

# Genome cross-referencing and XREFdb: Implications for the identification and analysis of genes mutated in human disease

Douglas E. Bassett Jr.<sup>1,4</sup>, Mark S. Boguski<sup>4</sup>, Forrest Spencer<sup>2</sup>, Roger Reeves<sup>3</sup>, Su-hyon Kim<sup>2</sup>, Theresa Weaver<sup>2</sup> & Philip Hieter<sup>1</sup>

Comparative genomics approaches and multi-organismal biology are valuable tools for genetic analysis. Cross-species connections between genes mutated in human disease states and homologues in model organisms can be particularly powerful, as model-organism gene function data and experimental approaches can shed light on the molecular mechanisms defective in the disease. We describe a project that is systematically identifying novel expressed sequence tag (EST) sequences that are highly related to genes in model organisms and mapping them to positions on the mouse and human maps. This process effectively cross-references model organism genes with mapped mammalian phenotypes, facilitating the identification of genes mutated in human disease states via the positional candidate approach. A public database, XREFdb (<http://www.ncbi.nlm.nih.gov/XREFdb/>), disseminates similarity search, mapping and mammalian phenotype information and increases the rate at which these cross-species connections are established.

The positional cloning approach has been used in the identification of 84 genes so far that, when mutated, cause disease in humans. This method involves linkage and/or translocation breakpoint analysis followed by a search for genes in the candidate region and, ultimately, the identification of the gene of interest. Once the gene is cloned and its sequence is determined, a similarity search is performed to identify orthologues and family members in other species (in particular, model organisms). These results frequently provide the first clues toward the elucidation of human gene function and can offer powerful insight into the disease process itself. Such cross-species relationships also allow available function data for model organism genes to be immediately applied to the study of their correlates in more complex eukaryotes. Furthermore, the experimental advantages of model organisms become available for functional analysis of the gene product of interest and elucidation of the molecular mechanisms defective in the human disease state. Studies of the *Saccharomyces cerevisiae* *MEC1* and *TEL1* genes, for example, have provided valuable insight into the function of the human *ATM* gene (mutated in ataxia telangiectasia), and further analysis is in progress utilizing the yeast experimental system<sup>1</sup>. Similarly, the *Drosophila melanogaster* *patched* gene aided scientists in understanding the underlying disease process in nevoid basal cell carcinoma syndrome, caused by mutations in the human homologue, *PTC2*.

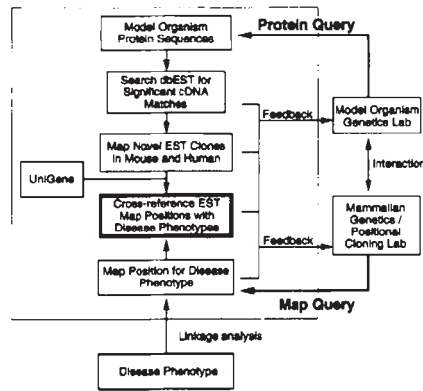
Although a proven method, cloning a gene mutated in a human disease state based solely on genomic position can be extremely labour- and resource-intensive. The approach has been expedited in a number of instances, however, by the availability of biological information on model organism proteins related to the human disease gene product. In the case of hereditary non-polyposis colon cancer (HNPCC), for example, patient tumour cell lines showed a particularly high frequency of DNA replication errors (RER+). The observation of similar replication errors in yeast *msh2* and *Escherichia coli mutS* mutants led two groups<sup>3,4</sup> to clone the human orthologue of these model organism genes, *hMSH2*, which maps to the same region as the disease phenotype and is mutated in HNPCC patients.

The XREF project seeks to establish similar connections in a systematic fashion. Mammalian cDNAs that are highly related to genes of known function in model organisms are first identified and then placed on the mouse and human maps. This promises to accelerate the identification of genes mutated in human disease by providing candidate genes for phenotypes mapped to the same location as the cDNA. More importantly, these candidates will in many cases be attached to functional information derived from experimentation on orthologues and family members in one or more model organisms. Analysis of genes discovered using this approach will thus benefit both from this body of pre-existing data and from the future prospect of utilizing experimental approaches available in model organisms to forward studies of the human gene and the disease process itself. XREFdb, a database designed to establish cross-references between model organism genes and mammalian phenotypes was established as an e-mail based community resource in September of 1994. Since its introduction on the World Wide Web in May of 1995, the cross-referencing approach and XREFdb<sup>5</sup> have aided investigators in identifying human homologues of model organism genes<sup>6</sup>, cross-referencing model organism genes with mammalian map positions<sup>7</sup> and establishing links between model organism genes and human disease states<sup>8</sup>. We report here a number of important advances in the cross-referencing method and database that have created a tool of particular utility and value to investigators working to identify genes that cause disease when mutated. The latest version of XREFdb, available on the World Wide Web (<http://www.ncbi.nlm.nih.gov/XREFdb/>), includes a new map position-based query tool to catalyse the positional identification of candidate genes for human genetic diseases. In addition, the system now integrates the vast resource of expressed sequence tag mapping data generated by the Radiation Hybrid Mapping Consortium, believed to represent over 16,000 human genes<sup>9</sup>. As a result, model organism and human biologists can now make direct use of this public resource

<sup>1</sup>Department of Molecular Biology and Genetics, <sup>2</sup>Center for Medical Genetics, <sup>3</sup>Department of Physiology, The Johns Hopkins University School of Medicine, 725 N. Wolfe Street, Baltimore, Maryland 21205-2185, USA. <sup>4</sup>The National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA. Correspondence should be addressed to P.H.

**Fig. 1** The XREF project scheme.

XREFdb is systematically cross-referencing functional information about genes in model organisms with positions on the mouse and human genetic maps. Model organism protein sequences are searched against dbEST using the TBLASTN program to identify significant EST matches. Novel ESTs are then mapped to positions on mouse and human chromosomes by the XREF project. These map data, supplemented with EST mapping data from the International RH Mapping Consortium deposited in UniGene, are used to establish potential connections with disease phenotypes in OMIM mapped to the same position via linkage analysis. All information generated by the project is freely accessible, and similarity search, EST mapping and cross-reference information is automatically fed back on a monthly basis to all XREFdb queriers who have expressed an interest in a related gene or map position for a phenotype.



to clone mammalian homologues of genes of interest, establish connections between model organism genes and disease phenotypes and identify candidate genes for human genetic disorders.

### Rationale: model organisms and human disease

Cross-referencing model organism genes with mammalian phenotypes involves integration of EST mapping data, mammalian phenotype data and information about model organism genes. These data, when linked together using a relational database system, create a powerful tool for gene discovery and, in particular, disease gene identification. The scheme by which XREFdb and cDNA mapping data from the XREF project can aid in the identification of human disease genes is outlined in Fig. 1. Open reading frames from model organisms are compared on a monthly basis with human, mouse and rat cDNA sequences in the EST division of GenBank (dbEST<sup>10</sup>) using the BLAST algorithm (TBLASTN)<sup>11</sup>. Those ESTs that exhibit significant sequence similarity with model organism proteins may represent mammalian orthologues or gene family members. Novel EST matches (those not already represented by a fully sequenced gene in GenBank) are prioritized for mapping in mouse and human by the XREF project based on the statistical significance of the sequence similarity. Mapping information generated by the project is combined with EST mapping data generated by the RH Mapping Consortium and used to effectively cross-reference genes in model organisms with positions of related genes on the mammalian maps. More importantly, XREFdb links these model organism genes to mapped human phenotypes in the Online Mendelian Inheritance in Man (OMIM)<sup>12</sup> database to suggest potential connections between these genes and human disease states.

The successful use of model organism protein sequences as probes to identify genes mutated in human disease depends on the frequency with which human disease genes have orthologues or family members in model organisms. To estimate how frequently human disease genes can be expected to have model organism correlates, we have searched all examples of disease genes identified by positional cloning to date against all protein sequences in the public databases for each of five model organisms: *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and

*Escherichia coli* (Table 1). The results reveal that 93% (78/84) have matches with at least moderate statistical significance (BLASTX  $P$  value  $\leq 1.0e-10$ ) in one or more of these five model organisms. A particularly large fraction (81%, 68/84) have matches in mouse at this significance level, despite the fact that complete sequences for only 3% of mouse genes are currently present in GenBank. There is the result of a strong bias, as mouse homologues of genes mutated in human disease are frequently cloned soon after, or in parallel with the identification of the human gene for the purpose of generating transgenic animal models. Excluding mouse, 70% (59/84) of positionally cloned genes match a protein in one or more of the other model organisms with moderate significance, and approximately half (41/84, 49%) have matches at least as significant as the human NF1 / yeast Ira2p pair ( $P \leq 6.2e-39$ , corresponding to high-very high statistical significance in Table 1), for

which cross-species functional complementation has been demonstrated<sup>13</sup>.

Extrapolating from these results, we expect that a newly discovered human disease gene will have approximately a two thirds chance of being significantly related to one or more genes in *D. melanogaster*, *C. elegans*, *S. cerevisiae* and/or *E. coli* given the current status of genome sequencing in these organisms—a figure that strongly supports the potential for cross-referencing as a means for disease gene identification. Furthermore, we can expect this fraction to climb even higher as more genomic sequence data are collected, allowing the establishment of new connections between model organism genes and human disease. The complete genomic sequences of *D. melanogaster* (currently 5% complete) and *C. elegans* (65% complete) in particular, can be expected to contribute substantially to this increase. Although their complete genomic sequences are not yet available, there are already many cases where disease genes identify similar genes in these model eukaryotes, while no significant homology exists with any open reading frame (ORF) in the finished yeast genome (see aniridia in Table 1, for example).

The results in Table 1 for *S. cerevisiae* and *E. coli* are of particular interest, as these genomes are completely sequenced and publicly available<sup>14,15</sup>. Thus, for each sequence searched against these databases, the complete set of matches is determined and the best match is known with certainty. Of these human disease genes, 8% (7/84) have *E. coli* matches of high or very high statistical significance, and 18% (15/84) have yeast matches at the same stringent significance level. The ability to identify the absolute best match in yeast or *E. coli* for a given disease gene allows investigators to take advantage of these experimental systems to study a disease process using the most highly conserved gene or genes present in these species. The availability of the complete sequence for an organism also allows the entire set of proteins to be used to probe the EST database for related human cDNAs, potentially aiding in the discovery of new genes involved in disease states using the cross-referencing approach. Presently, 39% (1688/4334) of all *S. cerevisiae* protein queries in XREFdb identify one or more EST matches when searched against dbEST with moderate statistical significance or better (TBLASTN  $P$  value  $\leq 1.0e-10$ ). The yeast example indicates the potential for cross-

**Table 1 legend, opposite page.** Summary of similarity search results for each of the 84 examples of human disease genes positionally cloned so far compared with five model organism-specific protein databases. The frequency with which these significant similarities are observed justifies the use of an EST-based cross-referencing approach for the identification of genes mutated in human disease. Matches highlighted in blue are of very high statistical significance, having BLASTX  $P$  values  $< 1.0e-80$ . Green-highlighted matches have high statistical significance (BLASTX  $P$  values between  $1.0e-80$  and  $6.2e-39$ ). Blue and green matches are at least as significant as the human NF1 / *S. cerevisiae* Ira2p relationship ( $P = 6.2e-39$ ). (This cutoff is of particular biological significance, as the human NF1 cDNA complements the yeast *ira2* mutant.) Matches highlighted in yellow have moderate statistical significance (BLASTX  $P$  values between  $6.2e-39$  and  $1.0e-10$ ). Cells are left blank in cases where no match is currently present in a particular model organism at  $P \leq 1.0e-10$  for a given positionally cloned gene. For each organism, the approximate number of genes for which sequence data are publicly available is provided, along with an estimate of the total gene repertoire and percentage of genes sequenced. The NCBI unique identifier is provided in the gene name column for the *E. coli* ORF exhibiting moderate sequence similarity with the human DTD gene, which was not named as of the preparation of this manuscript. A version of this table which includes precise BLASTX  $P$  values for each match is available on the World Wide Web (<http://www.ncbi.nlm.nih.gov/Bassett/modelorgs/>) and will be updated as positional cloning and model organism genomic sequencing projects progress.

**Table 1 • Positionally cloned genes and model organism proteins**

Disease	Homo Sapiens Positionally Cloned Disease Genes			M. musculus 2500 / 80000 (3%)		D. melanogaster 900 / 17000 (5%)		C. elegans 9100 / 14000 (65%)		S. cerevisiae 6000 / 6000 (100%)		E. coli 4300 / 4300 (100%)	
	MIM#	Gene Symbol	GenBank Accession#	Gene Symbol	GenBank Accession#	Gene Symbol	GenBank Accession#	Gene Symbol	GenBank Accession#	Gene Symbol	GenBank Accession#	Gene Symbol	GenBank Accession#
Aarskog-Scott syndrome	305400	FGD1	U11690	FGD1	U22325			<i>CGD9.1</i>	Z68159				
Achondroplasia	100800	FGFR3	M58051	FGFR3	S56291	<i>FRI</i>	D14976	<i>EG-15</i>	U39761	<i>IPL1</i>	U07163		
Adenomatous polyposis coli	175100	APC	M74088	APC	M88127	<i>APC</i>	U77947	<i>K04G2.8</i>	Z75712				
Adrenoleukodystrophy, X-Linked	300100	ALD	Z21876	ALDP	Z36337			<i>T02D1.c</i>	Z83319	<i>PXA1</i>	U17065	<i>YDDA</i>	X71917
Agammaglobulinemia, X-Linked	300300	XLA	X65892	<i>Usa8</i>	U26991								
Alzheimer disease, type 3	104311	PS1	L76517	PS1	L76517			<i>SEL-12</i>	U35660				
Alzheimer disease, type 4	600759	PS2	L44577	PS2	U57324			<i>SEL-12</i>	U35660				
Amyotrophic lateral sclerosis	105400	SOD1	K00065	SODC	X06683	<i>SLD</i>	Z19561	<i>SCD-1</i>	Z29135	<i>SCD1</i>	J03279		
Aniridia	106210	PAX6	M77844	PAX6	M93650	<i>EYE</i>	X79493	<i>VAR-3</i>	U31537				
Ataxia telangiectasia	208900	ATM	U26455	ATM	U43678	<i>MEI-41</i>	U34925	<i>T06E4.3</i>	Z70756	<i>TEL1</i>	Z35849		
Barth syndrome	302060	BTHS	X92762					<i>ZK909.2</i>	Z68303	<i>P8659.5</i>	U40829		
Bloom syndrome	219000	BLM	U39817					<i>K02F.1</i>	U69952	<i>SCS1</i>	Z35849	<i>REP2</i>	M30198
Breast & ovarian cancer, early onset, type 1	113705	BRCA1	U14680	BRCA1	U36475								
Breast cancer, early onset, type 2	600185	BRCA2	U43746	BRCA2	U65594								
Ceroid lipofusidosis, infantile neuronal	256730	INCL	U44772					<i>F44C.4.5</i>	U50313				
Chediak-Higashi Syndrome	214500	CHS	U67615	<i>Beige</i>	U70015			<i>T10F.1</i>	Z46242	<i>YCR032g</i>	X59075		
Chondrodysplasia punctata	302950	ARSE	X83573	STS	U37545			<i>T12.F.1</i>	U63160			<i>YDU</i>	L10328
Chorioideremia	303100	CHM	X78121	<i>GDB</i>	L36314	<i>GQY</i>	L03209	<i>GD1</i>	U00002	<i>GDI1</i>	S69371	<i>1571</i>	AE000247
Chronic granulomatous disease	306400	NCF1	M55067	NCF1	L11455								
Congenital adrenal hyperplasia	201910	CYP21	M26856	CYP21	M64933	<i>Cyt P450</i>	U44753	<i>C4PC8.4</i>	U61945				
Congenital adrenal hypoplasia	300200	AHC	S74720	<i>AKC1</i>	U41568	<i>SVP</i>	M28863						
Cystic fibrosis	219700	CFTR	M28668	CFTR	M69493	<i>MDR50</i>	L07065	<i>F21G1.2</i>	Z81016	<i>YCF1</i>	L35237	<i>MDL</i>	L08627
Dystrophic dysplasia	222600	DTD	U14528	STCB	D42049			<i>ZK287.2</i>	Z19757	<i>SUL1</i>	Z35134	<i>1787457</i>	AE000219
Duchenne muscular dystrophy	310200	DMD	M18533	<i>DMD</i>	M68859	<i>SG10.0</i>	M92688	<i>W04D2.1</i>	Z75552				
Emery-Dreifuss muscular dystrophy	310300	STA	X82434										
Epidemiologic palmo-plantar keratoderma	144200	KRT7	X75015	<i>KIC1</i>	L99193	<i>LAMC</i>	L07933	<i>W10G6.3</i>	Z81140				
Fanconi syndrome, renal	300009	CLCN5	X91906	CLCN5	X78874			<i>G07H4.2</i>	Z68334	<i>GRT1</i>	Z93117		
Fragile histidine triad	601153	FHT	U46922							<i>D8740.15</i>	U28374		
Fragile X syndrome	309550	FMR1	S65791	FMR1	L23971								
Fragile site mental retardation, type 2	309548	FMR2	U48436	<i>LAF-4</i>	U34361								
Friedreich ataxia	229300	FRDA	U43747					<i>F59G1.5</i>	U53332				
Glycerol kinase deficiency	307030	GK	L13943	<i>GK</i>	U48403			<i>R11F4.1</i>	U38378	<i>GPK</i>	X69049	<i>GPK</i>	M18393
Gonadal dysgenesis	306100	SFY	L08063	<i>SFY</i>	U03645	<i>SOXDP</i>	X96419	<i>F40E10.2</i>	Z69792				
Greenough granular dystrophy, type 1	122200	BIGH1	M77349	<i>Beta ig hd</i>	L19632								
Hemochromatosis	235200	HFE	U60319	<i>HFE</i>	U66849								
Hereditary multi-infarct dementia	125310	NOTCH3	X79439	<i>Notch3</i>	X74760								
Hereditary multiple exostoses	133700	EXT1	S79639	EXT1	X96639			<i>T12F6.3</i>	Z73425				
Hereditary non-polyposis colon cancer	120436	MLH1	U07418	MLH1	U59881					<i>MRH1</i>	U67187	<i>MRH1</i>	Z11831
Hereditary non-polyposis colon cancer	120436	MSP2	U03911	<i>MSP2</i>	X81143	<i>SPEL1</i>	U17893	<i>ZK1127.11</i>	U58758	<i>MSP2</i>	M84170	<i>MRUS</i>	M64739
Hereditary pancreatitis	276000	TRYP1	U70137	<i>TRYP</i>	X04573								
Hermansky-Pudlak syndrome	203300	HPS	U65676										
Holt-Oram syndrome	142900	TBX5	Y09445	<i>TBX5</i>	U57330	<i>TS</i>	M81796	<i>F21111.3</i>	U11279				
Huntington disease	143100	HD	L12392	<i>HD</i>	U24233								
Hyperkplexia	149400	GLRA2	X52009	<i>GLRA1</i>	S73717	<i>GLUC</i>	U58776	<i>QRP</i>	U40573				
Hypophosphatemic rickets, X-linked	307800	XPH	U60475	<i>PHX</i>	U39119			<i>ZK37.6</i>	Z69894				
Kallmann syndrome	308700	KAL	M97252										
Lissencephaly	247200	LIS1	L13385	<i>LIS1</i>	L34659	<i>TAF180</i>	U06460	<i>T03F6.1</i>	Z81113	<i>YPL151C</i>	Z73507		
Long QT Syndrome, type 1	192500	KVLQT1	U40990	<i>KVLQT1</i>	U70968	<i>SH</i>	M17211	<i>C25B8.2</i>	U41556				
Lowie syndrome	309000	CCRL	M88162	<i>S-P</i>	U39203			<i>C16C2F</i>	Z81536	<i>YJL020C</i>	Z47047		
Machado-Joseph disease	109150	MJD1	S75313					<i>E28B.6</i>	Z81971				
Marfan syndrome	154700	FBN1	L13923	<i>FBN1</i>	U22493	<i>CFR</i>	M33753	<i>ZK483.1</i>	U13046				
Maturity onset diabetes of the young	600496	TCF1	X59869	<i>TCF1</i>	X61385			<i>POP-1</i>	U37532				
Mcl.eod syndrome	314850	XK	Z32684										
Menkes syndrome	309400	MPK	X69208	<i>MPK</i>	U71691					<i>CCZ2</i>	L36317		
Multiple endocrine neoplasia 2A	171400	RET	M57464	<i>RET</i>	X67012	<i>DRE1</i>	S70576	<i>EG-15</i>	U30761	<i>PHO85</i>	Y00867		
Myotonic dystrophy	160900	DM	L19268	<i>DM</i>	Z39115	<i>NOR</i>	Z35103	<i>NOR</i>	Z34989	<i>N1727</i>	X92517		
Neurofibromatosis, type 1	162200	NF1	M89914	<i>NF1</i>	L19379			<i>C07B5.1</i>	Z42626	<i>RA2</i>	M33729		
Neurofibromatosis, type 2	101000	NF2	L11353	<i>NF2</i>	L27099	<i>Merlin</i>	U49724	<i>C01GR.5a</i>	U89439				
Nevoid basal cell carcinoma syndrome	109400	PTC	U59464	<i>PTC</i>	U46155	<i>PTC</i>	M20999	<i>ZK675.1</i>	Z46812	<i>LPA11</i>	U33335		
Norrie disease	310600	NDP	X65882	<i>NDP</i>	X83794								
Obesity	164160	OB	U18915	<i>OB</i>	U18812								
Ocular albinism	300500	OAI	Z48804	<i>OAI</i>	U63918								
Pallister-Hall syndrome	146510	GLI3	M34366	<i>GLI3</i>	X95255	<i>CD</i>	X54360	<i>TBA.7</i>	M93956	<i>YJL056C</i>	Z49331		
Pancreatic carcinoma	260350	DPF1	U44378	<i>MAOP2</i>	U69530	<i>AMAD</i>	U10328	<i>SM4.7</i>	U34396				
Polycystic kidney disease, type 1	173900	PKD1	L33243										
Polycystic kidney disease, type 2	173910	PKD2	U50928					<i>ZK945.8</i>	Z48544				
Progressive myoclonus epilepsy	254800	EPM1	U46692	<i>Cystatin B</i>	U59807								
Retinitis pigmentosa, type 3, X-linked	312610	RP3	X97668			<i>BLU1</i>	X58530	<i>F07C3.4</i>	U50308				
Retinoblastoma	180200	RB1	M15400	<i>RB1</i>	M26391	<i>REB</i>	X96975						
Rieger Syndrome	180500	REG	U69961	<i>OTLX2</i>	U80036	<i>AL</i>	L08401	<i>UNC-30</i>	U69961				
Simpson-Golabi-Behmel syndrome	312870	SGS1	L47125	<i>S-gpp-41</i>	X83577	<i>Daily</i>	U31985						
Spinal muscular atrophy	253300	SMA	U18423					<i>C41G7.1</i>	Z81048				
Spinocerebellar ataxia 1	164400	ATX1	X79204	<i>ATX1</i>	X83542								
Spinocerebellar ataxia 2	183090	ATX2	U70323	<i>ATX2</i>	U70676								
Stargardt disease	248200	ABCR	U88667	<i>ADCI</i>	X75926	<i>MCPS</i>	M59076	<i>C48D4.4</i>	Z29117	<i>YPL226W</i>	Z73582	<i>59R3</i>	AE000181
Thomsen disease	160800	CLC1	Z25884	<i>SMCT1</i>	X67895			<i>E24F6.11</i>	U20543	<i>GEF1</i>	Z23117		
Treacher Collins syndrome	154500	TCOF1	U40847										
Tuberous sclerosis	191090	TSC	X75621	<i>TSC2</i>	U39818								
Von Hippel-Lindau disease	193300	VHL	L15409	<i>VHL</i>	U12579								
Waardenburg syndrome	193500	PAX3	U02309	<i>PAX3</i>	X59258	<i>GSH-D</i>	M14944	<i>R08B4.2</i>	Z68008				
Werner syndrome	277700	WFN	L76937					<i>E23A.2.2</i>	Z38112	<i>SCS1</i>	U22341	<i>REP2</i>	M30198
Wilms tumor	194070	WT1	X51630	<i>WT1</i>	M55512	<i>Stripe</i>	U42403	<i>MUA-1</i>	U80952	<i>FZF1</i>	X78104		
Wilson disease	277900	WND	U11700	<i>ATP7B</i>	U38477	<i>CA-P60A</i>	M62892	<i>B0365.3</i>	Z81028	<i>CCZ2</i>	L36317	<i>TR34</i>	AE000154
Wiskott-Aldrich syndrome	301000	WASP	U12707	<i>WASP</i>	U54788			<i>C07G1.4</i>	U58751	<i>YOR181W</i>	Z75089		

= Very high statistical significance  
 = High statistical significance  
 = Moderate statistical significance

Table 2 • Cross-referencing model organism genes with mammalian phenotypes using XREFdb's protein query option

Protein/ORF Name	Model Organism Protein				GenBank Accession# for Related Mammalian EST probe	XREF Mouse Locus Symbol	Mouse Map Position		Inferred Human Position	Disease(s) Mapped to Region		
	Organism	GenBank Accession	Function	BLASTX P-value			Chromosome	Distance from Top		Map position	Description	MIM#
ZK742.1 Crm1p	<i>C. elegans</i>	U84855	Unknown Chromosome structure	2.80E-44	T07021	<i>D11Xf4</i>	11	19 cM	2pter-p11.1	2p25.3	Melanoma	273300
	<i>S. cerevisiae</i>	D13039		2.50E-36						2p24-p21	Spastic paraplegia 4	182601
										2p23-p22	Deafness, autosomal recessive 9	601071
										2p21	Holoprosencephaly 2	157170
										2p21-p16	Drusen, radial	128600
										2p21	Glaucoma 3, primary infantile	231300
										2p13.3-p13.1	Limb-girdle muscular dystrophy 2B	253601
ZC302.2 Hir1p	<i>C. elegans</i>	Z73978	Unknown	4.40E-14	T08110	<i>DSXf55</i>	5	79 cM	12q22-qter	12q	Oler-Rendu-Weber syndrome 2	600376
	<i>S. cerevisiae</i>	L03838	Transcription factor, histone transcription regulator	1.30E-09						12q13.2-q24.1	Fibrosis of the extraocular muscles	135700
										12q22-qter	Noonan syndrome 1	183950
										12q22	Testicular tumors	273300
										12q23-q24.1	Unilateral mammary syndrome	181450
										12q23-q24.1	Darier disease	124200
Orc1p	<i>S. cerevisiae</i>	U34860	DNA replication origin recognition	1.80E-20	T98858	<i>D4Xf216</i>	4	123 cM	1p35.3-p12	1p	Pheochromocytoma	171300
										1p36-p34	Peutz-Jeghers syndrome	175200
										1p36.1-p34	Schwartz-Jampel syndrome	255800
										1p32	Deafness, autosomal dominant 2	600101

Five model organism protein queries are shown (leftmost column) that significantly match three ESTs in dbEST mapped by the XREF project. Disease phenotypes overlapping the human map position to which each particular EST clone has been mapped are listed in the rightmost column, omitting those diseases for which a causal mutation in a particular gene has already been identified. The human gene represented by each XREF locus is a candidate gene for any disease state mapped to the same location. Function data known for a related model organism gene with highly significant sequence similarity to the mapped EST can suggest a particularly appealing candidate. Some suggestive cross-references among these XREF loci include melanoma with *D11Xf4* (EST similar to Crm1p, a yeast chromosome structural protein), testicular tumours with *DSXf55* (EST similar to Hir1p, a yeast transcription factor known to regulate histone transcription), and Peutz-Jeghers syndrome (hamartomatous intestinal polyposis) with *D4Xf216* (EST similar to Orc1p, a component of the yeast DNA replication origin recognition complex). These loci represent a subset of the over 360 total mapped so far by the XREF project using EST probes exhibiting significant sequence similarity with genes in human, mouse, fruit fly, nematode, bacteria, and other organisms. All EST sequences chosen for mapping by the XREF project are novel (not represented by a full length cDNA sequence in GenBank) as of the time of selection. BLASTX P values were obtained by searching the DNA sequence of the mapped EST against the non-redundant protein database maintained at NCBI. Phenotype descriptions, MIM numbers, and map positions are from the Online Mendelian Inheritance in Man (OMIM) database. Protein descriptions were obtained from Swiss-Prot database, release 34.0 and GenBank, release 98.0.

referencing model organism biology with mammalian maps and phenotypes, and preliminary data available for *D. melanogaster* and *C. elegans* suggest that the cross-referencing method will become increasingly valuable as more genome sequence from model multicellular eukaryotes becomes available — offering the opportunity for discovery of human disease genes that have no close family members or orthologues in *S. cerevisiae* using this approach.

### XREFdb and genome cross-referencing

XREFdb is designed to give mammalian geneticists, researchers working with genes in model organisms and positional cloners alike opportunities to establish connections that can lead to disease gene identification. The XREF project has currently mapped 256 cDNA probes exhibiting similarity to genes in mouse, fruit fly, nematode, yeasts, bacteria and other organisms to positions on the mouse and human genetic maps (F.S. *et al.*, manuscript in preparation). ESTs selected for their novelty and significant sequence similarity to genes in one or more model organisms are genetically mapped in mouse using the Jackson Laboratory BSS interspecific backcross DNA panel<sup>16</sup> and to human chromosomes using a somatic cell hybrid mapping panel from the National Institute of General Medical Sciences (NIGMS)<sup>17</sup>. Using the rich resource of available synteny data, defining local regions of conserved map order of homologous genes between mouse and human, the XREFdb software automatically infers subchromosomal human positions based upon the mouse mapping results. The database then uses these human map positions to connect each mapped EST clone to all disease phenotypes known to map to the same region according to the latest version of the OMIM database. Within XREFdb, EST clones are also linked back to the model organism genes for which they were selected as XREF

mapping candidates, as well as to other genes exhibiting significant sequence similarity. XREFdb allows users to access current cross-referencing information using either protein sequences or mammalian map positions of interest and suggests potential connections between model organism genes and mutant phenotypes in mammals.

### Accessing XREFdb via protein query

The XREFdb protein query option (Fig. 1) allows researchers to establish connections from genes of interest in model organisms to their human homologues and, ultimately, to corresponding human phenotypes. Users can query XREFdb with a protein sequence of interest from any organism. These queries are stored by the cross-referencing database, and each sequence is automatically searched against dbEST on a monthly basis. For each mapped EST match, XREFdb contains all relevant positional data in mouse and human, as well as data for human disease phenotypes located in the same region. The monthly execution of searches is important, as dbEST is growing at a rapid pace and each EST sequence deposited represents a new chance that a homologue for a gene of interest will be represented in the database. New matches are also flagged by XREFdb such that users of the system can rapidly identify and analyse the latest similarity search information within the context of matches that have been inspected previously. Table 2 provides examples of protein queries for which related ESTs have been mapped by the XREF project. Human diseases that map to regions overlapping those of the mapped ESTs can suggest links between model organism genes and mammalian phenotypes. The *S. cerevisiae* Crm1p protein believed to play a role in chromosome structure<sup>18</sup>, for example, has a significant match with an EST (GenBank no. T07021) mapped by the XREF project to a region of chromosome 2 known to harbour

a melanoma predisposition locus. Thus the gene from which this EST derives is a logical candidate for the cancer phenotype by virtue of both position and suggested function.

**Accessing XREFdb via map query**

XREFdb also accepts queries based on map position (Fig. 1). Mammalian geneticists and positional cloning laboratories can use this feature to great benefit if a particular mouse or human subchromosomal location contains a gene known to give rise to a particular phenotype or disease when mutated. XREFdb will accommodate standing queries of mouse or human map regions, tracking them and automatically feeding back data on new ESTs mapped to the regions of interest each month, along with function and similarity search information for model organism genes related to each EST sequence. These mapped ESTs, linked to functional data for similar model organism genes, can allow mammalian geneticists to identify likely candidate genes for a given disease and test them directly, possibly averting the resource-intensive gene hunt which comprises much of the cost of the typical positional cloning endeavor.

Since the rate-limiting step in positional cloning is typically the identification of candidate genes, expediting the candidate identification process will accelerate positional identification of genes mutated in human disease. Box 1 highlights the newly implemented MAP query option and illustrates the application of cross-referencing and XREFdb as a positional cloning tool. Investigators with linkage, translocation breakpoint, or other data placing a disease locus within a specific map region in either human or mouse can query the location of interest using XREFdb. The database will automatically suggest potential candidate genes from the EST database that map to the region. Model organism genes similar to these positional candidates are also reported, allowing for prioritization of each candidate based upon biological function information inferred from genes that exhibit significant sequence similarity with the ESTs.

The information in Table 2 and Box 1, along with similar cross-referencing data for hundreds of other XREF loci, is readily available through XREFdb and is updated on a monthly basis. The cross-referencing database also integrates EST mapping data from the Radiation Hybrid Mapping Consortium and the UniGene database<sup>20</sup> for use in establishing connections between model organism genes and mammalian phenotypes. XREFdb allows both model organism geneticists and mammalian geneticists aimed at cloning genes via positional information to make use of this powerful combination of protein function, similarity search, mapping and phenotype data to establish cross-references and, potentially, identify genes mutated in human disease states.

**Box 1 • XREFdb and XREF loci as a positional candidate cloning tool: querying by map position.**

Human Disease Phenotype			XREF Mouse Locus Symbol	Inferred Human Position	GenBank Accession for EST probe	BLASTX P-value	Related Model Organism Proteins			
Description	MIM#	Cytological Map Position					Model Organism	Protein/ORF Name	GenBank Accession	Function
Progressive external ophthalmoplegia, autosomal dominant with mitochondrial deletions	157640	10q	<i>D19Xf200</i>	10q23.1-q24.33	T50188	1.10E-36	<i>C. elegans</i>	W02B12	Z88521	Unknown
						1.40E-20	<i>S. cerevisiae</i>	Msr4p	X56444	Mitochondrial carrier protein
						6.40E-23	<i>D. melanogaster</i>	28S protease regulatory complex subunit	S78502	28S protease regulatory complex subunit
						6.80E-13	<i>S. cerevisiae</i>	Sun1p	D78022	Proteasome component
						2.40E-41	<i>C. elegans</i>	F10F2.1	Z46242	Unknown
						2.50E-31	<i>S. cerevisiae</i>	YCR032W	X59075	Unknown
Hepatocellular carcinoma	165320	2q14-q21	<i>D2Xf187</i>	2cen-q13	T03062	2.20E-42	<i>M. musculus</i>	RPL27A	X05021	Ribosomal protein
						7.60E-28	<i>S. cerevisiae</i>	Cyh2p	X01573	Ribosomal protein
						1.10E-71	<i>C. elegans</i>	T19A5.2	U63163	Ser/Thr protein kinase
						3.20E-43	<i>S. cerevisiae</i>	Ntk1p	D29980	Ser/Thr protein kinase
						2.80E-16	<i>C. elegans</i>	ZK656.9	Z70783	Unknown
						4.60E-18	<i>S. cerevisiae</i>	Tic4p	L12722	Transcription factor
Renal cell carcinoma	312390	Xp11.2	<i>D2Xf187</i>	2q11.1-q21.3	L28607	2.30E-17	<i>S. cerevisiae</i>	Bub1p	L32027	Cell cycle checkpoint control
						9.10E-15	<i>C. elegans</i>	R06C7.8	Z71266	Unknown
						3.40E-33	<i>S. cerevisiae</i>	N2005	X77114	Beta transthyretin
						1.70E-11	<i>C. elegans</i>	C36B1.5	Z80215	Unknown
						6.90E-13	<i>S. cerevisiae</i>	Sun1p	U51007	Proteasome component
						1.10E-71	<i>C. elegans</i>	T19A5.2	U59153	Ser/Thr protein kinase
Renal cell carcinoma	312390	Xp11.2	<i>DXXf176</i>	Xpter-q25	H09726	1.10E-71	<i>C. elegans</i>	T19A5.2	U59153	Ser/Thr protein kinase
						9.20E-43	<i>S. cerevisiae</i>	Ntk1p	D29980	Ser/Thr protein kinase
						1.80E-63	<i>C. elegans</i>	C36D10.1	U21324	Unknown
						1.30E-53	<i>S. cerevisiae</i>	YKL013C	X74152	Unknown
						2.10E-30	<i>S. cerevisiae</i>	YKL121W	Z28121	Unknown
						1.00E-35	<i>C. elegans</i>	F43G6.8	Z50070	Unknown
Renal cell carcinoma	312390	Xp11.2	<i>DXXf310</i>	Xp22.2-p11.21	N28768	1.00E-35	<i>C. elegans</i>	F43G6.8	Z50070	Unknown
						1.00E-33	<i>S. cerevisiae</i>	YER051W	U18796	Unknown
						4.40E-44	<i>S. cerevisiae</i>	Cbf5p	L12351	Centromere/microtubule binding protein

The table illustrates, using several example XREF loci, how a mammalian geneticist could make use of XREFdb to identify candidate genes for a given phenotype. Researchers interested in identifying the gene mutated to give rise to hepatocellular carcinoma which has been mapped to 2q11.1-q21.3 in human, for example, could query this map position using the cross-referencing database. ESTs mapped to the region by the XREF project and the RH mapping consortium would be retrieved and viewable using the World Wide Web, with links to genes in model organisms related to each EST sequence. In this case, there is an EST in the region of interest (D2Xf187) that was selected for mapping based upon its significant similarity with *S. cerevisiae* Bub1p, a protein involved in cell cycle checkpoint control<sup>19</sup>. Although the XREF project has also mapped several other ESTs to this region, the gene from which this EST derives could be considered a particularly provocative candidate based on established connections between cell cycle control and cancer. Similarly, a map query in XREFdb for chromosome 10q, to which a form of progressive external ophthalmoplegia featuring characteristic mitochondrial deletions has been mapped, reveals an interesting link to an XREF-mapped EST similar to Msr4p, a yeast mitochondrial protein. The XREF project has also mapped an EST clone similar to an *S. cerevisiae* centromere and microtubule binding protein (Cbf5p) to a renal cell carcinoma locus, providing a candidate gene for this phenotype. As with protein queries, XREFdb retains and tracks map queries for each user and flags new ESTs mapped to a region of interest every month.

Human phenotypes are used here as an example, but XREFdb also accommodates queries based on mouse map position. ESTs mapped by the XREF project with subchromosomal localizations overlapping the region to which each disease phenotype has been mapped are shown. All EST sequences chosen for mapping by the XREF project are novel (not represented by a full length cDNA sequence in GenBank) as of the time of selection. BLASTX P values were obtained by searching the DNA sequence of the mapped EST against the non-redundant protein database maintained at NCBI. Phenotype descriptions, MIM numbers and map positions are from the Online Mendelian Inheritance in Man (OMIM) database. Protein descriptions were obtained from Swiss-Prot database, release 34.0 and GenBank, release 98.0.

## Prospects

Model organisms have made numerous contributions in the field of human disease research. In particular, orthologs and family members of disease genes can be functionally characterized in model organisms — allowing us to learn more about the human disease gene. In cases where orthologs and family members of genes mutated in human disease are cloned across several species, the experimental advantages in each organism can be applied synergistically to characterize the homologous gene products and shed light on the molecular mechanisms of the human disease process. With the *S. cerevisiae* genome completely sequenced, the best yeast matches for each human gene mutated in a disease state can be determined with certainty, and, just as studies of *MEC1* and *TEL1* are forwarding ataxia telangiectasia research, these matches can be utilized to learn more about disease processes using the yeast experimental system<sup>21</sup>. As mouse, fly, and nematode sequencing projects progress, new matches to genes associated with human disease will be discovered. For these model organisms, proteins encoded by genes not yet sequenced will match human disease genes that previously had no model organism homologue, and moderately significant matches presently known (See Table 1) will in many cases be supplanted by more compelling matches. With completion of the sequences of these genomes, this approach for understanding human biology and disease processes will become increasingly powerful. Establishing the links between model organism and human disease at an early time point is critical, and we encourage investigators to access and make use of XREFdb for that purpose. Genome cross-referencing should open new doors for multi-organismal biology not only in terms of forwarding the functional analysis of disease gene products, but also in accelerating the identification of genes mutated in human disease states.

## Methods

**Cross-referencing database.** XREFdb is a publicly accessible database (<http://www.ncbi.nlm.nih.gov/XREFdb/>) implemented using the Sybase relational database management software. The database stores and cross-references protein sequence and function information, BLASTP and TBLASTN sequence similarity search data, EST mapping information and mammalian phenotype data. The World Wide Web interface to XREFdb is written in syperl, an extension of the perl programming language which allows direct access to a Sybase database. XREFdb allows users to establish an account containing proteins and map regions of interest. These queries are stored and tracked by the system, and each month new information related to each user's queries are flagged for closer inspection. A brief monthly e-mail report updates account holders on the status of each of their queries and potential cross-references.

A human, mouse and rat subset of the EST division of GenBank (dbEST), maintained at the National Center for Biotechnology Information, is generated and searched by XREFdb to identify mammalian ESTs similar to model organism proteins. The latest release of the Swiss-Prot protein database is imported into XREFdb as a source for public protein sequences. The non-

redundant nucleotide database at NCBI, which excludes EST sequences, is searched with EST queries using the BLASTN program. ESTs lacking exact matches with sequences in this database are considered novel and are evaluated for mapping by XREFdb based upon significance of sequence similarity with one or more model organism genes.

**cDNA Mapping.** EST clone inserts are isolated, labelled and used as a probe for Southern hybridization to filters prepared containing *EcoRI*, *TaqI* or *HindIII* digested DNA from The Jackson Laboratory BSS interspecific backcross DNA panel. RFLP data from this analysis is communicated to The Jackson Laboratory where a mouse chromosomal position is assigned using the Map Manager software (<http://mcbio.med.buffalo.edu/mapmgr.html>). Southern hybridization is also performed using each EST clone insert to probe DNA from NIGMS somatic cell hybrid panel #2, version 2 to determine chromosome localization in human. Human subchromosomal location is then inferred by the XREFdb software using regions of synteny between the mouse and human maps. Common markers are first identified between the BSS panel and the mouse chromosome committee maps. Markers mapped between and immediately outside the region bounded by the common markers that have also been mapped to positions in the human genome are then retrieved. The subchromosomal positions of these markers in human are used to generate one or more cytological ranges to which the locus is predicted to map in human. The chromosome assignment data is then used to confirm this inferred human position. For a complete description of the mapping method, see F.S. *et al.* (manuscript in preparation).

**Positionally cloned gene and model-organism protein similarities.** The cDNA sequences for 84 examples of positionally cloned human disease genes positionally cloned so far were searched using the BLASTX program against organism-specific subsets of the non-redundant protein database (nr) containing proteins from the following model organisms: *M. musculus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae* and *E. coli*. For each query, the best match in each model organism with a *P* value  $\leq 1.0e-10$  is reported. In cases where no match was present in a particular model organism at that significance level for a given positionally cloned gene, the cells are left blank. In those cases where a GenBank entry for the complete cDNA sequence for a positionally cloned gene does not exist, the GenBank accession number for the longest partial sequence available is listed. Default BLAST parameters were used for all searches with the exception that the *Z* parameter was set equal to the total number of letters in nr (69,232,276), to make all *P* values equal to those that would have been obtained had the entire database been searched. The SEG algorithm was used to mask low complexity subsequences. Model organism specific subsets of the non-redundant protein database were created using command-line Entrez (entrcmd, available as part of the NCBI toolkit at [ftp://ncbi.nlm.nih.gov/toolbox/ncbi\\_tools/](ftp://ncbi.nlm.nih.gov/toolbox/ncbi_tools/)), with the exception of the complete *E. coli* and *S. cerevisiae* protein sets, available at <ftp://ncbi.nlm.nih.gov/genbank/genomes/>.

## Acknowledgements

We thank S. Michaelis and D. Valle for comments on the manuscript. This work was supported by NIH grants HG00971-01 (P.H., M.B., F.S. and R. R.) and CA16519 (P.H.). D.B. is a student in the Predoctoral Training Program in Human Genetics at Johns Hopkins University (NIGMS P32GM07814) and an IRTA fellow at the National Institutes of Health.

- Morrow, D.M., Tagle, D.A., Shiloh, Y., Collins, F.S. & Hieter, P. *TEL1*, an *S. cerevisiae* homologue of the human gene mutated in ataxia telangiectasia, is functionally related to the yeast checkpoint gene *MEC1*. *Cell* **82**, 831–840 (1995).
- Hahn, H. *et al.* Mutations of the human homologue of *Drosophila* patched in the nevoid basal cell carcinoma syndrome. *Cell* **85**, 841–851 (1996).
- Fishel, R. *et al.* The human mutator gene homologue *MSH2* and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027–1038 (1993).
- Leach, F.S. *et al.* Mutations of a *MutS* homolog in hereditary non-polyposis colorectal cancer. *Cell* **75**, 1215–1225 (1993).
- Bassett, D.E. *et al.* Comparative genomics, genome cross-referencing and XREFdb. *Trends Genet.* **11**, 372–373 (1995).
- Chao, D.M. *et al.* A mammalian SRB protein associated with an RNA polymerase II holoenzyme. *Nature* **380**, 82–85 (1996).
- Banfi, S. *et al.* Identification and mapping of human cDNAs homologous to *Drosophila* mutant gene through EST database searching. *Nature Genet.* **13**, 167–174 (1996).
- Reifsnnyder, C., Lowell, J., Clarke, A. & Pillus, L. Yeast SAS silencing genes and human genes associated with AML and HIV-1 Tat interactions are homologous with acetyltransferases. *Nature Genet.* **14**, 42–49 (1996).
- Schuler, G. *et al.* A gene map of the human genome. *Science* **274**, 540–546 (1996).
- Boguski, M.S., Tolstoshev, C.M. & Bassett, D.E. Gene discovery in dbEST. *Science* **265**, 1993–1994 (1994).
- Altschul, S.F., Boguski, M.S., Gish, W. & Wootton, J.C. Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119–129 (1994).
- McKusick, V.A. *Mendelian Inheritance in Man* 11th edn. (Johns Hopkins University Press, 1994) (<http://www3.ncbi.nlm.nih.gov/Omim/>).
- Ballester, R. *et al.* The *NF1* locus encodes a protein functionally related to mammalian GAP and yeast IRA proteins. *Cell* **63**, 851–859 (1990).
- Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546–567 (1996).
- Blattner, F. *et al.* The *E. coli* Genome Project. *Microb. Comp. Genome* **1**, 357 (1996).
- Rowe, L.B. *et al.* Maps from two interspecific backcross DNA panels available as a community genetic mapping resource. *Mamm. Genome* **5**, 253–274 (1994).
- Drwina, H.L., Toji, L.H., Kim, C.H., Greene, A.E. & Mulivor, R.A. NIGMS human/rodent somatic cell hybrid mapping panels 1 and 2. *Genomics* **16**, 311–314 (1993).
- Funabiki, H., Hagan, I., Uzawa, S. & Yanagida, M. Cell cycle-dependent specific positioning and clustering of centromeres and telomeres in fission yeast. *J. Mol. Biol.* **217**, 23–37 (1991).
- Roberts, B.T., Farr, K.A. & Hoyt, M.A. The *Saccharomyces cerevisiae* checkpoint gene *BUB1* encodes a novel protein kinase. *Mol. Cell. Biol.* **14**, 8282–8291 (1994).
- Boguski, M.S. & Schuler, G.D. Establishing a Human Transcript Map. *Nature Genet.* **10**, 369–371 (1995).
- Hieter, P., Bassett, D.E. & Valle, D. The yeast genome—a common currency. *Nature Genet.* **13**, 253–254 (1996).