The finding that these individuals reach adulthood at all makes it unlikely that any one of these 5 loci is essential for morphogenesis.

Since we had previously demonstrated that the simultaneous deletion of the *tweety, dodo,* and *penguin* loci was without a morphological phenotype (7), these data *in toto* mean that deletions or disruptions of most of the 12 loci in this region result in sufficiently unobtrusive morphological changes under laboratory conditions that they are not readily detected in standard high-throughput genetic screens. For pragmatic purposes, all available strong morphological phenotypes are recovered by saturation mutagenesis. Most importantly, the frequency of recognizable phenotypic perturbations that arise from this region is congruent with the unpublished estimate from the 2.5-megabase *Adh* region (M. Ashburner and G. M. Rubin, personal communication), as well as from genome-wide compilations. *In toto*, these results indicate that 30% of *Drosophila* loci are lethals, and that an additional 20% or so can be mutated or deleted to produce morphological or behavioral phenotypes that are *relatively* easy to recognize in certain genetic backgrounds (14–16). This leaves roughly 50% of the coding capacity of the *Drosophila* genome as not being strongly reflected in morphological phenotype. This approximate figure is reliably transferable to *C. elegans, Mus musculus,* and probably to *H. sapiens*.

What then are the inferred biochemical properties of the proteins in this region, into what categories do they fall, and what insights do they or their relatives provide into functional transferability across phyla?

Functional Genomics. The *predicted* protein products of this region, and their closest relatives in other organisms, are described below (Fig. 2 and Tables 1 and 2). In brief, half of them are internally repetitive, have homopeptide repeats, or exhibit unusual charge clusters (10, 17). Most importantly, the levels of amino acid sequence identity between relatives in different phyla (Table 2), as judged by rigorous criteria (18), are sufficiently high that the comparable regions are likely to be structurally similar.

The predicted proteins fall into three categories: (*i*) those in which functional characteristics are relatively easy to infer for the *whole* protein—e.g., *flightless*, *dodo*, and *Helicase*; (*ii*) those in which a single domain, or motif, provides some evidence for potential biochemical characteristics—e.g., *small optic lobes*, *innocent bystander*, *waclaw*, *bobby sox*, and *sluggish*; and (*iii*) those in which functionality is obscure—e.g., *tweety*, *penguin*, *misato,* and *la costa*.

In the first category are found the *flightless* proteins, which are 1,256, 1,257, and 1,269 amino acids in length (fly, worm, and human respectively; Table 2). All share a conserved leucine-rich repeat protein–protein binding domain whose structure is known (19) and, in addition, share six other conserved domains, the first of which, F1, has demonstrated actin binding activity. Similarly, the fly, yeast, and human *dodo* proteins, (166, 172, and 163 amino acids, respectively), have high sequence similarities. The human *dodo* protein has been crystallized, demonstrated to have peptidylprolyl *cis-trans* isomerase activity, and shown to interact with a specific kinase (20, 21). Last, the *Helicase* protein shares high sequence similarity with over 50 ATP-dependent RNA helicases from bacteria to humans.

The second category contains proteins in which potential biochemical activities are inferable for *parts* of each protein, but the bulk of the protein is of unknown function. The predicted *small optic lobes* protein contains a region related to the catalytic subunit of calcium-activated neutral proteases, six unusual zinc finger motifs, and some homopeptide stretches. The *innocent bystander* and *waclaw* proteins share motifs that are diagnostic for guanine nucleotide binding proteins. The *bobby sox* protein contains an 80-amino acid high mobility

Table 1. *Inferred* major biochemical properties of the predicted proteins (boldface) and other distinguishing features: domains, motifs, and repetitive regions

Gene	Protein properties
tweety	Repeats of glutamine, glycine, aspartic acid, and proline in C terminus
flightless	**Actin filament binding; protein–protein binding;** 6 gelsolin-like domains, 17 repeats of a leucine-rich motif
dodo	**Peotidylprolyl *cis-trans* isomerase; WW domain; nuclear localization sequence; docking surface for NIMA kinase**
penguin	Repeats of glutamine, alanine, glycine, serine, and threonine; sequence similarity to sperm activating peptide 1 precursor (denoted sapp); sequence similarity to *pumilio* repeats of *Drosophila* (denoted pum)
small optic lobes	**Calcium-activated neutral protease; zinc finger motifs of the form $WXCX_2CX_3NX_5KCX_2C$**
innocent bystander	**GTPase; G1, G3, and G4 motifs diagnostic for guanine nucleotide-binding proteins**
waclaw	**Elongation factor, GTPase; G1, G3, G4 motifs, and diagnostic G2 elongation factor motif**
bobby sox	**Transcription factor, HMG domain (DNA binding and DNA bending);** homopeptide amino acid stretches
sluggish	**Proline oxidase**
Helicase	**DEAD-box family; ATP-dependent RNA helicases**
misato	α-, β-, and γ-tubulin motifs; myosin-like motif
la costa	Signal peptide; 24 repeats of collagen-like sequence; 9 repeats of serine/threonine motif (TESS motif)

group (HMG) DNA-binding and DNA-bending domain, flanked by a number of homopeptide amino acid stretches and miscellaneous short repeats. The *sluggish* protein is a proline oxidase (4), which shares a functionally homologous catalytic domain with the yeast proline oxidase (Table 2). Nevertheless, the yeast and fly proteins differ markedly in size, by nearly 200 amino acids.

The third category contains proteins with a combinatorially novel juxtaposition of domains, motifs, and repetitive elements (Fig. 2; Table 1). The *tweety* protein, for example, contains homopeptide runs of small and/or polar amino acids that occur disproportionately in proteins implicated in human neurological disorders, such as Huntington's disease, spinocerebellar ataxia type 1, spinobulbular muscular atrophy, and denatatorubral-pallidoluysian atrophy 1 (10). The *penguin* protein is also a chimera of repetitive motifs, whereas the *misato* protein contains motifs found in different α-, β-, and γ-tubulins, as well as two regions that are related to part of a hinge region of the myosin family (8). The *la costa* protein is almost totally repetitive, consisting of 24 variant repeats of a collagen-like sequence and 9 copies of a threonine/serine repeat that is found in diverse human proteins.

Evolutionary Genomics. The detailed analysis of these proteins reveals that six of them are absent from *S. cerevisiae* as the *equivalent full length proteins* (Table 2). Many of the individual domains and motifs are present in this yeast, but they do not occur in the same combinations as in the fly. For example, there is no multidomain *flightless* protein in yeast, and none of the yeast proteins that have an LRR domain, such as adenylyl cyclase, has additional actin binding domains or *vice versa*. Similarly, there are no multidomain relatives of the *small optic lobes* protein having the combination of zinc finger motifs and a protease domain, although separate yeast proteins have these individual regions. There are also no recognizable yeast relatives of *bobby sox*, although yeast does contain a single 99-amino acid nonhistone chromosomal protein that constitutes a solo HMG domain. There are also no close *tweety*, *misato,* or *la costa* relatives (8), although a number of yeast proteins (such as glucoamylase), have the TESS motif found in *la costa*. Again, the results from this region are congruent with our examination of *Drosophila* databases, which indicates that a significant proportion of fly proteins do not occur in yeast. Thus outside of that overlapping cohort of proteins that is shared by yeast and *Drosophila*, and the common components of overlapping networks, functional transfer from yeast to fly has its limitations even at this most basal level.

Transferability of Function Across Phyla. How useful are data of this type in accelerating knowledge across phyletic lines? The answer to this question depends on the *level* at which comparisons are made.

For example, the yeast, fly, and human *dodo* proteins are functionally interchangeable in yeast, are localized in the nuclear speckle, have similar mutant phenotypes at the *cellular level*, and down-regulate G_2/M-specific cell cycle kinases (7, 9, 20). However, at the *organismal level* the transferability is difficult, because whereas the yeast protein is essential for cell division, the fly protein is nonessential (7). Despite assertions that the human *dodo* protein (confusingly renamed Pin1) is an essential and critical one (20), neither data from human diseases nor knockout data from the mouse are yet available to substantiate this claim. Although the crystallography of the protein and its cell biology are well described, it is not possible to *predict* the resultant fly or human mutant phenotypes from a knowledge of the yeast data.

As with the previous example, data transferability within the *innocent bystander* family, with its members in yeast, fly, mouse, and humans, is excellent at the protein and cellular levels. The best-characterized fly member is the essential *peanut* locus, which leads to cell division defects in cytokinesis when mutated (11). At the cellular level, the *peanut* and *innocent bystander* proteins are colocalized in the cleavage furrows of dividing cells, in the cytoplasmic bridges connecting daughter cells after division, and in subsets of eye disk cells and larval nervous system cells (11, 22). In the mouse, an antibody to one of the mouse *innocent bystanders* (also known as *diff6*), also stains cleavage furrows, as well as growth cones of differentiating PC-12 cells (22).

In contrast to the *dodo* and *innocent bystander* examples, the transferability of function above the protein level for the *flightless*, *small optic lobes, bobby sox, sluggish,* and *Helicase* loci is currently problematic. The reasons for this are largely due to the difficulties involved in comparing complex phenotypes arising from different developmental processes and different neuroanatomies, across many different levels, in the worm, fly, and human being.

For example, the human *flightless* gene maps to 17p11.2, and the common clinical features of individuals heterozygous for a massive deletion of this region include brachycephaly, brachydactyly, mental retardation, short stature, self-destructive behavior, and facial dysmorphology with midface hypoplasia. The variable clinical features include sleep disturbance, cardiac defects, genital anomalies, eye abnormalities, hearing loss, seizures, hand anomalies, cleft lip, and cleft palate. This is hardly a surprise, because this chromosomal region is about 10 megabases in size and probably contains in excess of 100 genes. However, the usual fashionable attempt to depict a complex phenotype in the context of a single "candidate" gene, *flightless*, illustrates the extraordinarily simplistic interpretations that are sometimes invoked for the gene–phenotype transition (23), with little consideration of the importance of epigenetic processes, or the different genetic backgrounds that have

Table 2. Comparative genomics of the predicted proteins in the *flightless* region

D. melanogaster			*S. cerevisiae*				*C. elegans*				*H. sapiens*		
Gene	Deletion	Protein, aa	Gene	Protein, aa	Location	Iden/Sim, %	Gene	Protein, aa	Location	Iden/Sim, %	Gene/EST	Protein, aa	Iden/Sim, %
tty	Viable	836	—	—	—	—	F42E11.2	519	X	16/36	F11755	—	—
fli	**Lethal**	1,256	—	—	—	—	*fli*	1,257	III	49/70	*fli*	1,269	57/75
dod	Viable	166	ESS1	172	X	46/56	Y110A2	167	I	45/57	*dodo*	166	53/61
pen	Viable	737	YDR496c	657	IV	17/30	ZK945.3	766	II	21/32	KIAA0020	508	18/31
sol	Viable	1,597	—	—	—	—	—	—	—	—	R40581	—	—
iby	Viable	361	CDC 12	407	VIII	39/59	F07A5.7	757	I	17/36	*H5*, D28540	406	61/78
waw	Viable	679	L8003.7	645	XII	44/59	ZK1236.1	645	III	48/59	AA780390	—	—
bbx	Viable	769	—	—	—	—	T22H6.6	833	X	51/63	AA040785	—	—
slg	Viable	669	PUT1	476	XII	30/69*	F14E12.h	516	IV	36/46	F19541.1	826	32/47
Hlc	**Lethal**	566	YLR276c	594	XII	38/58	C24H12.4	653	II	45/62	R27090-2	483	41/52
mst	**Lethal**	574	—	—	—	—	—	—	—	—	R17341	—	—
lcs	Viable	145	—	—	—	—	H17B01.2	530	II	41/70	AA305561	—	—

The viability of organisms homozygous deficient for each fly gene; the lengths of the different proteins; their closest relatives at the amino acid sequence level; the genomic locations in the *S. cerevisiae, C. elegans*, and human genomes; and the levels of amino acid sequence identity (Iden) and similarity (Sim) relative to the *Drosophila* protein are as shown. EST, expressed sequence tag.
*Refers to the catalytic domain of *sluggish*.

enormous influences on human phenotypes. Furthermore, data from other phyla reveal the contextually diverse nature of *flightless* expression: worm *flightless* is expressed in certain pharyngeal epithelia and vulval muscles (24); fly *flightless* is found in ovaries, larval fat bodies, brain, and adult thorax (1, 6); and human *flightless* is found in heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas (23). The differences in epigenesis between these three organisms, particularly in the stereotyped fixed cell lineages of the worm, the blastoderm-based early embryogenesis of the fly, and the epigenetically more flexible human development, presently preclude a smooth transfer of information from one to the other except at the level of protein structure.

A similar "levels" problem in transferability is found in the case of the *bobby sox* family. This protein is most closely related to the human SOX9 protein, a putative transcription factor involved in the early development of the skeleton, in sex determination, and in the human dysmorphology syndrome, campomelic dysplasia (Table 2). The human phenotype is complex, with affected individuals commonly having a deformed pelvis, a missing pair of ribs, bowing and angulation of the long bones, cleft palate, micrognathia, and ear defects affecting the malleus, incus, stapes, and tympanum (12). If indeed *bobby sox* and SOX9 turn out to be functionally interchangeable, the useful data transfer is not likely to be in the realm of skeletal development, but rather in terms of protein structure.

In the examples of the defective *sluggish* gene, and one of its putative relatives, the *PRO/Re* gene of *Mus musculus*, elevated levels of free proline are correlated with sluggish behavior in both organisms. In addition, mice with elevated serum proline levels have generalized hair loss, are unusually cannibalistic, and are difficult to breed. In *Rattus norvegicus*, high free proline levels lead to neuronal cell death in the hippocampus, and excess proline is excitotoxic on pyramidal and granule cells. In humans, elevated levels of free proline are found in the blood and urine of patients with type I hyperprolinemia, and many of these cases are associated with mental retardation, convulsive disorders, renal abnormalities, hereditary deafness, and photogenic epilepsy. Again, insufficient overlapping data are yet available from any model organism to allow for a direct transfer of useful phenotypic information to humans.

The DEAD-box ATP-dependent RNA helicases are helpful in revealing a quite different, but most important, cross-phyletic result relating to essentiality and functional redundancy. Although there are at least a dozen members of the *Helicase* family in the fly (Berkeley *Drosophila* Genome Project), none are normally able to rescue the lethality caused by mutations at the *Helicase* locus. Thus the existence of a large number of closely related family members within the same genome is an unreliable predictor of the functional interchangeability of any family member.

Last, there are presently insufficient data on the *tweety, penguin, small optic lobes, misato*, and *la costa* loci, from any organism, to allow for meaningful cross-phyletic comparisons.

In summary, metazoan functional genomics is a radically different field from unicellular functional genomics, and it *depends absolutely on epigenetic context* and the so-called proximal and distal regulatory networks in which any protein is transiently engaged (25). Phenotypic *prediction* in the metazoa is not automatically derivable from protein function in unicellular organisms, and the structurally and functionally conservative yeast/fly/human *dodo* data set discussed earlier is an exemplar of the current predictive inadequacy at higher levels. These findings also place into perspective the popular but predictively unhelpful Rosetta stone analogy, in which the decipherment of three different scripts (Egyptian hieroglyphics, the cursive form of this hieroglyphics, and Greek), has been extended to the decipherment of biological systems, specifically to the model organisms–human case. For example, it is claimed that "the meaning of the sequence of the (human) disease genes is routinely deciphered by using information from yeast and worms" (26). As we have seen from the model organism data, and as is clear from a comparison of many human disease genes, such as the breast cancer ones, the decipherment is far from routine. In this sense the Rosetta stone analogy more closely parallels the original decipherment, which depended as much on a knowledge of Coptic, and the cartouches on the obelisk at Philae, as they did on the stone itself.

Predicting Human Disease Phenotypes by Means of Data Transfer from Model Organisms? The acid test for functional genomics is no different to any other field: that test is the robustness of its predictions. What then are the data that usefully impinge upon the very foundations of functional genomics, namely the issue of cross-phyletic transferability?

First, it is firmly established that organisms in different phyla share a core of conserved proteins, some core regulatory elements, and some core signal transduction networks (14, 27). It is also well documented that data transfer between phyla is powerful at the conserved protein, domain, and motif levels. However, once the comparisons encompass regulatory networks and epigenesis, or those components that are unique to each phylum [such as the exclusively vertebrate V(D)J com-

ponents and the many metazoan proteins that are not found in *S. cerevisiae*], there is considerable difficulty in transferability. Furthermore, as we have seen in this study, whenever a locus leads to small effects on fitness in one phylum, but drastic inviability problems in another, the transferability problem is nontrivial. Furthermore, the sheer developmental complexity of some human phenotypes, particularly the large number of neuropsychiatric disorders characterized by deviance in social, communicative, and cognitive development, such as Tourette's syndrome (28), indicates that model organisms are inappropriate comparators at these levels.

At the level of regulatory networks, it is obvious that a gene must be turned on where its protein is needed, but the converse is not necessarily true. There are excellent data which indicate that some protein expression can be a default outcome of regulatory networks, and where expression has little or no measurable phenotypic consequences at the organismal level (29–31). This finding has significant consequences for the functional interpretation of transcriptome analyses. In this context, the data described in this study, and those from the unpublished 2.5-megabase *Adh* region (Berkeley *Drosophila* Genome Project; M. Ashburner and G. M. Rubin, personal communication), highlight a looming challenge. Given the upcoming availability of exhaustive transcriptome, proteome, and genomic net data (25, 32, 33) and its extension to metazoan development, the critical issue will center on which of the changes found in networks are of functional significance to the organism. The solution to this problem will require not only expanded robogenomics, but experiments utilizing very specific modifications of parts of genomes in conjunction with modified transgenic material, all analyzed in developmental time and space with full transcriptome and proteome measurements. If unrefined pseudosolutions are to be avoided, the proving ground of functional genomics in metazoans will ultimately be the epigenesis of complex nervous systems in the context of the classic "levels" problem (34–36). Functional genomics will need to squarely confront subtle phenotypic effects and the marginal effects of multiple genomic changes on fitness, a field pioneered in yeast (31). In organisms with complex brains, and with experience-dependent brain plasticity, this will require experimental approaches to which *Drosophila* and *Mus* are well suited—namely, exquisitely controlled gene expression, misexpression, and screens using different genetic backgrounds (37–39). The sophisticated transgenically based methods now available in the fly, the genomic nets from yeast (33), the coherency pattern analyses being pioneered in vertebrates (25), and the synthetic neural modeling approach (34) all provide entry points into the integration of genomic perturbations and phenotypic analyses with transcriptome and proteome data in epigenetic contexts. These approaches should reveal those interacting components of networks that can be evaluated in clinical screens in humans with different genetic backgrounds.

We are grateful to K. Crossin, S. Delaney, R. Greenspan, D. Hayward, and F. Jones for advice and help, and A. Kasprzak, J. Mason, and D. Slifka for their efforts in the sequencing and RT-PCR work. G.L.G.M. was funded by Neurosciences Research Foundation; H.G.d.C. by National Science Foundation Grant DCB-9106129, and R.M. by the Australian National University.

1. Perrimon, N., Smouse, D. & Miklos, G. L. G. (1989) *Genetics* **121,** 313–331.
2. Delaney, S. J., Hayward, D. C., Barleben, F., Fischbach, K.-F. & Miklos, G. L. G. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 7214–7218.
3. Schuppler, U. C. (1992) Ph.D. thesis (Australian National University, Canberra).
4. Hayward, D. C., Delaney, S. J., Campbell, H. D., Ghysen, A., Benzer, S., Kasprzak, A. B., Cotsell, J. N., Young, I. G. & Miklos, G. L. G. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 2979–2983.
5. Campbell, H. D., Schimansky, T., Claudianos, C., Ozsarac, N., Kasprzak, A. B., Cotsell, J. N., Young, I. G., de Couet, H. G. & Miklos, G. L. G. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 11386–11390.
6. de Couet, H. G., Fong, K. S. K., Weeds, A. G., McLaughlin, P. J. & Miklos, G. L. G. (1995) *Genetics* **141,** 1049–1059.
7. Maleszka, R., Hanes, S. D., Hackett, R. L., de Couet, H. G. & Miklos, G. L. G. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 447–451.
8. Miklos, G. L. G., Yamamoto, M.-T., Burns, R. G. & Maleszka, (1997) *Proc. Natl. Acad. Sci. USA* **94,** 5189–5194.
9. Maleszka, R., Lupas, Á., Hanes, S. D. & Miklos, G. L. G. (1997) *Gene* **203,** 89–93.
10. Karlin, S. & Burge, C. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 1560–1565.
11. Neufeld, T. P. & Rubin, G. M. (1994) *Cell* **77,** 371–379.
12. Kent, J., Wheatley, S. C., Andrews, J. E., Sinclair, A. H. & Koopman, P. (1996) *Development (Cambridge, U.K.)* **122,** 2813–2822.
13. Burge, C. & Karlin, S. (1997) *J. Mol. Biol.* **268,** 78–94.
14. Miklos, G. L. G. & Rubin, G. M. (1996) *Cell* **86,** 521–529.
15. Ashburner, M., Thompson, P., Roote, J., Lasko, P. F., Grau, Y., Messal, M. El, Roth, S. & Simpson, P. (1990) *Genetics* **126,** 679–694.
16. Spradling, A. C., Stern, D. M., Kiss, I., Roote, J., Laverty, T. & Rubin, G. M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 10824–10830.
17. Karlin, S. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5593–5597.
18. Sander, C. & Schneider, R. (1991) *Proteins Struct. Funct. Genet.* **9,** 56–68.
19. Kobe, B. & Deisenhofer, J. (1995) *Nature (London)* **374,** 183–186.
20. Lu, K. P., Hanes, S. D. & Hunter, T. (1996) *Nature (London)* **380,** 544–547.
21. Ranganathan, R., Lu, K. P., Hunter, T. & Noel, J. P. (1997) *Cell* **89,** 875–886.
22. Longtine, M. S., De Marini, D. J., Valencik, M. L., Al-Alwar, O. S., Fares, H., De Virgilio, C. & Pringle, J. (1996) *Curr. Opin. Cell Biol.* **8,** 106–119.
23. Campbell, H. D., Fountain, S., Young, I. G., Claudianos, C., Hoheisel, J. D., Chen, K.-S. & Lupski, J. R. (1997) *Genomics* **42,** 46–54.
24. Lynch, A. S., Briggs, D. & Hope, I. A. (1995) *Nat. Genet.* **11,** 309–313.
25. Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L. & Somogyi, R. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 334–339.
26. Botstein, D. & Cherry, J. M. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 5506–5507.
27. Artavanis-Tsakonas, S., Matsuno, K. & Fortini, M. E. (1995) *Science* **268,** 225–232.
28. Leckman, J. F., Peterson, B. S., Anderson, G. M., Arnsten, A. F. T., Pauls, D. L. & Cohen, D. J. (1997) *J. Child Psychol. Psychiat.* **38,** 119–142.
29. Erickson, H. P. (1993) *J. Cell Biol.* **120,** 1079–1081.
30. Dickinson, W. J. (1988) *BioEssays* **8,** 204–208.
31. Thatcher, J. W., Shaw, J. M. & Dickinson, W. J. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 253–257.
32. Wilkins, M. R., Sanchez, J.-C., Williams, K. L. & Hochstrasser, D. (1995) *Electrophoresis* **17,** 830–838.
33. Brent, R. (1997) *Nat. Genet.* **16,** 216–217.
34. Edelman, G. M. (1987) *Neural Darwinism* (Basic Books, New York).
35. Miklos, G. L. G. (1998) *Daedalus,* in press.
36. Miklos, G. L. G. (1993) *J. Neurobiol.* **24,** 842–890.
37. Hay, B. A., Maile, R. & Rubin, G. M. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 5195–5200.
38. Rorth, P. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 12418–12422.
39. Brand, A. H. & Perrimon, N. (1993) *Development (Cambridge, U.K.)* **118,** 401–415.