ELSEVIER

Gene 195 (1997) 187–193

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

One-hundred and five new potential *Drosophila melanogaster* genes revealed through STS analysis

Christos Louis [a,b,*], Encarna Madueño [a,c], Juan Modolell [c], Mahmoud M. Omar [d], George Papagiannakis [a], Robert D.C. Saunders [d], Charalambos Savakis [a,e], Inga Sidén-Kiamos [a], Lefteris Spanos [a], Pantelis Topalis [a,b], Yong Qing Zhang [f], Michael Ashburner [f], Panayotis Benos [a,b,c], Viatcheslav N. Bolshakov [a,g], Peter Deak [d,h], David M. Glover [d], Siegrun Herrmann [c,f], Fotis C. Kafatos [a,b,g]

[a] *Institute of Molecular Biology and Biotechnology, FORTH, P.O. Box 1527, Heraklion, 71110, Crete, Greece*
[b] *Department of Biology, University of Crete, P.O. Box 1425, Heraklion, 71110, Crete, Greece*
[c] *Centro di Biologia Molecular Severo Ochoa, CSIC, Universidad Autonoma de Madrid, Madrid 28049, Spain*
[d] *Department of Anatomy and Physiology and CRC Cell Cycle Group, University of Dundee, Dundee, DD1 4HN, UK*
[e] *Division of Medical Sciences, Medical School, University of Crete, P.O. Box 1425, Heraklion, 71110, Crete, Greece*
[f] *Department of Genetics, University of Cambridge, Cambridge, CB2 2EH, UK*
[g] *European Molecular Biology Laboratory, Heidelberg, D-69117, Germany*
[h] *Institute of Biochemistry, Biological Research Center, P.O. Box 521, Szeged, H-6701, Hungary*

Received 20 December 1996; accepted 31 January 1997

Abstract

Complementation analysis had suggested that the *Drosophila melanogaster* genome contains approximately 5000 genes, but it is now generally accepted that the actual number is several times as high. We report here an analysis of 1788 anonymous sequence tagged sites (STSs) from the European *Drosophila* Genome Project (EDGP), totalling 463 kb. The data reveal a substantial number of previously undescribed potential genes, amounting to 6.1% of the number of *Drosophila* genes already in the sequence databases. © 1997 Elsevier Science B.V.

Keywords: Genome mapping; Contig analysis; Orphan genes; Gene families

1. Introduction

Although in classical terms genes are defined by functional analysis, genes more loosely identified as potential protein coding entities provide a significant alternative description of the genetic potentialities of a given organism. It is now widely appreciated that, because of the existence of redundant genes and genes with subtle mutant phenotypes, many more genes can be detected in eukaryotic organisms by molecular analysis than by classical genetics. For example, the yeast *Saccharomyces cerevisiae* was estimated to have about 1000 genes (Mortimer et al., 1989), but its total genomic sequence has revealed 6275 potential open reading frames (ORFs) (Dujon, 1996). The gene number can be expected to be higher in more complex organisms with larger genomes. Recent releases of the FlyBase (1996), the comprehensive database for the fruitfly, list about 9000 genes that have been identified so far using a variety of molecular as well as genetic procedures; Miklos and Rubin (1996), using a statistical re-evaluation of sequenced genes, arrived at a total number of 12 000. The 51 Mb of genomic sequence that is currently available from the worm *Caenorhabditis elegans* encompasses 9700 protein coding genes, and it is anticipated that its complete genome sequence might reveal 15 000 to 16 000 genes (S. Jones and R. Durbin, personal communication).

While total sequencing gives the full measure of a

* Corresponding author. Tel. +30 81 391119; Fax +30 81 391104; e-mail: louis@myia.imbb.forth.gr

Abbreviations: bp, base pair(s); EDGP, European *Drosophila* Genome Project; EST, expressed sequence tags; hyp., hypothetical; id., identity; kb, kilobase(s), or 1000 bp; ORF, open reading frame; put., putative; sim., similar; STS, sequenced tagged site(s).

PII S0378-1119(97)00138-8

genome's protein coding capacity, partial sequencing can also reveal previously undiscovered genes. In *Drosophila melanogaster* the gene number had been estimated as approximately 5000, by extrapolation from local saturation mutagenesis experiments and from the genetic evidence supporting the one gene–one chromosomal band hypothesis (Garcia-Bellido and Ripoll, 1978). However, over a decade ago it was already recognized that the 'question of the total gene number in *Drosophila* will, no doubt, eventually be solved by molecular analysis, not by statistical analysis of mutation data or saturation studies' (Lefevre and Watkins, 1986). More recent evidence from a variety of procedures, including transposon tagging and enhancer traps, homology screens and chromosomal walks, has revised upwards the estimate of gene number in the fruitfly; extrapolation from sequencing substantial chromosomal regions has suggested an actual number that might be several times as high as the earlier estimates (Miklos and Rubin, 1996). Widely scattered *Drosophila* chromosomal sequence tagged sites (STSs; Olson et al., 1989) are being determined as part of physical mapping projects. The European *Drosophila* Genome Project (EDGP) is constructing a cosmid-based physical map of moderately high coverage, anchored to the cytogenetic map by in situ hybridization to polytene chromosomes (Sidén-Kiamos et al., 1990, Kafatos et al., 1991). As part of this project, we are determining short STSs from both ends of cytologically mapped cosmids, aiming to establish semi-random sequence landmarks at an average distance of approx. 50 kb and thus make our physical map independent of the library used for its construction (Madueño et al., 1995). Here we report that STSs of this type are a rich source for detecting previously unknown potential *Drosophila* genes.

2. Experimental

The physical map prepared by the EDGP includes contigs of overlapping cosmid clones as well as cosmids that have not been attached to others according to our strict fingerprint criteria. The subset of mapped clones chosen for STS analysis includes both representatives of contigs and unattached clones that have been assigned to chromosomal sites by in situ hybridization to polytene chromosomes (Sidén-Kiamos et al., 1990; Kafatos et al., 1991); clones with mostly repetitive sequences are excluded from this analysis. We aim at a semi-random distribution of STSs, by including a selection of clones from each contig as well as unattached clones (Madueño et al., 1995). The analysis reported here is based on single-pass, single-strand sequences of 1788 STSs, with a cumulative length of 463 139 bp and an average size of 259 bp/STS, which have been submitted to the EMBL nucleotide library and are also available through FlyBase. Of these, 568 STSs are X-linked (Madueño et al., 1995) while 909 are derived from chromosome 2 and 311 from chromosome 3. Release 47.0 of the EMBL nucleotide library plus daily updates until 31 May 1996 [including the STS and expressed sequence tags (EST) subsets] and release 33.0 of the SWISS-PROT database were searched with the BLASTN and BLASTX programs, respectively (Altschul et al., 1990). BLASTX was also used to search a beta version of TREMBL, an unannotated protein database containing the translation of EMBL nucleotide entries that have not yet been processed into SWISS-PROT (Bairoch and Apweiler, 1996). The criteria used for accepting database hits were scores of ≥ 200 for nucleic acids and ≥ 75 for proteins, in both cases combined with a $P \leq 10^{-4}$. Most of the hits were obtained by protein analysis. When a hit was detected with the SWISS-PROT database but not the EMBL nucleotide library, the STS was further analysed for confirmation of the hit. This analysis included the determination of ORF, the presence of putative protein sequence motifs and comparison of the entire STS and peptide hit sequences using a variety of programs (Devereux et al., 1984). Forty of the 102 'SWISS-PROT but not EMBL' hits were rejected on the basis of the similarities being due only to the presence of simple repeats in the query sequences. Of the remaining 62 hits in this class, 51 were corroborated and 11 were rejected by further analysis, suggesting that the standard score and probability criteria used to screen for potential hits were pragmatically reasonable. Comparable results were obtained for TREMBL hits that did not simultaneously hit a corresponding EMBL nucleotide library entry.

The database searches revealed that 1436 STSs (80.4%) had no significant hits. 112 STSs (6.3%) were derived from 92 already sequenced fruitfly genes by the criterion of sequence identity or near identity (consistent with polymorphisms or errors expected from single-pass sequence analysis). These hits represent additional links of the cosmid physical map to pre-existing *Drosophila* maps. Of the remainder, 107 (6.0%) showed similarities to known *D. melanogaster* repetitive sequences including transposable elements, histone genes and ribosomal RNA genes, and 27 others (1.5%) corresponded to previously known anonymous fruitfly sequences (STSs and P1 clones).

Interestingly, 106 STSs (5.9%) identified 105 novel *Drosophila* putative genes. Of these, 13 STSs hit new members of gene families that are already represented in the fruitfly. All of these were putative genes with localized similarities but extensive divergence from known and sequenced *D. melanogaster* genes. In all but two cases the previously known gene and the cosmid from which the STS was made mapped to different chromosomal sites. The new putative gene encoding a phosphoglycerate kinase homologue was mapped in the cytological proximity of the original gene (23AB and

Table 1
STSs identifying new putatives genes of *D. melanogaster*[a]

STS	Accession no.	Cytol. loc.	Protein/gene
Drosophila melanogaster			
33A10T	Z32066	8F-9A	*kelch* (36E1-3), ring canal BTB domain protein (Q04652)
85A5T	Z32328	22F	*Notch* (3C7), cell-adhesion transmembrane receptor protein (P07207)
169C9S	Z50433	23AB	*Phosphoglycerate kinase* (23A1-2) (Q01604)
164G3S	Z50413	27F–28A	*zinc finger homeodomain 1* (100A1-2) (P28166)
159E11S	Z71115	38B	*Serendipity-α* (99D4-8), plasma membrane associated protein (P07666)
169C4S	Z50431	47EF	*η-Trypsin* (47D-F) (P42279)
103E9T	Z71007	47F	*Serendipity-α* (99D4-8), plasma membrane associated protein (P07666)
73C1S	Z70947	50EF	E239.9-1 mRNA (Q23992)
61D6S	Z70922	51C	*Cytochrome P450-6a2* (42D) (P33270)
90G3T	Z70987	54D	*Suppressor of variegation 205* (29A), chomatin binding protein (P05205)
114B9T	Z50257	56F–57A	*ζ-Trypsin* (47D-F) (P42280)
110D4S	Z50246	58BC	*scabrous* (49D1-3), fibrinogen-like protein (M60065)
12E3T	Z31819	67C	*Serine protease 2* (99D1) (P17205)
Escherichia coli			
42D3T	Z50572	67D	ATP-dependent RNA helicase SrmB (P21507)
174E4S	Z50448	100C	Polyribonucleotide nucleotidyltransferase (P05055)
Sulfolobus solfataricus			
90E6T	Z50693	62B; 48E	Elongation factor 1-α (P35021)
Blastocladiella emersonii			
92F9T	Z50701	31A	cAMP-dependent protein kinase regulatory chain (P31320)
Yarrowia lipolytica			
25C7S	Z32012	19BC	Hyp. protein in alkaline extracellular protease 3′ region (P09379)
Saccharomyces cerevisiae			
170H11T	Z32461	6C	β-Transduction protein TUP1P (Q06440)
47C4T	Z32133	7D	Phenylalanine-tRNA synthetase β chain (P15625)
38C6T	Z32096	14E	Hyp. 108.4 kDa protein in BET1–PAN1 intergenic region (P40559)
38H5T	Z32101	17D; 86E	Hyp. 90.2 kDa zinc finger protein in CCA1–ADK2 intergenic region (P39956)
155F7S	Z31877	18CD	Mitochondrial RNA splicing protein MSR4 (P23500)
156G4T	Z31880	18D	Isocitrate dehydrogenase (P28241)
26A8S	Z32019	22A	DOA1 protein (P36037)
136D12S	Z71068	46B	Put. ATP-dependent RNA helicase Drs1 (P32892)
153H6S	N. S.b	48E	Pre-mRNA splicing factor PRP8 (P33334)
53B6T	Z70907	48E	Pre-mRNA splicing factor PRP8 (P33334)
159A11T	Z71114	49A	Paired amphipathic helix protein Sin3 (P22579)
85C3T	Z70980	55F–56A	Methionyl-tRNA synthetase (P00958)
Schizosaccharomyces pombe			
11G9S	Z31791	10A1.2	ATP-dependent RNA helicase Prh1 (Q03319)
Allium porrum			
192G1S	Z50482	34A	DnaJ homologue 2 (Q03363)
Caenorhabditis elegans			
99C2T	Z71002	23AB	F38B6.6 put. protein, sim. to *E. nidulans* bimA gene product(Q20144)
166H10T	Z50420	23C	C38H2.2. put. protein (Q18515)
143G7T	Z50352	28E	F11A10.4 put. protein (Q19338)
36C2T	Z70879	43F–44A	D2013.9 put. protein (Q09512)
71C4S	N. S.	44A	F45E12.3 put. protein, sim. to LIN-19 (Q20428)
98D7T	Z70998	44F	YK45B9.3 cDNA. (Q09263)
96E3T	Z50707	44F–45A	C44C1.5 put. protein, sim. to *S. cerevisiae* hyp. 51.3 kDa protein in SMY2–RPS101 region (Q18610)
98C9S	Z70995	45D	F35G12.4. put. protein, sim. to β-transducin (Q20059)
64B3S	Z70929	48F	Deoxyribose-phosphate aldolase (Q19264)
16A7S	Z70818	52C; 52D	UNC-89 (Q17362)
145A2S	Z71091	52D	T11B7.4 put. protein (Q17878)
99D3T	N. S.	60A	Hyp. 112.3 kDa protein K02A2.3 in chromosome II (Q20144)
15A8S	Z31888	62B	T08A11.1. put. protein (Q22331)
34H9S	Z70869	67CD	F47A4.5. put. protein (Q20500)
Manduca sexta			
49G4T	Z70901	49A	Diuretic hormone receptor (P35464)
Tenebrio molitor			
87D3T	Z50689	57BC	Trehalase (P32359)
Lucilia cuprina			
47H9T	Z32139	13D	Serine proteinase (Q25232)

Table 1 (*continued*)
STSs identifying new putatives genes of *D. melanogaster*[a]

STS	Accession no.	Cytol. loc.	Protein/gene
Xenopus laevis			
50C11S	N. S.	38E	Defender against cell death 1 (DAD1) (P46967)
19F12S	Z31984	66C	P58 (Q91671)
Petromyzon marinus			
83C11T	Z32319	16EF	Fibrinogen γ-chain (P04115)
Mus musculus			
131F2S	Z31823	2B7–8	Putative embryonic H β 58 protein (P40336)
43C6S	Z32119	11BC	KIZ-1 (P53668)
176C7T	Z31931	13E	TCP-1-containing cytosolic chaperonin ζ-subunit (CCT ζ) (P80317)
181H6T	Z31947	17DE	Cytochrome P450 IIF2 (naphthalene hydroxylase) (P33267)
149G7T	Z50375	23CD	Seryl-tRNA synthetase (P26638)
100F7T	Z71005	28E	Laminin α-1 chain (P19137)
176G11T	Z71149	31A	Protein kinase RCK (Q04859)
186F5S	Z50468	34C	Mitogen-activated protein kinase P38 (P47811)
110D6T	Z31762	34D	ATP-binding cassette transporter 1 protein (P41233)
49G2T	N. S.	38D	Phosphatidylcholine-sterol acyltransferase precursor (P16301)
68G9S	Z50637	47EF	cDNA clone 330330 (W14980)
87D3S	Z50688	57BC	Put. p4-6 protein, sim. to *C. elegans* hyp. protein F10B5.4 (P46686)
83E10S	Z50674	60A	DNA ligase I (P37913)
1H4T	N. S.	61B	Zinc finger protein MFG1 (P16372)
Rattus norvegicus			
30B8S	Z32055	2F	RAB geranylgeranyltransferase α-subunit (Q08602)
176C8S	Z31932	10F	Insulin degrading enzyme (NDR convertase) (P35559)
57H4T	Z50609	34F–35A	GTP-binding protein Rab 14 (P35287)
186F2T	Z50467	55E	RSEC6 (Q62825)
17F2S	Z31944	60E	Membrane-type matrix metalloproteinase (Q10739)
115A8S	Z50260	66CD	Putative Unr protein (P18395)
Oryctolagus cuniculus			
110E6T	Z31764	8D	Low-density lipoprotein receptor (P20063)
39C10S	Z50560	98F	Phosphoserine aminotransferase (P10658)
Sus scrofa			
100B6T	Z31728	9A	Prolyl endopeptidase (P23687)
63D12S	Z70925	35E	Aminopeptidase N (P15145)
34F4T	N. S.	68D	Aldose reductase (P80276)
Bos taurus			
7B3S	Z32298	25A	Farnesyl transferase α-subunit (P29702)
86G11T	Z50687	57F	cGMP-gated cation channel protein (Q03041)
12E9T	Z50304	58CD	Alkaline phosphatase (P09487)
78F12T	Z50652	66A	Leucine aminopeptidase (P00727)
44A3S	Z50573	66D	Cytosol aminopeptidase (P00727)
169A12T	Z50430	96B	Dipeptidyl aminopeptidase IV-like protein (P42659)
Homo sapiens			
79G8T	Z32297	2B	Glutamate (NMDA) receptor ζ-subunit 1, long form precursor (NR1)(P35437)
63B12S	Z32226	2B10–14	Archain (P48444)
84H4T	Z31739	4C	CBP80 mRNA (X80030)
105D2S	Z31739	5C	TAT interactive protein mRNA (U40989)
143C1T	Z31845	13B	Dihydrolipoyl transacylase (P11182)
193C11T	Z32471	15F–16A	KIAA0033 mRNA (Q15055)
29G11T	Z32048	18BC	Partial cDNA sequence of clone c-20d10 (F07258)
167B11S	Z71128	26A	Hepsin (cell surface serine protease) (P05981)
172F5T	Z50477	29A	1α-*N*-Acetylglucosaminidase (U43573)
55A11T	Z70911	29A	DNA-Repair protein complementing XP-G (Xeroderma pigmentosum) cells (P28715)
110H6T	N. S.	30E	Mitochondrial 3-ketoacyl-CoA thiolase (P42765)
53D10S	Z70908	33B	Myosin VIIA (USH1B) (U39226)
69B8T	N. S.	37	Insulin gene enhancer protein ISL-1 (P47894)
185H6T	Z71153	45F	Uroporphyrinogen decarboxylase (P06132)
49G4S	Z70900	49A	Dystrobrevin-δ (U26742)
39C11S	Z79888	52E	Aspartyl β-hydroxylase (Q12797)
68G6S	Z70938	53D	Putative oncoprotein DEK (P35659)
22G12T	Z50522	55CD	High-density lipoprotein binding protein (Q00341)
145E2S	Z50361	60C	NEDD-4 related protein (KIAA0093) (P46934)

Table 1 (*continued*)
STSs identifying new putatives genes of *D. melanogaster*[a]

STS	Accession no.	Cytol. loc.	Protein/gene
18G3S	Z31958	67C	40 kDa peptidyl-prolyl cis-trans isomerase (Cyclophilin-40) (Q08752)
3H2T	N. S.	87C	Nuclear factor NF45 mRNA (U10323)
25H1T	Z50525	96DE	Protein phosphatase PP2A, 130 kDa regulatory subunit (Q06190)
59D7T	Z70916	99F	Prolyl 4-hydroxylase α-subunit (P13674)

[a]For each putative new gene, the EDGP designation of the STS is shown in the first column, followed by its accession number in the EMBL nucleotide library, its cytological location on polytene chromosomes and a brief description of the homologous protein or gene. The STSs are sorted by the organism in which the highest homology score was detected. For convenient reference to the *Drosophila* cytological map, within each organism set the STSs are listed in the order of cytological location. The descriptions of homologies are accompanied by the accession number of the hit (in parentheses). For *D. melanogaster* hits, the name of the known gene and its cytological location (in parentheses) are also shown. Additional information about EDGB STSs can be obtained from FlyBase at the WWW, at the URL http://morgan.harvard.edu.
[b]The STS sequence is not submitted yet to the EMBL nucleotide library.

23A1-2, respectively; Table 1). One gene encoding a trypsin homologue was mapped within the cytological location of the previously identified cluster of trypsin encoding genes (47EF and 47D-F, respectively; Table 1). Fig. 1 shows the conceptual products of these 13 putative genes aligned with their previously known fruitfly cognates, using the BESTFIT program of the GCG Sequence Analysis package (Devereux et al., 1984). The remaining 93 STSs encoded putative proteins showing similarities with the products of 92 genes from other species. The genes on that list included 20 genes identified by double hits (EMBL nucleotide and SWISS-PROT or TREMBL protein databases); 43 genes that were identified by SWISS-PROT but not EMBL searches (a yeast gene homologue was hit by two different STSs, 153H6S and 53B6T); and 22 genes that were identified by TREMBL but not EMBL searches and were corroborated upon closer examination. Furthermore, the list included four highly significant EMBL hits that were not accompanied by relevant entries in either the SWISS-PROT or TREMBL database; the most probable explanation is that there were no conceptual translations for any homologues of these DNAs in the two protein databases at the time of the searches. Finally, one STS (193C11T) was closely similar to a human anonymous cDNA and one (68G9S) to a mouse cDNA encoding peptides of unknown function. Table 1 lists these 106 STSs derived from previously unknown *D. melanogaster* putative genes, together with their presumed function as indicated by the protein hit. They include prokaryotic as well as eukaryotic homologues, with a phylogenetic distribution that is skewed by the high prevalence of sequences from certain organisms.

3. Discussion

It is noteworthy that even in the very well-studied genome of *D. melanogaster*, 5.9% of the STSs hit novel but recognizable putative genes. Thus, at least in compact genomes like that of the fruitfly, a byproduct of STS-based mapping projects may be the discovery of new genes by homology searches. The number of genes hit is significantly higher than expected. The 463 kb of DNA encompassed in the STSs is only 0.39% of the euchromatic genome, and yet it overlaps with 92 out of the 1700 sequenced *Drosophila* genes that are in the databases (5.4%). Similarly, besides the 105 'hits' listed in Table 1, we probably encountered but did not recognize additional genes: approximately one-third of eukaryotic ORFs correspond to pioneer (orphan) proteins (Dujon, 1996), and pioneer *Drosophila* proteins would not have been recognized as database hits. Thus, if our DNA sequences were a random sample of a euchromatic genome 120 Mb in length, they would imply the existence of at least 43 000 *Drosophila* genes. Since our sequences are short we do not know whether any of the hits correspond to pseudogenes, but sequencing studies to date have revealed very few pseudogenes in the fruitfly. Only one of the new sequences (C in Fig. 1) shows internal stops suggestive of a pseudogene (or of the presence of a mini-intron). However, we do not believe that our sequences are a random sample. For example, all of our STSs are adjacent to genomic *Sau* 3A sites, and are derived from approx. 40 kb fragments enriched in unique DNA; they have an average composition of 44% G+C, as opposed to 43% for the fruitfly genome as a whole. Our current interpretation is that, for various systematic reasons, our method of obtaining STSs tends to sample gene-rich regions preferentially. Additional studies are necessary to estimate more accurately the number of currently uncharacterized fruitfly genes, but the present study represents an intriguing sampling of this unknown genetic universe.

The 13 hits that show homology but not identity with known *Drosophila* genes are a diverse lot, emphasizing that a wide variety of *Drosophila* genes are members of gene families. Their distinctness is established both by sequence comparisons and by their cytogenetic locations. They include, for example, two new homologues of the plasma membrane associated protein gene *Sry-α*, at

Fig. 1. Protein alignments between conceptually translated STSs (bottom line of each alignment) and TREMBL or SWISS-PROT hits representing known *D. melanogaster* genes (top line). The sequences are presented in the order shown in Table 1. The numbers flanking the top line indicate the positions of the amino acid residues in the known protein; numbers flanking the bottom line indicate nucleotide positions in the STS. Only two sequences (C and L) had significant hits to the nucleotide library as well as to the protein database; the former has stop codons (asterisks) in all three reading frames, and may represent a pseudogene or an alignment corrupted by sequencing errors and a mini-intron. The per cent similarity and identity are indicated, dashes are inserted for best alignment, and only the regions of highest similarity are shown in the alignments.

different sites of the second chromosome. Three *Sry* genes (α, β and δ) were already known to be clustered near the tip of chromosome *3*. The newly discovered putative genes also encode a novel member of the Notch family of receptors, a new zinc finger protein and two new trypsin homologues, one within a cluster of nine previously known trypsin genes and one at a separate site.

The 91 potential genes that show homologies to sequences of other organisms are also diverse. The largest classes encode enzymes of intermediate metabolism (23), nucleic acid synthesis or modification enzymes including transcription factors, RNA helicases, splicing factors, tRNA synthetases and DNA modification enzymes (20), enzymes of potential signal tranduction pathways (7), and membrane proteins (5). It is worth mentioning here that a 'reverse' study was recently published, in which potential *Drosophila* genes were searched for using BLASTX searches, showing that out of approximately 255 000 human ESTs used to probe the databases, 66 were found to be very similar to a diverse set of already sequenced fruitfly genes (Banfi et al., 1996).

The newly encountered *Drosophila* putative genes amount to as much as 6.1% of the fruitfly genes that exist in the sequence databases. Therefore, they will be of significant interest for *Drosophila* researchers. In recent years, it has become obvious that the fruitfly is an excellent model system for higher eukaryotes. For example, it shares most types of regulatory genes which are represented with a higher multiplicity in vertebrates, and which can be analysed conveniently in the fruitfly because of its genetic tractability. Thus, the present findings should also be of interest to non-*Drosophila* workers.

Acknowledgement

This work was supported by grants from the European Union (SCIENCE and Human Capital and Mobility programmes) and the Medical Research Council, as well as institutional funds from the Fundacion Ramon Areces to the CBMSO, and from the Greek General Secretariat for Research and Technology to IMBB.

References

Altschul, S.F., Gisch, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic Local Alignment Search Tool. J. Mol. Biol. 215, 403–410.

Bairoch, A., Apweiler, R., 1996. The SWISS-PROT protein sequence databank and its new supplement TREMBL. Nucleic Acids Res. 24, 21–25.

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

Devereux, J., Haeberli, P., Smithies, O., 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12, 387–395.

Dujon, B., 1996. The yeast genome project: what did we learn? Trends Genet. 12, 263–270.

FlyBase, 1996. The *Drosophila* database. Nucleic Acids Res. 24, 53–56.

Garcia-Bellido, A., Ripoll, P., 1978. The number of genes in *Drosophila melanogaster*. Nature 273, 399–400.

Kafatos, F.C., Louis, C., Savakis, C., Glover, D.M., Ashburner, M., Link, A., Sidén-Kiamos, I., Saunders, R.D.C., 1991. Integrated maps of the *Drosophila* genome: progress and prospects. Trends Genet. 7, 155–161.

Lefevre, G., Watkins, W., 1986. The question of the total gene number in *Drosophila melanogaster*. Genetics 113, 869–895.

Madueño, E., Papagiannakis, G., Rimington, G., Saunders, R.D.C., Savakis, C., Sidén-Kiamos, I., Skavdis, G., Spanos, L., Trennear, J., Adams, P., Ashburner, M., Bolshakov, V.N., Glover, D.M., Kafatos, F.C., Louis, C., Majerus, T., Modolell, J., 1995. A complete physical map of the X chromosome of *Drosophila melanogaster* assembled by cosmid contigs and sequenced tagged sites. Genetics 139, 1631–1647.

Miklos, G.L.G., Rubin, G.M., 1996. The role of the genome project in determining gene function: insights from model organisms. Cell 86, 521–529.

Mortimer, R.K., Schild, D., Contopoulou, C.R., Kans, J.A., 1989. Genetic map of *Saccharomyces cerevisiae*, Edition 10. Yeast 5, 321–403.

Olson, M., Hood, L., Cantor, C., Botstein, D., 1989. A common language for physical mapping of the human genome. Science 245, 1434–1435.

Sidén-Kiamos, I., Saunders, R.D.C., Spanos, L., Majerus, T., Trenear, J., Savakis, C., Louis, C., Glover, D.M., Ashburner, M., Kafatos, F.C., 1990. Towards a physical map of the *Drosophila melanogaster* genome: mapping of cosmid clones within defined genomic divisions. Nucleic Acids Res. 18, 6261–6270.