Classical Oncogenes and Tumor Suppressor Genes: A Comparative Genomics Perspective

Oxana K. Pickeral*[†], Jonathan Z. Li*, Ian Barrow*[‡], Mark S. Boguski*[†], Wojciech Makałowski* and Jiong Zhang*[†]

*Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892; [†]Department of Molecular Biology and Genetics, The Johns Hopkins University School of Medicine, Baltimore, MD; [‡]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892

Abstract

We have curated a reference set of cancer-related genes and reanalyzed their sequences in the light of molecular information and resources that have become available since they were first cloned. Homology studies were carried out for human oncogenes and tumor suppressors, compared with the complete proteome of the nematode, *Caenorhabditis elegans*, and partial proteomes of mouse and rat and the fruit fly, *Drosophila melanogaster*. Our results demonstrate that simple, semi-automated bioinformatics approaches to identifying putative functionally equivalent gene products in different organisms may often be misleading. An electronic supplement to this article[1] provides an integrated view of our comparative genomics analysis as well as mapping data, physical cDNA resources and links to published literature and reviews, thus creating a "window" into the genomes of humans and other organisms for cancer biology. *Neoplasia* (2000) **2,** 280–286.

Introduction

Seventeen years ago, bioinformatics and cancer research intersected in a way that profoundly altered biologist's view of computers and databases as biomedical research tools. A long-forgotten chapter in the history of this field is the computer-based discovery that the viral oncogene *sis* was "homologous" (80% identical) to human platelet-derived growth factor [1,2]. This singular event provided a dramatic demonstration that great advances in understanding the pathophysiology of disease could be made by searching and aligning sequence data. Since that time, this process of discovery has been successfully repeated countless times, often aided by cross-species sequence comparisons (e.g., Refs. [3,4]). The Human Genome Project and associated developments have engendered more "global" views of biology where either entire genomes, or large functional components thereof, may be analyzed in toto rather than one gene at a time.

In the present work, we provide an integrated view of classical cancer genes by assembling information resources for, and performing new analyses of 101 oncogene and tumor suppressor gene products. The new analyses include the "comparative genomics" of human cancer-related genes with their homologs in four important model organisms: mouse and rat, the nematode, *Caenorhabditis elegans*, whose genome was completed in late 1998 [5] and the fruit fly, *Drosophila melanogaster* for which a substantial database of protein sequences was already available before anticipated publication of the complete sequence. The results from these cross-species comparisons show that simple quantitative comparisons, i.e., BLAST searches, are not a reliable guide for identifying functionally equivalent gene products but rather just the first step in assessing whether or not a particular organism is the most appropriate model for specific studies of cancer biology.

With the emergence of complete genomes and/or comprehensive gene catalogs for a variety of organisms, molecular sequence data have become the common currency of biomedical research. The sheer quantity and complexity of these data, however, are daunting: in GenBank, there are currently about 6 billion bases in approximately 5.7 million sequence records representing more than 50,000 different biological species. Furthermore, these data are often complicated by redundancy, and uneven or outdated annotation. In the electronic supplement to this work, we have built a "window" into the human genome through which one can view a non-redundant and consistent picture of molecular genetic properties of 101 genes involved in neoplasia. The reference sequences contained in our collection have been used in the design and construction of the "lymphochip" gene expression array (L. Staudt, personal communication) that has recently been used to discover distinct types of diffuse large B-cell lymphomas [6].

Materials and Methods

Oncogenes and tumor suppressor genes were selected for analysis using a published collection [7] of these genes as a guide. Many genes may be implicated in various neoplastic phenomena. However, we used very strict criteria for inclusion of sequences in our study. Only those genes that have been shown to be tumorigenic or specifically expressed (in activated form) in at least one type of tumor cell, or that display either specific mutations or complete loss of expression in at least some human cancers, were included. These rigorous criteria prevented the project from becoming so open-ended that virtually all genes having anything to do with cell cycle control, signal transduction, or indeed any pleiotropic effect of the transformed phenotype would have to be considered. A somewhat less stringent, but more inclusive, approach was taken by the CGAP project (www.ncbi.nlm.nih.gov/ncicgap/) subsequent to the appearance of our web site.

D. melanogaster and *C. elegans* protein databases used for BLAST searches were created using the **formatdb** program (ftp://ncbi.nlm.nih.gov/blast/server/README). The *D. melanogaster* protein file (6592 sequences), and two *C. elegans* protein files— NematodePep 17 (19,126 sequences) and "October_Proteins.pep" (19,099 sequences) were obtained with the assistance of M. Ashburner and R. Durbin, respectively, and were downloaded on 21 April 1999 from the following sites:

ftp.ebi.ac.uk/pub/databases/edgp/sequence_sets/nuclear_cds_set.embl.v2.0.Z
(note that this set is no longer available and has been updated to nuclear_cds_set.embl.v2.3.Z)

ftp.sanger.ac.uk/pub/databases/nematodepep/nematodepep16

ftp.sanger.ac.uk/pub/C.elegans_sequences/SCIENCE98/October_Proteins.pep.gz.

It is important to note that the latter set was the one used for most publications in the nematode genome issue of *Science*, 11 December 1998.

Similarity searches were performed by BLASTP program [8] (also, http://www.ncbi.nlm.nih.gov/BLAST/) with default parameters. One hundred and four queries corresponded to the 101 genes in our data set. (When different amino acid records were found for the same gene due to alternative splicing, both sequences were used in blast searches.) The best match was selected based on its local alignment (HSP) score plus an alignment length criterion applied to the matching query and database protein sequences. When several high-scoring candidates were present, the "best" were selected based on the e-value, the percentage identity of the HSP, the relative positions of HSPs within the query and the subject proteins, the presence of multiple high-scoring HSPs aligning to the same domain in the query protein, knowledge of domain function, global alignment scores, and multiple alignment results.

Global alignments were computed using the **align** program [9] and the BLOSUM50 scoring matrix with default gap penalties [10]. For multiple sequence alignments, the **clustalify** utility from the SEALS package [11] was used (command line parameters were: clustalify -mode = align -multiple_type = protein -multiple_endgaps -save [file names]).

Results and Discussion

Comparative Genomics

Results of the cross-species analyses are summarized in Table 1. Mouse or rat orthologs were retrieved from the HOVERGEN database [12], together with their corresponding percentage identities derived from global alignments with their human counterparts. Fly and nematode homologs were selected independently, using the BLASTP program to search a database of *D. melanogaster* proteins and a database of *C. elegans* proteins, respectively, as described in Materials and Methods section. BLAST parameters included the e-value cutoff of e-05, and the best "candidate ortholog" was selected from the top five matches, based on the score of the match and the differences between the lengths of the query and subject proteins. Once the best match was selected, protein sequence identity in a global alignment between the human query and the *D. melanogaster* or *C. elegans* match was calculated, if the length of the matching protein was within 20% of the query protein length.

An additional test of putative orthology applied in this study was reciprocal BLAST analysis. For the best matches selected in fly and nematode, BLASTP searches using these sequences as queries were performed against a database if all sequences classified as "vertebrata" in GenBank as of 17 August 1999. The high scores obtained in the initial BLASTP search (using the human protein query) were compared with the high scores from the reciprocal BLASTP search. If these scores differed by more than 20%, the corresponding fly or nematode match was deemed unlikely to be the ortholog of the initial human query protein. Matched sequences that did not satisfy this reciprocal BLAST criterion are identified by an asterisk preceding the GenBank accession number in Table 1.

The values for protein percentage identities in pairwise, cross-species alignments are provided in Table 1. Only matches that passed both the length comparison and the reciprocal BLAST criteria are included in the following summary statistics. Percentage identities for human-rodent alignments ($n = 90$) ranged from 57.3% to 99.5%, with a mean of 89.9% (SD 8.7) and a median of 92.4%. This mean value is not significantly different from the mean values (85.4% SD 12.6 and 88.0%, SD 11.8) previously reported for much larger human-mouse ($n = 1196$) [13] and human-rat ($n = 1212$) [14] data sets, respectively.

Percentage identities for human-fly alignments ($n = 40$) ranged from 19.6% to 78.8%, with a mean of 42.2% (SD 14.5) and a median of 41.2%. Percentage identities for

Table 1. Homologs of the Human Oncogenes and Tumor Suppressors in Rodents, Flies and Nematodes.

Gene symbol	Hum_acc	Mouse_acc	Prot_id	Fly_acc	Prot_id	Nem_acc	Prot_id
HRAS	J00277	Z50013	99.5	M16429	76.4	ZK792.6	74.6
KRAS2	M54968			M16429	78.8	ZK792.6	77.1
NRAS	X02751	M12124	98.4	M16429	75.7	ZK792.6	74.1
EGFR/ERBB-1	X00588			AF109077	35.2	ZK1067.1	27.7
ERBB2/HER2/NEU	M11730	X03362#	87.3	AF109080	32.5	*ZK1067.1	27.1
ERBB3/HER3	M34309	U29339#	90.4	AF109079	31.6	ZK1067.1	25.6
ERBB4/HER4	L07868	AF041838	96.6	AF109077	34.3	ZK1067.1	26.7
RAF1	X03484	M15427	98.3	X07181	44.2	Y73B6A.A	
E2F1	M96577	L21973	86.2	X78421		Y48C3A.T	20.4
GTBP/MSH6	U28946	U42190	86.1	*U17893		Y47G6A_242.C	
CRK	D10656	S72408	98.7	AF112976	42.9	Y41D4A_3457.B	
MLH1	U07418	U80054#	86.9	AF068257	46.0	T28A8.7	33.5
JUN	J04111	J04115	97.3	X54144	31.1	T24H10.2	
JUNB	X51345	U20735	93.6	X54144	30.5	T24H10.2	
JUND	X56681	J05205	95.1	X54144	32.1	T24H10.2	
DCC	X76132	X85788	96.5	U71001	32.2	T19B4.7	26.4
TAL1	M61108	M59764	93.6	AL024485	27.1	T15H9.3	19.0
ERG	M17254	S66169*	98.0	*X68259	26.6	T08H4.3	33.1
FLI1/ERGB2	X67001	X59421	85.6	*X68259	24.3	T08H4.3	33.2
ETS1	J04101	X53953	97.3	X69166		*T08H4.3	23.3
ETS2	J04102	J04103	92.1	*X68259	26.8	*T08H4.3	24.7
CDKN1B/KIP1	U10906	U09968	88.3			T05A6.2	
CDKN1C/KIP2	U22398	U22399	61.4			T05A6.2	20.2
ABL1	X16416	J02995	82.2	M19692		M79.1	34.5
CBL	X57110	X57111	93.1	AJ223175		M02A10.3	
APC	M74088	M88127	90.6	U77947	24.9	K04G2.8B	
MSH2	U03911	X81143	92.4	U17893	41.5	H26D21.2	28.9
MSH3	U61981	M80360	81.5	*U17893	23.1	*H26D21.2	
FGR(SRC2)	M19722	X16440	86.3	D42125	53.7	F49B2.5	46.2
FYN	M14333	U35365#	99.3	D42125	55.0	F49B2.5	48.5
HCK	M16591	J03023	90.1	D42125	52.3	F49B2.5	45.2
LCK	M36881	X03533	96.4	D42125	49.5	F49B2.5	43.4
LYN	M16038	M57696	95.9	D42125	51.0	F49B2.5	43.9
SRC	AF077754	M17031	98.9	D42125	54.0	F49B2.5	47.9
YE1	M15990	X67677	96.3	D42125	53.0	F49B2.5	48.3
ROS1	M34353	X81650	80.5	M34545	23.3	*F49B2.5	
MAX	X66867	M63903	98.1	U77369	41.6	F46G10.6	
MYC	J00120	X00195	91.6	U77370		*F46G10.6	
PIM1	M24779	M13945	93.9	*AL031765		F45H7.4	36.0
CDK4	U37022	L01640	94.7	X99510	43.7	F18H3.5B	36.3
MET	J02958	Y00671	89.6	*U18351		F11E6.8	
BCR	Y00661			*AL031884		C38D4.5	
ELK1	M25269	X87257	85.7	*M88475		C37F5.1	27.3
ELK3	Z36715	Z32815	91.4	*M20408		*C37F5.1	28.6
BRCA1	U14680	U31625	57.3	*AJ001514		C36A4.8	
VAV1	X16316	X64361	93.1	*L12446		C35B8.2	27.6
AKT1	M63167	X65687	98.1	Z26242	56.8	C12D8.10B	52.7
AKT2	M95936	U22445	98.1	Z26242	57.2	C12D8.10B	51.2
BCL3	M31732	AF067774	82.3	L03367	19.6	C04F12.3	
NF1	M89914	L10370*	98.5	L26500	54.0	*ZK899.8D	
PTCH/PTC	U43148	U46155	96.1	X17558	31.6	*ZK6751	25.2
CCND1	M64349	S78355	93.2	U41808		*Y38F1A5	
CCND2	M90813	M83749	92.4	U41808		*Y38F1A5	
CCND3	M92287	U43844	94.9	U41808		*Y38F1A.5	
WNT1	X03072	M11943	98.9	M17230		*W01B6.1	38.3
WNT2	X07876			X64735	38.9	*W01B6.1	40.4

Table 1. (continued)

Gene symbol	Hum_acc	Mouse_acc	Prot_id	Fly_acc	Prot_id	Nem_acc	Prot_id
THRA	M24898	M25804#	94.1	*X51548		*T01B10.4	23.6
MADH4/DPC4	U44378	U79748	99.2	AF019753		*R12B2.1	31.4
CDH1/E-CAD	Z13009	X06115	81.7	*AB002397		*R10F2.1	
FER	J03358	U76762	92.8	X52844	36.4	*M79.1	
FES/FPS	X06292	X12616	90.0	*X52844	36.7	*M79.1	
EPHA1	M18391	U18084*	80.3	*AF146648	33.9	*M03A1.1	25.5
MCC	M62397					*K12F2.1	
PMS1	U13695			*AF068271	23.7	*H12C20.2A	23.0
PMS2	U13696	U28724	74.7	AF068271	40.9	*H12C20.2A	33.6
CSF1R	X03663	X06368	74.6	*X74031	26.1	*F58A3.2	23.1
KIT	X06182	Y00864	92.7	*X74031	25.6	*F58A3.2	23.5
RET	M57464	X67812	85.8	D16401		*F58A3.2	23.5
WT1	X51630	M55512	96.7	*U42402		*F56F11.3	21.7
MOS	J00119	J00372	74.7	*K01042	21.1	*F33E2.2	
MYB	U22376			X05939	31.4	*F32H2.1B	
TGFBR2	D50683	D32072	91.9	*L22176	29.9	*F29C4.1	23.3
CDKN2B/INK4B	AF004819			*AF132196		*D2021.8	
RB1	M15400	M26391	91.1	*AL031583	22.9	*C32F10.2	20.3
MCF2/DBL	X12556			*D86546		*C14A11.3	
TIAM1	U16296	U05245	94.8	D86546		*C11D9.1	
FGF3/INT2	X14445	Y00848	82.4	*U82273		*C05D11.4	
FGF4/HSTF1	J02986	X14849	81.2	*U82273		*C05D11.4	
FGF6/HST2	X63454	X51552	93.4	*U82273		*C05D11.4	
NF2	L11353	L28176	98.2	U49724	46.0	*C01G8.5A	41.6
NTRK3/TRKC	U05012	L14445#	97.1	AF037164	26.7	*C01G6.8	22.9
TRK	M23102	M85214#	86.5	*AF037164	28.1	*C01G6.8	22.7
NFKB2	X61498			AF053614	23.8	*B0350.2B	
CDKN2A/INK4A	L27211	AF059567	85.3	*AF132196			
REL	X75042	X60271	75.6	M23702	29.5		
MAS1	M13150	X67735	88.9	*M77168			
MYCN	Y00664	X03919	85.4	*U77369			
MYCL1	M19720	X13944	90.4	U77370			
BCL2	M14745	L31532	89.0				
BRCA2	U43746	U65594	59.1				
CDKN1A/WAF1	U03106	U09507	78.2				
FOS	V01512	V00727	93.7				
FOSB	L49169	AF093624	95.6				
FOSL1/FRA1	X16707	AF017128	90.0				
FOSL2/FRA2	X16706	X83971	95.1				
PDGFB	M12783	M84453	99.2				
SKI	X15218	U14173*	92.8				
THRB	X0470	S62756	95.8				
TP53	X54156	X00741	76.8				
VHL	AF010238	U12570	84.5				

NOTE. Included for each human gene in this set are: its official gene symbol (column 1), GenBank accession no. (column 2), rodent (mouse or rat) GenBank accession no (column 3), protein percentage identity between the human and rodent proteins (column 4) *D. melanogaster* GenBank accession no. (column 5), protein percentage identity between the human and fly proteins (column 6), *C. elegans* nematodepep identification no. (column 7), protein percentage identity between the human and nematode proteins (column 8). Percentage identity is only reported if the protein length for the other organism is within 20% of the human query length. An asterisk preceding the GenBank no. in columns 5 and 7 denotes that the match did not meet the "reciprocal BLAST" criterion (see text). The framed boxes correspond to the 19 groups that share the same matches in *D. melanogaster* or *C. elegans* (see text).

human-nematode alignments ($n = 28$) ranged from 19.0% to 77.1%, with a mean of 39.6% (SD 16.0) and a median of 35.3%. The mean value (39.6%) for cancer-related genes shared by humans and nematodes is somewhat lower than the mean value (49.1%, SD 17.1) previously reported for a much larger ($n = 819$) set of human-nematode orthologs

[15]. However, the magnitude of the variances indicates that these mean values are not significantly different.

It appears that the mean value (42.2%) for protein conservation between human cancer proteins and their putative fly orthologs is somewhat higher than the degree of sequence conservation (39.6%) for human-nematode

cancer gene products. However, the large variances show that these values are not significantly different. A more telling fact distinguishing flies from nematodes in their relationship to humans is that a larger number ($n = 40$) of putative human-*D. melanogaster* orthologs were found than human-*C. elegans* orthologs ($n = 28$) even though the latter proteome is essentially complete and the *D. melanogaster* data set represented only about 20% of the complete proteome at the time of our analysis.

One of the most striking examples of differences between the best matches to a human query in *D. melanogaster* and *C. elegans* was for the NF1 gene product, neurofibromin. NF1 is a tumor suppressor gene mutated in neurofibromatosis (OMIM [Online Mendelian Inheritance in Man] number 162200), an autosomal dominant disorder characterized by café-au-lait spots and fibromatous tumours of the skin. NF1 homologs, IRA1 and IRA2, are known in yeast [16] and resemble human and *D. melanogaster* NF1 more closely than the best-scoring match from the complete nematode proteome (Table 1).

The best nematode candidate for neurofibromin homolog is a protein annotated as "similar to GTPase-activating protein" (GenBank protein id 3947665). Alignment studies (not shown) indicate that only the central region of this nematode protein aligns with human neurofibromin and furthermore, the nematode protein is only half of the size of both the human and fly NF1 gene products that are nearly identical in size. Interestingly, reciprocal BLASTP analysis shows that there is another human protein that is more similar to nematode GTPase-activating protein, namely ras-

GAP-like protein (gi 105589). Phylogenetic analysis of selected neurofibromin homologs (data not shown) suggests that the ras-GAP-like protein is the ortholog of the nematode protein (id 3947665) and that an ortholog of the NF1 gene is entirely missing from the nematode genome. This finding excludes *C. elegans* as a model organism for study of neurofibromin biology.

Interestingly, in several cases, multiple human genes produced the same "best match" in *D. melanogaster* or *C. elegans*. Based on this, 19 groups of genes (at least two in a group) were determined. The largest group consists of eight human proteins, that include SRC, FYN, YES1, LYN, HCK, FGR (SRC2), LCK, and ROS1 (Table 1). A multiple alignment of all protein sequences in this cluster (eight human proteins, eight rodent proteins, two *D. melanogaster* proteins, and one protein from *C. elegans*) shows that ROS1 and its homologs differ significantly from the rest of the group. The same *C. elegans* match, F49B2.5, was the "candidate ortholog" for all human queries in this group. The same *D. melanogaster* match, D42125, was the "candidate ortholog" for seven of the human queries (ROS1 produced a different best match in *D. melanogaster*, namely M34545). Most notably, F49B2.5 did not satisfy the "reciprocal blast criterion" for ROS1, a finding that decreases the likelihood of this match being the real *C. elegans* ortholog of the human ROS1. The remaining homologs in this group satisfy both the reciprocal BLAST criterion and the length criterion. All of these computed findings are entirely consistent with experimental evidence that ROS1 is not "functionally orthologous" with the rest of the SRC cluster genes.
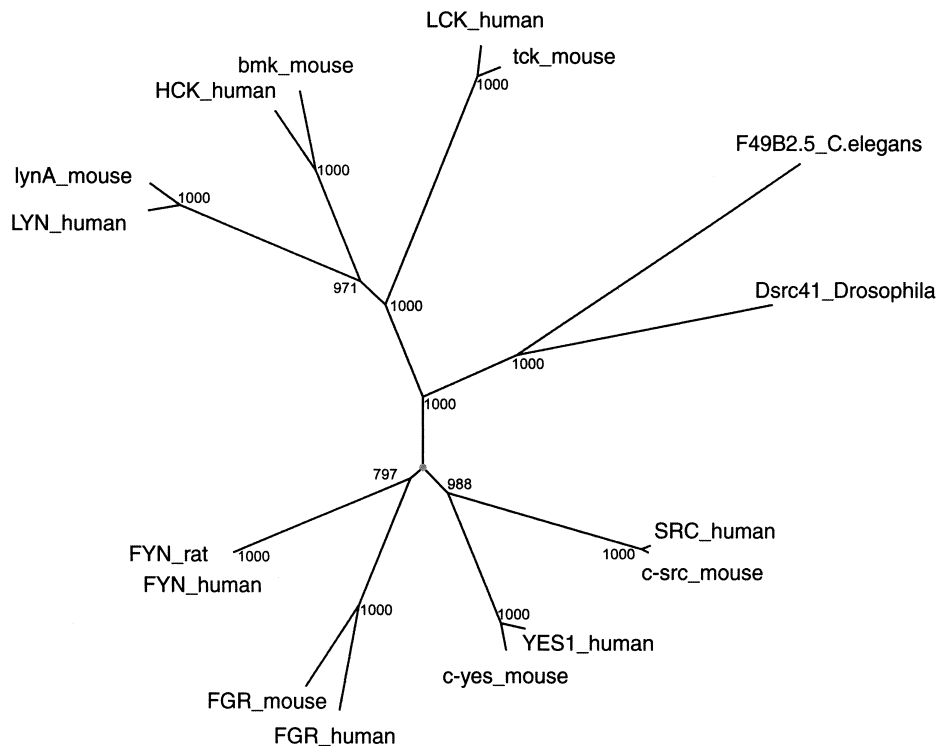

Figure 1. *Neighbor joining tree for the proteins that share fly and nematode matches with the human SRC family of genes. Seven human, seven rodent, one D. melanogaster and one C. elegans gene products belong to this "cluster." Numeric values at branch points indicate the bootstrap values for 1000 tree replications.*
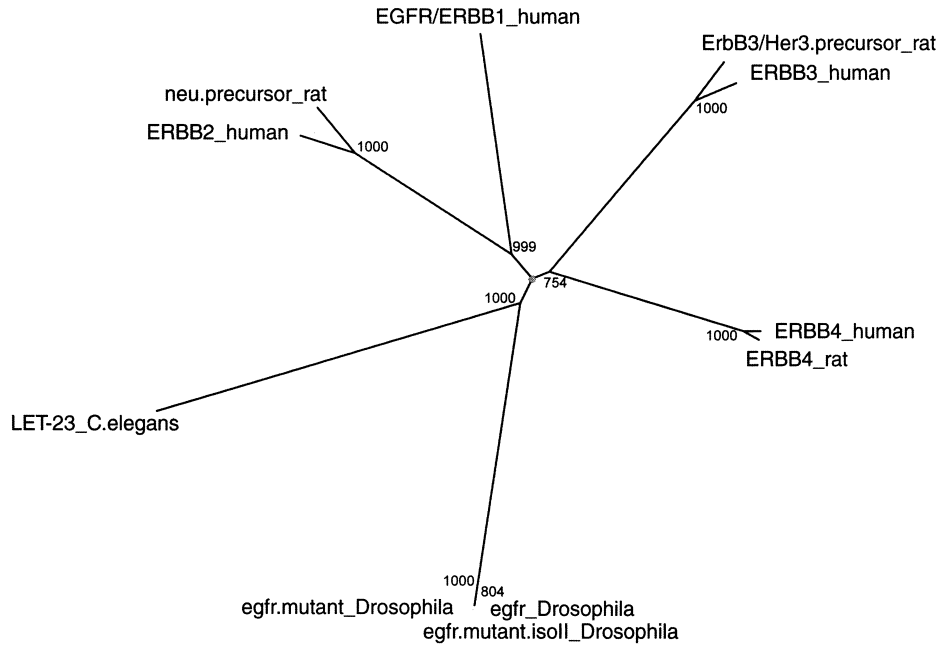
Figure 2. *Neighbor joining tree for the proteins that share fly and nematode matches with the human EGFR/ERBB-1 genes. Numeric values at branch points indicate the bootstrap values for 1000 tree replications.*

Figure 1 shows a dendrogram of the seven SRC family human genes, their rodent orthologs, plus one *D. melanogaster* and one *C. elegans* homolog. This neighbor joining tree was calculated using the CLUSTAL_X program [17]. The bootstrap values were 1000 on all human-rodent nodes, and at least 970 on other nodes with the exception of the FYN genes where the bootstrap value was 797. This tree shows with high confidence that several duplications of the ancestral gene for this family occurred following the divergence of Nematodes and Arthropodes, but before the mammalian radiation.

Multiple alignments of the human proteins, and all available corresponding best matches from rodent, fly, and nematode were produced using CLUSTALW program [17] using default parameters (complete results are available in the electronic supplement from the "Comparative Genomics Table"). If a gene was found to belong to one of the 19 groups, a single multiple alignment was produced for all protein sequences in the group. For example, the ERBB cluster consists of 11 genes, four from human (EGFR/ERBB1, ERBB2, ERBB3, ERBB4), three rat genes, three *D. melanogaster* genes, and one gene from *C. elegans* (Figure 2). The three rodent genes appear to be the orthologs of ERBB2-4, respectively, whereas all *D. melanogaster* genes and the Nematode match are orthologous to all four human ERBB genes. The presence of multiple matches in *D. melanogaster* and a single match in *C. elegans* is due to the presence of multiple sequenced alleles in the fly sequence database. In this case, the three fly matches include epidermal growth factor receptor, mutant epidermal growth factor receptor, and mutant epidermal growth factor receptor isoform ii. Thus, effectively, there is only one candidate ortholog each in *D. melanogaster* and *C. elegans* corresponding to the four ERBB genes in human. Thus, this example illustrates that other cases of apparent over-representation of *D. melanogaster* matches may be explained by a larger number of well-studied alleles and the existence of large mutant collections.

Despite the power and scope of computational, comparative genomics methods to infer or predict gene function, these methods must be carefully applied and their results considered in the broader context of experimental evidence that often includes or implicates pathways of interacting gene products. Indeed, organizing large-scale sequence analysis around a coherent biological subject or system, as we have done in the present work, provides a more meaningful framework in which to evaluate the results. These considerations are becoming critically important as we struggle to provide accurate annotation for the rapidly emerging, complete genome sequences of human and other organisms and to use this information to plan and direct experiments that will take maximal advantage of "model organisms" for gaining insights into human biology and disease.

Acknowledgements

Appendix

Guide to the electronic supplement at www.ncbi.nlm.nih.gov/CBBresearch/Boguski/Neoplasia_Supplement/

The cancer gene set is arranged as two tables, a "Gene Information Table" and a "Comparative Genomics

Table" that may alternately be selected by a pull-down menu on the "Gene List" page. Both contain extension hypertext links to more detailed information. The Gene Information Table begins with the official HUGO (Human Genome Organization) gene symbol and ends with the common name of the gene or gene product. Columns 2 and 3 contain OMIM (Online Mendelian Inheritance in Man) record numbers and GenBank accession numbers, respectively. The link to OMIM provides access to a textual knowledge base containing expert reviews of the literature. The link to GenBank provides a reference sequence for the mRNA (or gene) and usually represents the most complete ("full-length") sequence available, although this is not necessarily the first published report of the sequence. Column 4 shows the length (in kilobases) of the mRNA for ease of comparison with the size of the longest cDNA/EST clone available from public sources (columns 6 and 7). The EST link is provided to the clone with the longest cDNA insert, as extracted from the dbEST [18] records.

Column 5 includes a *LocusLink* identifier that, for each gene, points to a complete list of all existing mRNA, EST and STS sequences and associated annotation. LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink) is a new resource at the National Center for Biotechnology Information and contains descriptive information about genetic loci [19]. It presents information on official nomenclature, gene and gene product name aliases, sequence accession numbers, phenotypes, Enzyme Commission Nomenclature (EC) numbers, UniGene [20,21] clusters, relevant web sites and other information.

Any of the columns in the Gene Information Table may be included or excluded from the display using check boxes following the "Select Columns:" option. The table can also be text-searched by gene symbols or product names and the corresponding line in the table is highlighted when a match occurs.

The "Comparative Genomics Table" is similar to Table 1 in the printed article but also includes hypertext links to the multiple sequence alignments, as described in the text, including those used to compute the dendrograms in Figures 1 and 2.

Most of the genes (80%) in our collection have been placed on the integrated radiation hybrid map of the human genome [22] and links to GeneMap'99 (http://www.ncbi.nlm.nih.gov/genemap/)are provided. For the 20 genes not present on this map, a cytogenetic location is given, based on data in the corresponding OMIM records. An overview of the map locations of cancer genes on each human chromosome, or the genome as a whole, is provided through a selection box just above and to the left of the online table or a menu selection on the left side bar of the home page.

References

[1] Doolittle RF, Hunkapiller MW, Hood LE, Devare SG, Robbins KC, Aaronson SA, and Antoniades HN (1983). Simian sarcoma virus oncogene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science* **221**, 275–277.

[2] Waterfield MD, Scrace GT, Whittle N, Stroobant P, Johnsson A, Wasteson A, Westermark B, Heldin CH, Huang JS, and Deuel TF (1983). Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus. *Nature* **304**, 35–39.

[3] Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, and Kolodner R (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer [published erratum appears in Cell 1994 Apr 8;77(1):167]. *Cell* **75**, 1027–1038.

[4] Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Peltomaki P, Sistonen P, Aaltonen LA, Nystrom-Lahti M, de la Chapelle A, Kinzler KW, Vogelstein B, et al. (1993). Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* **75**, 1215–1225.

[5] Consortium CES (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018.

[6] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, and Staudt LM (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling [see comments]. *Nature* **403**, 503–511.

[7] Hesketh R (1997). In *The Oncogene and Tumour Suppressor Gene FactsBook* (2nd ed). Academic Press, San Diego.

[8] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

[9] Myers EW, and Miller W (1989). Approximate matching of regular expressions. *Bull Math Biol* **51**, 5–37.

[10] Henikoff S, and Henikoff JG (1993). Performance evaluation of amino acid substitution matrices. *Proteins* **17**, 49–61.

[11] Walker DR, and Koonin EV (1997). SEALS: a system for easy analysis of lots of sequences. *ISMB*, **5**, 333–339.

[12] Duret L, Mouchiroud D, and Gouy M (1994). HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res* **22**, 2360–2365.

[13] Makalowski W, Zhang J, and Boguski MS (1996). Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res* **6**, 846–857.

[14] Makalowski W, and Boguski MS (1998). Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci USA* **95**, 9407–9412.

[15] Wheelan SJ, Boguski MS, Duret L, and Makalowski W (1999). Human and nematode orthologs— lessons from the analysis of 1800 human genes and the proteome of *Caenorhabditis elegans*. *Gene* **238**, 163–170.

[16] Ballester R, Marchuk D, Boguski M, Saulino A, Letcher R, Wigler M, and Collins F (1990). The NF1 locus encodes a protein functionally related to mammalian GAP and yeast IRA proteins. *Cell* **63**, 851–859.

[17] Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, and Higgins DG (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882.

[18] Boguski MS, Lowe TM, and Tolstoshev CM (1993). dbEST— database for "expressed sequence tags". *Nat Genet* **4**, 332–333.

[19] Maglott DR, Katz KS, Sicotte H, and Pruitt KD (2000). NCBI's LocusLink and RefSeq. *Nucleic Acids Res* **28**, 126–128.

[20] Boguski MS, and Schuler GD (1995). Establishing a human transcript map. *Nat Genet* **10**, 369–371.

[21] Schuler GD (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* **75**, 694–698.

[22] Deloukas P, et al. (1998). A physical map of 30,000 human genes. *Science* **282**, 744–746.