Analysis

A Survey of Human Disease Gene Counterparts in the *Drosophila* Genome

Mark E. Fortini,* Marian P. Skupski,‡ Mark S. Boguski,§ and Iswar K. Hariharan‖

*Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104; ‡Celera Genomics, Rockville, Maryland 20850; §National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894; and ‖Massachusetts General Hospital Cancer Center, Charlestown, Massachusetts 02129

The recent sequencing of the *Drosophila* genome as a collaborative effort between the Berkeley *Drosophila* Genome Project (BDGP)[1] and Celera Genomics provides an unparalleled opportunity to assess the prevalence of human disease gene counterparts in the fly genome (Adams et al., 2000; Rubin et al., 2000). Previous surveys based upon limited data available during earlier phases of the sequencing project have suggested that >50%, and perhaps as many as 75%, of human disease genes are conserved in *Drosophila* (Banfi et al., 1996; Berkeley *Drosophila* Genome Project, 1999; Pickeral et al., 2000). With the virtually complete sequence now in hand, we were able to perform a more comprehensive survey, resulting in the finding that 178 out of 287 human disease genes (62%) appear to be conserved in the fly (Rubin et al., 2000). The major findings of this survey have been presented and briefly discussed (Rubin et al., 2000). Here we describe in more detail the manner in which we conducted this survey, and the limitations of large-scale computational methods for whole-genome searches of this type. We summarize the results of our searches for several different categories of human disease genes in the fly, and their implications for the utility of *Drosophila* for the analysis of human disease gene function.

Survey Design and Methodology

Constructing the Human Disease Gene List

The core component of our survey is a list of 287 human disease genes representing several different classes of diseases, including cancer, neurological diseases, cardiovascular diseases, malformation syndromes, hematological, immune, endocrine, renal, and metabolic disorders (Table I). This list was compiled by scanning the Online Mendelian Inheritance in Man database (OMIM; http://www.ncbi.nlm.nih.gov/omim/) as well as medical textbooks and review articles listing classes of human disease genes. The criterion for inclusion in the final list was that the human gene must actually be mutated, altered, amplified, or deleted in human subjects with the disease. From our initial set of >800 human genes associated with diseases, over half were eliminated because they did not meet this criterion. Genes potentially linked to a human disease solely by cell culture experiments, yeast two-hybrid interaction screens, model organism studies, or similar approaches were excluded from our analysis. Each human disease gene on the final list was confirmed by checking OMIM or published literature sources, and was placed in the most relevant disease category on the list. For human disease genes in which different paralogs have been associated with disease, such as Ras family members, rhodopsins, and some HOX and PAX gene family members, a single example was chosen to represent the group and redundant paralogs were eliminated from the list. In some cases, assignment of a gene to a particular category was somewhat arbitrary, since altered gene function may result in different diseases or a syndrome characterized by multiple organ involvement or a complex pathophysiology. For example, human *Notch* gene mutations cause both cancer (*Notch1* rearrangements in T cell acute lymphoblastic leukemia) and neurological disease (*Notch3* point mutations in CADASIL). The final list of 287 human disease genes is not meant to be comprehensive; in fact, there are currently estimated to be 1,000 human disease genes defined by at least one allelic variant each (Antonarakis and McKusick, 2000). However, our list does represent a large set of genes mutated in a wide variety of human diseases, adjusted to prevent biasing the survey towards certain common gene families.

Address correspondence to Mark E. Fortini, Department of Genetics, University of Pennsylvania School of Medicine, Stellar-Chance Laboratories, Room 709C, 422 Curie Boulevard, Philadelphia, PA 19104. Tel.: (215) 573-6446. Fax: (215) 573-9411. E-mail: fortini@mail.med.upenn.edu

[1]*Abbreviations used in this paper:* ATM, ataxia telangiectasia; *ATR, ATM-related*; BDGP, Berkeley *Drosophila* Genome Project; OMIM, Online Mendelian Inheritance in Man database; VWF, von Willebrand factor; *WAS*, Wiskott-Aldrich syndrome.

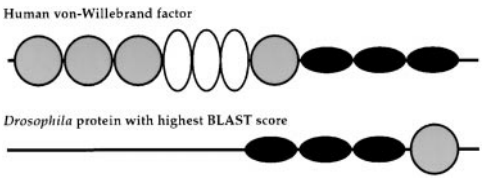

Figure 1. InterPro alignment of VWF to the protein that it most resembles in the *Drosophila* genome. Representation of domains of the human VWF protein and the most similar predicted *Drosophila* protein (not drawn to scale). The *Drosophila* ORF with the best sequence similarity to human VWF (bottom) has several VWF-C domains (black), a VWF-D domain (gray), and a very low E value when aligned to human VWF (top). However, it has no VWF-A domains (white), which bind GPIIb, collagen and heparin, and are crucial for the function of the human protein. Thus the *Drosophila* predicted protein is unlikely to represent a true homologue of human VWF.

Bulk BLAST Searches and Analysis

The initial step in the survey was to perform bulk BLASTP searches in which the 287 human disease protein sequences were used as queries to search a database consisting of all predicted proteins encoded by the yeast, worm, and fly genomes (38,860 total sequences; Rubin et al., 2000). To control for potential frameshift errors in the *Drosophila* genomic sequence, which might result in an open reading frame of one actual gene being split and assigned to two or more predicted genes, searches against a six-frame translation of the entire genome using TBLASTN were also conducted with the disease gene query sequences. Only two cases were detected in which matches to the genomic sequence were better than to the predicted protein data set.

These BLAST searches provided us with an initial set of BLAST scores and amino acid sequence alignments of each human disease protein to the top five matches in each of the three fully sequenced genomes. To assess the likelihood that a given human disease protein has a counterpart encoded in the fly genome, each set of alignments was visually inspected and compared with one another across the different species. Since the bulk BLAST searches were done with an aggregate set of 38,860 target sequences representing all yeast, worm, and fly proteins, using a normalized database size setting equal to the largest possible proteome size for all three genomes combined (the z parameter in BLAST, derived by adding the nucleotide lengths of all the genomes and dividing by three), the resulting E values in the different organisms could be compared directly despite their nonequivalent genome sizes. Cross-species inspection of the best alignments for each query protein aided our judgments by revealing whether specific protein domains tend to be highly conserved in all species, whether the fly possesses a much better match to the query than do yeast and worm, and whether a clear pattern of conserved residues can be discerned in the best alignments from the different species. In addition, the overall domain compositions of some fly proteins were compared with their human query proteins using InterPro, a database of protein sequence motifs (Apweiler et al., 1999). Supporting biological information, such as published reports of known fly and human gene homologues as well as the consensus opinion of the *Drosophila* research community, was also taken into consideration. For example, many human disease genes were already known to be homologous (and in some cases named after) well-characterized fly genes, including *patched* (basal cell nevus syndrome), *Sonic hedgehog* (holoprosencephaly 3), *PAX6* (aniridia), *Notch3* (CADASIL multi-infarct dementia), *Jagged1* (Alagille syndrome), *Twist* (Saethre-Chotzen syndrome), *GLI3* (Greig cephalopolysyndactyly), *Diaphanous 1* (autosomal dominant deafness), and *Eyes absent 1* (Melnick-Fraser brachiootorenal dysplasia).

Evaluation of Questionable Cases

The above strategy allowed ~200 of the 287 genes on the list to be scored as present or absent in the fly genome in a relatively straightforward manner. For ~90 genes, however, the analysis was more problematic. For some of these cases, the query protein sequence was very short or the level of amino acid identity in the best BLAST alignments was not very impressive, resulting in a poor E value. A particularly striking example is the fly p53 gene, which exhibits a low degree of similarity to human p53 family members (E value = 2×10^{-8}), but which was judged to be a homologue because it shows a conserved organization of functional domains, and its DNA-binding domain contains conserved residues at positions that are commonly mutated in human cancers (Walker et al., 1999). In other cases, fly proteins with very high similarity to the human query sequence, often over a considerable portion of the protein, were considered nonhomologous due to their different domain organization. The best fly protein alignment to the von Willebrand Factor (VWF) protein, for example, extended over much of the query sequence and had an E value of 1×10^{-106}. However, the *Drosophila* protein lacks the VWF type A domains that are crucial for the function of the human protein (Fig. 1). For the fly protein most similar to human E-cadherin (implicated in gastric cancer), supporting biological evidence for the fly protein led us to conclude that it should be considered a cadherin homologue even though its domain structure is not quite identical to mammalian E-cadherin.

A common problem that arose in the large-scale BLAST searches was that many human disease proteins belong to large families of closely related yet functionally distinct proteins, such as kinases, cell adhesion molecules, and certain classes of transcription factors. Inspection of BLAST scores and alignments was often insufficient to determine whether the most related protein detected in a given species represents a true homologue or merely a related protein. *Drosophila* possesses ~300 protein kinases and >100 homeodomain proteins, for example, and identification of many of these proteins as exact homologues to known proteins in other organisms is often not possible (Rubin et al., 2000). A particular kinase or homeodomain protein implicated in human disease was not considered by us to possess a *Drosophila* homologue simply because it shows some similarity to various fly kinases or homeodomain proteins, without convincing conservation beyond the generic kinase or homeodomain motifs. In these cases,

Table I. Summary of the Human Disease Gene Survey in Drosophila

Present	Absent	Present	Absent
Cancer		**Cardiovascular**	
Chronic Myeloid Leukemia (ABL1)	Breast Cancer (BRCA1)	Atrioventricular Conduction Defects (CSX)	
Acute Myeloid Leukemia (DEK)	Breast Cancer (BRCA2)	High Density Lipoprotein Deficiency 1 (ABCA1)	
Adenomatous Polyposis Coli (APC)	B-Cell Lymphoma 3 (BCL-3)	Long QT 1 (KCNQ1)	
Oncogene (AKT2)	Leukemia (FMS)	Long QT 2 (KCNH2)	
Ataxia Telangiectasia (ATM)	Platelet Derived Growth Factor (PDGFB)	Long QT 3 (SCN5A)	
Basal Cell Nevus (PTC)	Oncogene (ETS1)	Familial Hypertrophic Cardiomyopathy (MYH7)	
B-Cell Lymphoma 2 (BCL-2)	Fibroblast Growth Factor (FGF3)		
Bloom (BLM)	Fanconi's Anemia A (FANCA)	**Endocrine**	
Burkitt's Lymphoma (MYC)	Fanconi's Anemia C (FANCC)	Diabetes (INS)	Adrenal Hypoplasia (NROB1)
Chk2 Protein Kinase (CHK2)	Fanconi's Anemia G (FANCG)	Diabetes (INSR)	Androgen Insensitivity (AR)
Chronic Myeloid Leukemia (BCR)	Oncogene (KIT)	Hyperinsulinism (ABCC8)	Adrenal Hyperplasia III (CYP21A2)
Cyclin D1 (CCND1)	T-cell Leukemia (LCK)	Hyperinsulinism (KCNJ11)	Diabetes Insipidus (AVP)
Cyclin-dependent Kinase 4 (CDK4)	P53 Regulator (MDM2)	Hypothyroidism (SLC5A5)	Diabetes with Hypertension (PPARG)
Epidermal Growth Factor Receptor (EGFR)	Renal Cancer (MET)	Leydig Cell Hypoplasia (LHCGR)	Dwarfism (GH1)
Oncogene (ERBB2)	Thyroid Cancer (NTRK1)	Maturity-onset Diabetes of the Young 1 (HNF4A)	Dwarfism (GH2)
E-Cadherin (CDH1)	Tumor Suppressor (P16 INK4A)	Maturity-onset Diabetes of the Young 2 (GCK)	Gonadal Dysgenesis (SRY)
Ewing Sarcoma (FLI-1)	Tumor Suppressor (P14 ARF)	McCune-Al bright (GNAS1)	Hypothyroidism (TRH)
Colon Cancer (MSH2)	Wilm's Tumor (WT1)	Noninsulin-dependent Diabetes (PCSK1)	Maturity-onset Diabetes of the Young 3 (TCF1)
Colon Cancer (MSH3)		Pendred (PDS)	Maturity-onset Diabetes of the Young 4 (IPF1)
Colon Cancer (MSH6)		Vitamin D-Resistant Rickets (VDR)	Maturity-onset Diabetes of the Young 5 (TCF2)
Colon Cancer (MLH1)			Obesity (LEP)
Colon Cancer (PMS2)			Obesity (LEPR)
Lymphoma (MCF2)			Obesity (MC4R)
Pancreatic Cancer (DPC4/MADH4)			Obesity (POMC)
Multiple Endocrine Neoplasia 1 (MEN1)			Thyroid Hormone Resistance (THRA)
Multiple Endocrine Neoplasia 2A (RET)			Thyroid Hormone Resistance (THRB)
Multiple Exostosis 1 (EXT1)			Thyrotropin Deficiency (TSHB)
Multiple Exostosis 2 (EXT2)			
Neurofibromatosis 1 (NF1)		**Renal**	
Neurofibromatosis 2 (NF2)		Alport (COL4A5)	Hypomagnesemia (CLDN16)
Nijmegen Breakage 1 (NBS1)		Bartter (SLC12A1)	Nephronophthisis 1 (NPHP1)
Nucleoporin (NUP214)		Congenital Nephrosis (NPHS1)	Polycystic Kidney 1 (PKD1)
Tumor suppressor (P53)		Dent (CLCN5)	Pseudohypoaldosteronism (NR3C2)
Tumor suppressor (PTEN)		Diabetes Insipidus, Nephrogenic (AQP2)	Williams-Beuren (ELN)
Oncogene (RAS)		Gitelman (SLC12A3)	
Oncogene (REL)		Hyperoxaluria 1 (AGXT)	
Retinoblastoma (RB1)		Hypophosphatasia (ALPL)	
Peutz-Jeghers (STK11)		Polycystic Kidney 2 (PKD2)	
Stem Cell Leukemia (TAL1)		Renal Tubular Acidosis (ATP6B1)	
Tuberous Sclerosis 1 (TSC1)		Hypophosphatemia (PHEX)	
Tuberous Sclerosis 2 (TSC2)			
Von Hippel Lindau (VHL)		**Hematological**	
Xeroderma Pigmentosum A (XPA)		Chediak-Higashi (CHS1)	Essential Thrombocythemia (THPO)
Xeroderma Pigmentosum B (ERCC3)		Diamond-Blackfan Anemia (RPS19)	Lymphohistiocytosis (PRF1)
Xeroderma Pigmentosum D (XPD)		G6PD Deficiency (G6PD)	Hemophilia A (Factor VIIIC)
Xeroderma Pigmentosum F (XPF)		Hereditary Spherocytosis (ANK1)	Hemophilia B (Factor IX)
Xeroderma Pigmentosum G (XPG)		Megaloblastic Anemia (SLC19A2)	Osler-Rendu-Weber (ENG)
Neurological		Myeloperoxidase Deficiency (MPO)	α-Thalassemia (HBA1)
Adrenoleukodystrophy (ABCD1)	Ceroid Lipofuscinosis (CLN2)	Thrombophilia (PLG)	β-Thalassemia (HBB)
Alzheimer (PS1)	Charcot-Marie-Tooth 1A (PMP22)	Wiscott-Aldrich (WAS)	δ-Thalassemia (HBD)
Alzheimer (APP)	Charcot-Marie-Tooth 1B (MPZ)		ε-Thalassemia (HBE)
Amyotrophic Lateral Sclerosis (SOD1)	Creutzfeldt-Jakob (PRNP)		Von Willebrand (VWF)
Angelman (UBE3A)	Emery-Dreifuss Muscular Dystrophy (EMD)		
Aniridia (PAX6)	Fukuyama Muscular Dystrophy (FCMD)	**Immune**	
Best Macular Dystrophy (VMD2)	Limb-Girdle Muscular Dystrophy 2E (BSG)	Bare Lymphocyte (ABCB3)	Bare Lymphocyte (RFXAP)
Ceroid Lipofuscinosis (PPT)	Machado-Joseph (SCA3/MJD1)	Bare Lymphocyte (RFX5)	Bare Lymphocyte (MHC2TA)
Ceroid Lipofuscinosis (CLN3)	Myotonic Dystrophy (DM1)	Bruton Agammaglobulinemia (BTK)	Chronic Granulomatous (NCF1)
Choroideremia (CHM)	Naito-Oyanagi (DRPLA)	Chronic Granulomatous (CYBB)	Immunodeficiency (CD3G)
Deafness, Hereditary (MYO15)	Nemaline Myopathy 2 (NEB)	Immunodeficiency (DNA ligase 1)	Severe Combined Immunodeficiency (IL2RG)
Deafness, X-linked (TIMM8A)	Neuroaminidase Deficiency(NEU1)	Severe Combined Immunodeficiency (JAK3)	Severe Combined Immunodeficiency (IL7R)
Deafness, Autosomal Dominant (DIAPH1)	Norrie (NDP)	Severe Combined Immunodeficiency (ZAP70)	Severe Combined Immunodeficiency (RAG1)
Dementia, Multi-Infarct (NOTCH3)	Ocular Albinism (OA1)		Severe Combined Immunodeficiency (RAG2)
Duchenne Muscular Dystrophy (DMD)	Parkinson (SNCA)		T-cell immunodeficiency (CD3E)
Emery-Dreifuss Muscular Dystrophy (LMNA)	Myoclonic Epilepsy (CSTB)		X-linked Lymphoproliferative (SH2D1A)
Familial Encephalopathy (PI12)	Retinitis Pigmentosa 3 (RPGR)		
Fragile X (FMR1)	Retinitis Pigmentosa 2 (RP2)	**Metabolic**	
Friedreich Ataxia (FRDA)	Spinal Cerebellar Ataxia 1 (SCA1)	CPT2 Deficiency Myopathy (CPT2)	Hypocalciuric Hypercalcemia (CASR)
Frontotemporal Dementia (TAU)	Spinal Cerebellar Ataxia 7 (SCA7)	Carnitine Deficiency (SLC22A5)	Hemochromatosis (HFE)
Huntington (HD)	Usher 2A (USH2A)	Citrullinemia (ASS)	Lesch-Nyhan (HPRT1)
Limb-Girdle Muscular Dystrophy 2A (CAPN3)		Cystinuria (SLC3A1)	
Limb-Girdle Muscular Dystrophy 2B (DYSF)		Galactokinase Deficiency (GALK1)	
Lissencephaly, X-linked (DCX)		Gaucher (GBA)	
Lowe Oculocerebrorenal (OCRL)		Liddle (SCNN1G)	
Miller-Dieker Lissencephaly (PAF)		Liddle (SCNN1B)	
Myotubular Myopathy 1 (MTM1)		Menkes (ATP7A)	
Oculopharyngeal Muscular Dystrophy (PABPN1)		Niemann-Pick Type C (NPC1)	
Oguchi Type 2 Nightblindness (RH KIN)		Severe Combined Immunodeficiency (ADA)	
Parkinson, Juvenile (PARK2)		Trimethylaminuria (FMO3)	
Parkinson (UCHL1)		Variegate Porphyria (PPOX)	
Spinal Cerebellar Ataxia 2 (SCA2)		Wernicke-Korsakoff (TKT)	
Spinal Cerebellar Ataxia 6 (CACNA1A)			
Spinal Muscular Atrophy (SMN1)		**Other**	
Stargardt (ABCA4)		α-1-Antitrypsin Deficiency (PI)	Pulmonary Alveolar Proteinosis A(SFTPB)
Tay-Sachs (HEXA)		Corneal Dystrophy (TGFBI)	Congenital Chloride Diarrhea (DRA)
Thomsen Myotonia Congenita (CLCN1)		Cystic Fibrosis (ABCC7)	Ehlers-Danlos Type IV (COL3A1)
Wilson (ATP7B)		Cystinosis (CTNS)	Familial Mediterranean Fever (MEFV)
Malformation Syndromes		Darier-White (SERCA)	Keratoderma (KRT9)
Achondroplasia (FGFR3)	Aarskog-Scott (FGD1)	Finnish Amyloidosis (GSN)	Mcleod (XK)
Alagille (JAG1)	Beckwith-Wiedemann (CDKN1C)	Glycerol Kinase Deficiency (GK)	Monilethrix (KRTHB1)
Barth (TAZ)	Cerebral Cavernous Malformation (CCM1)	Hereditary Pancreatitis (PRSS1)	Monilethrix (KRTHB6)
Chondrodysplasia Punctata 1 (ARSE)	Cockayne 1 (CKN1)	Hermansky-Pudlak (HPS)	Osteogenesis Imperfecta (COL1A1)
Orofacial Cleft (OFC1)	Hand-Foot-Genital (HOXA13)	Hyperexplexia (GLRA2)	Spondyloepiphyseal Dysplasia (COL2A1)
Coffin-Lowry (RPS6KA3)	Immunodeficiency, Facial Anomalies (DNMT3B)	Open Angle Glaucoma A (GLC1A)	Vohwinkel (LOR)
Diastrophic Dysplasia (SLC26A2)	Laterality, X-linked (ZIC3)	Marfan (FBN1)	
Ectrodactyly EEC 3 (P63)	Opitz (MID1)	Wolfram (WFS1)	
Greig Cephalopolysyndactyly (GLI3)	Rubinstein-Taybi (CREBBP)		
Holoprosencephaly 3 (SHH)	Septooptic Dysplasia (HESX1)		
Holoprosencephaly 2 (SIX3)	Treacher Collins-Franceschetti (TCOF1)		
Holt-Oram (TBX5)	Venous Malformations (TEK)		
Kallmann (KAL1)			
Melnick-Fraser (EYA1)			
Nail-Patella (LMX1B)			
Renal-Coloboma (PAX2)			
Rieger Type 1 (PITX2)			
Saethre-Chotzen (TWIST)			
Simpson-Golabi-Behmel (GPC3)			
Townes-Brocks (SALL1)			
Waardenburg (PAX3)			
Zellweger (PEX1)			

The 287 human disease genes are grouped according to disease classification, with genes that possess likely *Drosophila* homologues listed in the lefthand column and genes that lack apparent *Drosophila* homologues listed in the righthand column within each disease category. In most cases, the disease name is given, followed by the human gene abbreviation in parentheses, although some cancer genes are listed by gene name only due to their pleiotropic involvement in numerous cancers. Additional information about this survey, including the GenBank accession numbers, OMIM numbers, and BLAST E value scores for the human disease genes tested can be found in Rubin et al. (2000) and at http://www.sciencemag.org/feature/data/1049664t1.shl. Our current estimate of 178 out of 287 disease gene homologues in the fly genome differs slightly from the estimate of 177 out of 289 reported in Rubin et al. (2000) due to the deletion of two redundant entries in the initial survey (CDKN2C and P16-INK4A; see Fig. 1 of Rubin et al., 2000), and the transfer of the Von Hippel Lindau (VHL) gene into the homologue detected category.

it was useful to perform reverse BLAST searches, in which the best fly candidate homologues were used as query sequences to search a database of all available human sequences. If the fly protein showed a significantly better match to the original human query than to other related proteins, we considered it to be a likely homologue. This approach was successful in confirming that *Drosophila* has homologues of Peutz-Jeghers kinase, leutenizing hormone/choriogonadotropin receptor precursor (mutated in Leydig cell hypoplasia), Wolfram syndrome protein, Kallmann disease protein, and numerous other human disease proteins.

Limitations of the Analysis

In any large-scale computational survey of this type, it is inevitable that some inaccuracies and errors will arise. Although we performed extensive manual cross-checking of our results, our effort should be considered a first-pass survey of human disease-related loci in *Drosophila*, which will be corrected and refined as the completed fly genome sequence is further analyzed. Certain aspects of the sequence data would be expected to cause potential homologues to be missed. The genome sequence currently contains >1,000 relatively small gaps, some of which could contain disease gene homologues that we failed to detect. The Machado-Joseph disease gene (*SCA3/MJD*) has an identified homologue in *C. elegans* but not in *D. melanogaster*, for example, suggesting that a fly counterpart could exist but might have gone undetected in our survey. Furthermore, the gene prediction algorithms used to analyze the fly genome sequence are known to be error-prone, incorrectly predicting 5′ and 3′ coding exons and intron–exon boundaries, splitting single genes into two or more predicted genes, and merging adjacent genes into a single predicted transcript (Reese et al., 2000). Such errors might be expected to result in artificially lower BLAST scores, incorrect domain organizations for some predicted proteins, or even the failure of an actual fly protein to be predicted with enough accuracy to be detected at all by our BLAST search protocol. Finally, for human disease genes that have not been well studied or are completely novel, the lack of supplementary data makes it difficult to evaluate manually the BLAST alignments and InterPro domain structure predictions, or to use other biological clues to guide the analysis.

Results of the Survey

In our survey of 287 human disease genes, a total of 178 (62%) were found to have likely homologues in *Drosophila*. Inspection of the different classes of genes indicates that some categories are better represented than others (Table I). Categories with a high representation of homologues in *Drosophila* include the genes for cancer (47 of 65, 72%), neurological diseases (38 of 59, 64%), malformation syndromes (22 of 34, 65%), metabolic diseases (14 of 17, 82%) and renal diseases (11 of 16, 69%). A small number of genes implicated in cardiac diseases all had likely homologues in *Drosophila* (6 of 6). Underrepresented in the fly genome were likely homologues of the genes implicated in endocrine (12 of 31, 39%) and hematological diseases (8 of 18, 44%) as well as diseases of the immune system (7 of 17, 41%).

Cancer Genes in Drosophila

A high proportion of cancer genes have homologues in *Drosophila*. Many of these homologues have been identified by workers in the field who have cloned these genes using hybridization to mammalian sequences or more recently by searching for sequence similarity in databases of *Drosophila* genomic sequence or EST sequences. Our first look at the complete sequence of the genome allowed us to identify additional homologues and also to make tentative statements about classes of cancer genes that appear to be absent in *Drosophila*. Two groups of cancer genes that appear to be absent in *Drosophila* are the genes mutated in breast cancer (*BRCA1* and *BRCA2*) and the genes mutated in Fanconi's anemia (*FANCA*, *FANCC*, and *FANCG*), a disease characterized by anemia, chromosomal instability and a predisposition to cancer. Also apparently absent are homologues of *mdm2* and *p19ARF* (or *p14ARF*) which regulate the levels of the p53 protein in mammalian cells (Pomerantz et al., 1998; Zhang et al., 1998).

One of the genes identified as a result of the sequencing effort is a convincing homologue of the human *menin* gene (*MEN1*). Mutations in *menin* are found in the multiple endocrine neoplasia type 1 syndrome, a familial cancer syndrome characterized by varying combinations of tumors in the parathyroid glands, the pancreatic islets, the anterior pituitary, as well as a variety of other tissues. These tumors often secrete the hormones of the tissue of origin (e.g., insulin and growth hormone). The *menin* gene encodes a nuclear protein of 610 amino acids which is thought to bind to and inhibit the function of the JunD transcription factor in humans (Agarwal et al., 1999). The biological function of Menin and the relationship between *menin* mutations and endocrine tumors are poorly understood. An alignment of the human and the putative *Drosophila* Menin proteins is shown in Fig. 2. The proteins display 34% identity and 47% similarity over their entire length. Of a panel of eleven naturally occurring germline or somatic missense mutations in *MEN1* (Agarwal et al., 1999), ten affect residues that are conserved in the *Drosophila* protein, suggesting that residues crucial for function are conserved between the two proteins. In contrast, the *C. elegans* genome does not appear to have a *menin* homologue.

Also identified in our survey was a *Drosophila* p53 family member, the sequence and function of which have been described in two recent publications (Brodsky et al., 2000; Ollmann et al., 2000). The levels of mammalian p53 are thought to be regulated by two distinct pathways. One involves Mdm2 and p19ARF, which do not appear to have homologues in the *Drosophila* genome, as noted above. The second pathway, which leads to the phosphorylation of p53, appears to be activated in response to DNA damage and involves the ATM and Chk2 kinases (reviewed by Caspari, 2000). Homologues of the *ATM* and *chk2* genes are clearly found in *Drosophila*, indicating that this pathway may represent the more evolutionarily ancient mode of p53 regulation. Consistent with this notion, *Drosophila* p53 appears to be necessary for apoptosis in response to DNA damage (Brodsky et al., 2000; Ollmann et al., 2000).

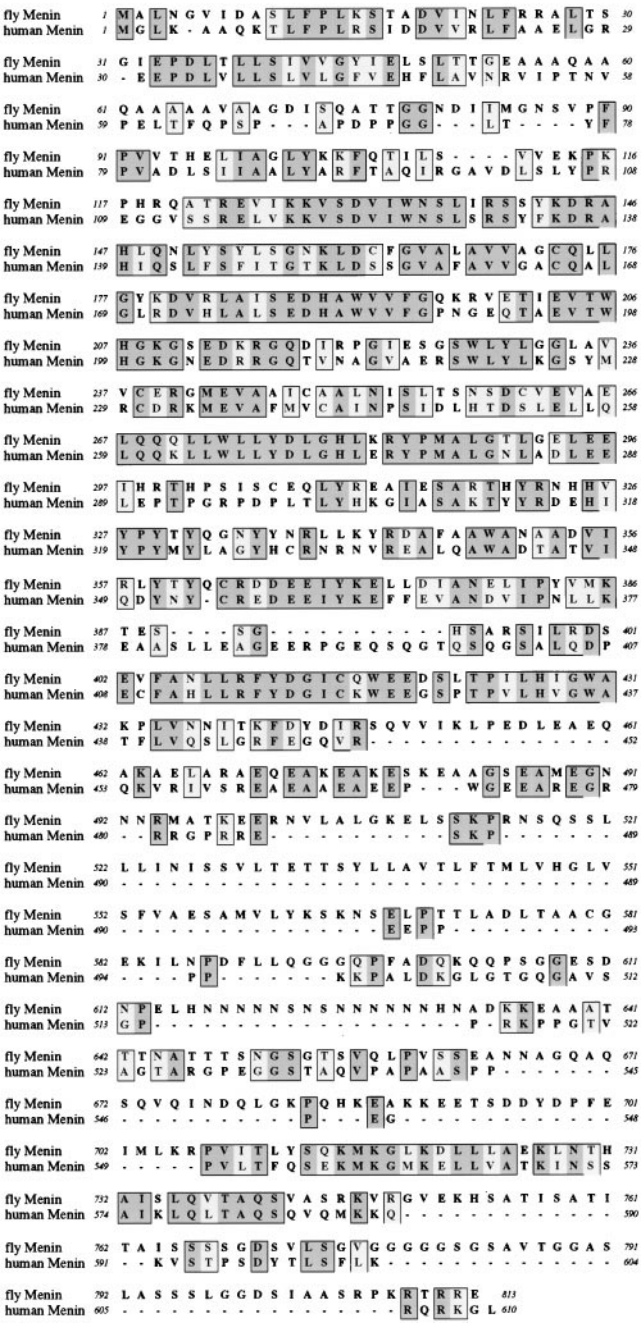

Figure 2. Alignment of the human and *Drosophila* Menin proteins. The human (bottom line) and *Drosophila* (top line) Menin proteins were aligned using the MacVector 6.5 ClustalW program with the BLOSUM 30 matrix and an open gap penalty of 10. The fly protein is 34% identical and 47% similar to the human protein over its entire length.

Furthermore, only one p53-related protein is predicted in the fly genome, indicating that it may represent an ancestral prototype of the mammalian p53/p63/p73 family.

Mutations in the two related genes *EXT1* and *EXT2* have been implicated in human multiple exostosis syndromes, which are characterized by abnormal bony outgrowths. The *Drosophila* homologue of *EXT1* is the *tout-velu* gene, which has been shown to be necessary for the diffusion of the Hedgehog protein that functions as a morphogen in many tissues (Bellaiche et al., 1998). The sequencing effort has identified a homologue of the *EXT2* gene. One might speculate that this *EXT2* homologue may have a role in facilitating the diffusion of Hedgehog or perhaps a different morphogen, such as a Wnt protein family member, in *Drosophila* tissues.

A likely homologue of the STK11 kinase was also identified (Fig. 3). This gene is mutated in Peutz-Jeghers disease (Jenne et al., 1998), in which affected individuals present clinically with oral pigmentation and multiple intestinal polyps. Its mechanism of action and likely substrates are not known, and the identification of a *Drosophila* homologue should enable genetic screens analogous to those that have been employed to successfully identify signaling components downstream of protein kinases such as Sevenless and Raf (Simon et al., 1991; Dickson et al., 1996).

Ataxia telangiectasia (ATM) is a syndrome characterized by loss of coordination (ataxia) as well as multiple cutaneous capillary malformations (telangiectasia). The gene product of the *ATM* locus has been implicated in activating signaling pathways in response to DNA damage. The *Drosophila mei-41* gene has previously been shown to function as an ATM homologue in many respects (Hari et al., 1995). The *Drosophila* genome contains an additional gene identified by the genomic sequencing effort that appears to be more similar to *ATM* than is *mei-41* (Sekelsky et al., 2000). Thus these two genes in *Drosophila* may have roles analogous to those of the *ATM* and the *ATM-related* (*ATR*) genes in mammals.

Neurological Genes in Drosophila

Out of 59 human neurological genes surveyed, 38 appear to be conserved in *Drosophila* (Rubin et al., 2000). Several of these genes were previously identified by molecular genetic studies in the fly, such as *Notch* (CADASIL syndrome), *Presenilin* (Alzheimer's disease), *Amyloid protein precursor-like* (Alzheimer's disease), *diaphanous* (nonsyndromic deafness, premature ovarian failure), *Superoxide dismutase 1* (amyotrophic lateral sclerosis), *rhodopsin kinase* (Oguchi stationary nightblindness type 2), and *huntingtin* (Huntington's disease). Functional studies of these fly loci have already provided insights into the molecular mechanisms regulated by their mammalian counterparts, in some cases recapitulating aspects of human neurological disorders and aging with startling fidelity (reviewed in Fortini and Bonini, 2000). The genome sequencing project has uncovered a number of additional fly homologues for other important neurological disease loci, holding promise for similarly informative analyses of these genes in *Drosophila.* Among the most highly conserved neurological disease genes in the fly are loci implicated in adrenoleukodystrophy (*ABCD1*), Angelman syndrome (*UBE3A*), Best macular dystrophy (*VMD2*), Duchenne muscular dystrophy (*DMD*), Emery-Dreifuss muscular dystrophy (*LMNA*), Fragile-X syndrome (*FMR1*), Limb girdle muscular dystrophy 2A and 2B (*CAPN3* and *DYSF*), Lowe oculocerebrorenal syndrome (*OCRL*), Miller-Dieker lissencephaly (*PAF*), juvenile-onset Parkinson's disease (*Parkin*), Stargardt disease

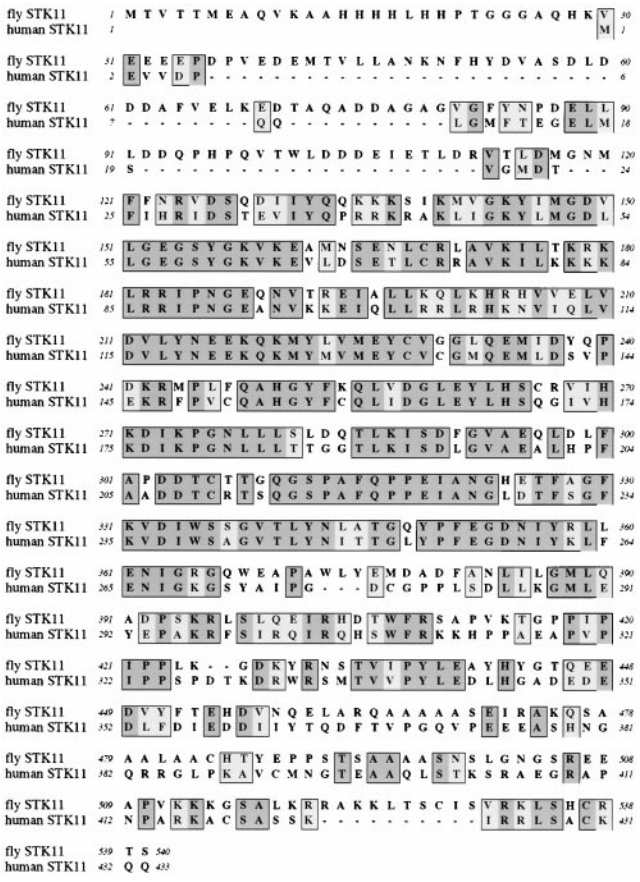

Figure 3. Alignment of the human and *Drosophila* Peutz-Jeghers STK11 kinases. The human (bottom line) and *Drosophila* (top line) Peutz-Jeghers disease STK11 kinases were aligned using the MacVector 6.5 ClustalW program with the BLOSUM 30 matrix and an open gap penalty of 10. The fly protein is 43% identical and 56% similar to the human protein over its entire length.

(*ABCA4*), and Tay-Sachs disease (*HEXA*). Our survey also identified a less well conserved yet convincing homologue of the human *tau* gene, mutations in which have recently been found to cause frontotemporal dementia with parkinsonism (reviewed in Lee and Trojanowski, 1999).

Also worth noting are the human neurological disease loci that we failed to detect in the fly genome. We were unable to find a fly counterpart of human prion protein gene (*PRNP*), despite extensive BLAST searches using different prion protein segments, or homologues of the Charcot-Marie-Tooth syndrome 1A and 1B loci, or the Parkinson's disease gene encoding α-synuclein. Directed neuronal expression of human α-synuclein in aged transgenic flies leads to the formation of Lewy bodies and other morphological defects reminiscent of human Parkinson's disease (Feany and Bender, 2000), indicating that the cellular mechanisms involved in Parkinson's disease neuropathophysiology may be conserved in *Drosophila* even though α-synuclein itself may be absent.

Several genes implicated in expanded polyglutamine repeat diseases, including *huntingtin*, *FRDA* (Friedreich ataxia), *SCA2* and *SCA6* (spinal cerebellar ataxia loci) are conserved in the *Drosophila* genome, although others, such as putative homologues of *SCA1*, *SCA3/MJD* (Machado-Joseph Disease) and *SCA7* were not found. Transgenic *Drosophila* models of expanded polyglutamine repeat diseases have already been developed by directed expression of human Huntingtin and SCA3/MJD proteins and shown to reproduce the nuclear inclusions of expanded repeat proteins that are characteristic of this class of illness (Jackson et al., 1998; Warrick et al., 1998). The identification of single-copy fly homologues of different expanded polyglutamine repeat genes raises the possibility of using *Drosophila* genetics to eliminate the function of the endogenous fly proteins. Such an approach should help elucidate the normal activities of these proteins and to determine whether the loss of normal protein activity may contribute to human disease pathophysiology, independent of or in conjunction with the expanded polyglutamine repeat inclusions. Moreover, the ability to produce easily observable phenotypes in *Drosophila* by either mutagenesis or overexpression of human disease homologues will allow the powerful genetic screening methods of this organism to be exploited to help identify protein partners and biochemical pathways involved in human pathophysiology. This approach has already met with some early successes, such as the suppression of polyglutamine repeat expansion-induced phenotypes by overexpression of molecular chaperones in *Drosophila* (Warrick et al., 1999; Kazemi-Esfarjani and Benzer, 2000).

Malformation Syndrome and Metabolic Disorder Genes in Drosophila

Almost two thirds of the genes implicated in malformation syndromes have likely *Drosophila* homologues. This finding is not surprising since many of these genes function in defining the body plan in the embryo and in patterning specific tissues. Indeed, some of these genes were originally cloned by virtue of their sequence similarity to patterning genes in *Drosophila*. These include the *Sonic hedgehog* gene (holoprosencephaly 3) and the *Eyes absent 1* (*EYA1*) gene which is mutated in Melnick-Fraser brachiootorenal dysplasia. Also frequently conserved were genes implicated in diseases caused by abnormalities in metabolism, presumably reflecting a conservation of many of the metabolic pathways between *Drosophila* and humans. Surprisingly, a likely homologue for the gene implicated in Lesch-Nyhan syndrome, hypoxanthine guanine phosphoribosyl transferase (*HPRT1*), was not found in the fly genome sequence. *C. elegans* appears to have an *HPRT* homologue, and many other enzymes involved in purine biosynthesis in mammals have *Drosophila* homologues. Our failure to detect a fly *HPRT* homologue might therefore represent a likely candidate for a gene that lies in one of the small gaps still remaining in the assembled *Drosophila* genome sequence.

Hematological and Immune Disease Genes in Drosophila

Genes that function in the mammalian immune system and in mammalian blood cells are significantly underrepresented in *Drosophila.* Not surprisingly, genes that function in acquired immunity, such as *RAG1* and *RAG2*, which are involved in immunoglobulin gene rearrange-

ment, are not found in *Drosophila*. The absence of RAG-like proteins is consistent with the fact that the *Drosophila* genome is not known to undergo any programmed DNA rearrangements, unlike several lower organisms such as yeast and bacteria or the mammalian immune system. In contrast, some of the genes that function in signaling pathways in hematopoietic cells, including *BTK* and *JAK3*, possess *Drosophila* homologues. Genes that function in oxygen transport via erythrocytes, such as the hemoglobin genes, and genes involved in blood coagulation do not have *Drosophila* homologues, reflecting fundamental differences in physiology between the two organisms. Flies do not possess a hemoglobin-based oxygen delivery system or clotting system resembling those of human blood, and they instead rely upon a branching tracheal system in which oxygen is delivered directly to tissues from the atmosphere, and in which soft tissues are protected from injury by a durable exoskeleton. However, human hematological disease genes that encode components or likely regulators of the cytoskeleton, such as the genes for Wiskott-Aldrich syndrome (*WAS*) and hereditary spherocytosis (*ANK1*), do possess *Drosophila* homologues.

Endocrine and Renal Disease Genes in Drosophila

Of the genes implicated in endocrine disorders, many components of insulin-mediated signaling pathways are found in *Drosophila*. Most other endocrine pathways do not appear to be conserved. Finally, despite major dissimilarities between the vertebrate kidney and insect Malpighian tubules, a significant proportion of genes mutated in congenital renal disorders have *Drosophila* homologues. Many of these genes encode proteins involved in fluid and electrolyte transport and their identification might encourage the study of Malpighian tubule development and function to gain insight into certain human kidney disorders.

In conclusion, *Drosophila* appears to represent a particularly good model organism for the study of genes implicated in many cancers, neurological disease, malformation syndromes, metabolic disorders and some renal diseases. Specific endocrine, immunological and hematological disease genes may require vertebrate model organisms such as mice, since relatively few of the known human disease genes in these categories are present in *Drosophila*. Most promisingly, our search for fruit fly homologues of 287 known human disease genes leads us to conclude that as additional human disease genes are discovered, it is more likely than not that a counterpart will be found in the *Drosophila* genome.

We thank Celera Genomics and the Berkeley *Drosophila* Genome Project for their hospitality and generosity in providing access to the *Drosophila* genome sequence data, and to Oxana K. Pickeral, Jiong Zhang, Peter M. Kuehl, Gena Heidary, and Glen A. Seidner for assistance with the protein database searches and sequence alignments.

M.E. Fortini is supported by grants from the National Institutes of Health (R01AG14583), the Alzheimer's Association, the Life and Health Insurance Medical Research Fund, the American Heart Association, and Merck & Co., Inc., and I.K. Hariharan is supported by grants from the National Institutes of Health (R01NS36084 and R01EY11632) and the American Cancer Society.

Submitted: 5 June 2000
Revised: 26 June 2000
Accepted: 26 June 2000

References

Adams, M.D., S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science*. 287:2185–2195.

Agarwal, S.K., S.C. Guru, C. Heppner, M.R. Erdos, R.M. Collins, S.Y. Park, S. Saggar, S.C. Chandrasekharappa, F.S. Collins, A.M. Spiegel, et al. 1999. Menin interacts with the AP1 transcription factor JunD and represses JunD-activated transcription. *Cell*. 96:143–152.

Antonarakis, S.E., and V.A. McKusick. 2000. OMIM passes the 1,000-disease-gene mark. *Nat. Genet.* 25:11.

Apweiler, R., T.K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, M.D.R. Croning, R. Durbin, et al. 1999. InterPro integrated resource of protein domains and functional sites. http://www.ebi.ac.uk/interpro/

Banfi, S., G. Borsani, E. Rossi, L. Bernard, A. Guffanti, F. Rubboli, A. Marchitiello, S. Giglio, E. Coluccia, M. Zollo, et al. 1996. Identification and mapping of human cDNAs homologous to *Drosophila* mutant genes through EST database searching. *Nat. Genet.* 13:167–174.

Bellaiche, Y., I. The, and N. Perrimon. 1998. *Tout-velu* is a *Drosophila* homologue of the putative tumour suppressor *EXT-1* and is needed for Hh diffusion. *Nature*. 394:85–88.

Berkeley *Drosophila* Genome Project. 1999. Sequence similarities of human inherited disease genes with *Drosophila* DNA. http://www.fruitfly.org/sequence/human2fly.html.

Brodsky, M.H., W. Nordstrom, G. Tsang, E. Kwan, G.M. Rubin, and J.M. Abrams. 2000. *Drosophila* p53 binds a damage response element at the *reaper* locus. *Cell*. 101:103–113.

Caspari, T. 2000. How to activate p53. *Curr. Biol.* 10:R315–R317.

Dickson, B.J., A. van der Straten, M. Domínguez, and E. Hafen. 1996. Mutations modulating Raf signaling in *Drosophila* eye development. *Genetics*. 142:163–171.

Feany, M.B., and W.W. Bender. 2000. A *Drosophila* model of Parkinson's disease. *Nature*. 404:394–398.

Fortini, M.E., and N.M. Bonini. 2000. Modeling human neurodegenerative diseases in *Drosophila*: on a wing and a prayer. *Trends Genet.* 16:161–167.

Hari, K.L., A. Santerre, J.J. Sekelsky, K.S. McKim, J.B. Boyd, and R.S. Hawley. 1995. The *mei-41* gene of *D. melanogaster* is a structural and functional homolog of the human ataxia telangiectasia gene. *Cell*. 82:815–821.

Jackson, G.R., I. Salecker, X. Dong, X. Yao, N. Arnheim, P.W. Faber, M.E. MacDonald, and S.L. Zipursky. 1998. Polyglutamine-expanded human Huntingtin transgenes induce degeneration of *Drosophila* photoreceptor neurons. *Neuron*. 21:633–642.

Jenne, D.E., H. Reimann, J. Nezu, W. Friedel, S. Loff, R. Jeschke, O. Müller, W. Back, and M. Zimmer. 1998. Peutz-Jeghers syndrome is caused by mutations in a novel serine threonine kinase. *Nat. Genet.* 18:38–43.

Kazemi-Esfarjani, P., and S. Benzer. 2000. Genetic suppression of polyglutamine toxicity in *Drosophila*. *Science*. 287:1837–1840.

Lee, V.M.-Y., and J.Q. Trojanowski. 1999. Neurodegenerative tauopathies: human disease and transgenic mouse models. *Neuron*. 24:507–510.

Ollmann, M., L.M. Young, C.J. Di Como, F. Karim, M. Belvin, S. Robertson, K. Whittaker, M. Demsky, W.W. Fisher, A. Buchman, et al. 2000. *Drosophila p53* is a structural and functional homolog of the tumor suppressor *p53*. *Cell*. 101:91–101.

Pickeral, O.K., J. Li, I. Barrow, M.S. Boguski, W. Makalowski, and J. Zhang. 2000. Classical oncogenes and tumor suppressor genes: a comparative genomics perspective. *Neoplasia*. In press.

Pomerantz, J., N. Schreiber-Agus, N.J. Liégeois, A. Silverman, L. Alland, L. Chin, J. Potes, K. Chen, I. Orlow, H.-W. Lee, et al. 1998. The *Ink4a* tumor suppressor gene product, $p19^{Arf}$, interacts with MDM2 and neutralizes MDM2's inhibition of p53. *Cell*. 92:713–723.

Reese, M.G., G. Hartzell, N.L. Harris, U. Ohler, J.F. Abril, and S.E. Lewis. 2000. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* 10:483–501.

Rubin, G.M., M.D. Yandell, J.R. Wortman, G.L.G. Miklos, C.R. Nelson, I.K. Hariharan, M.E. Fortini, P.W. Li, R. Apweiler, W. Fleischmann, et al. 2000. Comparative genomics of the eukaryotes. *Science*. 287:2204–2215.

Sekelsky, J.J., M.H. Brodsky, and K.C. Burtis. 2000. DNA repair in *Drosophila*: insights from the *Drosophila* genome sequence. *J. Cell Biol.* 150:F31–F36.

Simon, M.A., D.D.L. Bowtell, G.S. Dodson, T.R. Laverty, and G.M. Rubin. 1991. Ras1 and a putative guanine nucleotide exchange factor perform crucial steps in signaling by the Sevenless protein tyrosine kinase. *Cell*. 67:701–716.

Walker, D.R., J.P. Bond, R.E. Tarone, C.C. Harris, W. Makalowski, M.S. Boguski, and M.S. Greenblatt. 1999. Evolutionary conservation and somatic mutation hotspot maps of *p53*: correlation with p53 protein structural and functional features. *Oncogene*. 18:211–218.

Warrick, J.M., H.L. Paulson, G.L. Gray-Board, Q.T. Bui, K.H. Fischbeck, R.N. Pittman, and N.M. Bonini. 1998. Expanded polyglutamine protein forms nuclear inclusions and causes neural degeneration in *Drosophila*. *Cell*. 93:939–949.

Warrick, J.M., H.Y.E. Chan, G.L. Gray-Board, Y. Chai, H.L. Paulson, and N.M. Bonini. 1999. Suppression of polyglutamine-mediated neurodegeneration in *Drosophila* by the molecular chaperone HSP70. *Nat. Genet.* 23:425–428.

Zhang, Y., Y. Xiong, and W.G. Yarbrough. 1998. ARF promotes MDM2 degradation and stabilizes p53: *ARF-INK4a* locus deletion impairs both the Rb and p53 tumor suppression pathways. *Cell*. 92:725–734.