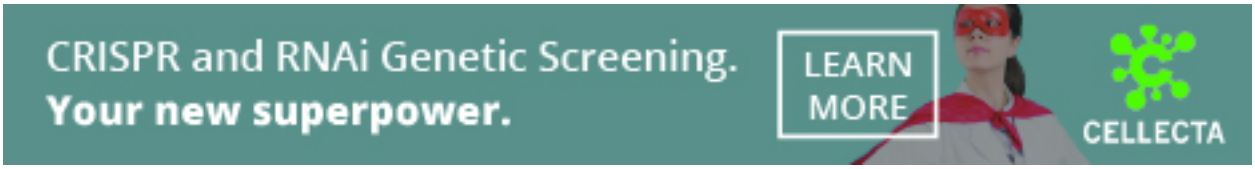
An expression-independent catalog of genes from human chromosome 22.

J A Trofatter, K R Long, J R Murrell, et al.

Genome Res. 1995 5: 214-224
Access the most recent version at doi:10.1101/gr.5.3.214



RESEARCH

An Expression-independent Catalog of Genes from Human Chromosome 22

James A. Trofatter,[1,3] Kimberly R. Long,[1,3] Jill R. Murrell,[1] Christy J. Stotler,[1] James F. Gusella,[1,2] and Alan J. Buckler[1,4]

Molecular Neurogenetics Unit, Departments of [1]Neurology and [2]Genetics, Massachusetts General Hospital and Harvard Medical School, Charlestown, Massachusetts 02129

To accomplish large-scale identification of genes from a single human chromosome, exon amplification was applied to large pools of clones from a flow-sorted human chromosome 22 cosmid library. Sequence analysis of more than one-third of the 6400 cloned products identified 35% of the known genes previously localized to this chromosome, as well as several unmapped genes and randomly sequenced cDNAs. Among the more interesting sequence similarities are those that represent novel human genes that are related to others with known or putative functions, such as one exon from a gene that may represent the human homolog of *Drosophila* Polycomb. It is anticipated that sequences from at least half of the genes residing on chromosome 22 are contained within this exon library. This approach is expected to facilitate fine-structure physical and transcription mapping of human chromosomes, and accelerate the process of disease gene identification.

A primary goal of the human genome initiative is the construction of fine-structure physical maps of the chromosomes in anticipation of full DNA sequence analysis. However, probably the most important purpose of this mapping, the identification and placement of human genes, can be carried out effectively before determining the complete sequence of the human genome, and can aid in increasing the resolution of the physical map. Low-resolution physical maps of human chromosomes have been described recently (Cohen et al. 1993), but considerably greater detail is needed to maximize their utility and proceed with large-scale sequencing. Identification of a representative set of genes or gene fragments corresponding to a specific genomic region would satisfy many of the requirements for finer mapping and would add a level of functional significance to these evolving maps. The resulting gene, or transcription maps, would provide a new framework for the study of structural, functional, and organizational aspects of chromosomes, and would lead to more efficient identification of genes involved in human disease. Consequently, recently the development of methods for rapid gene identification has received greater attention, and numerous strategies, including approaches based on hybridization and biological selection (Auch and Reth 1990; Duyk et al. 1990; Buckler et al. 1991; Lovett et al. 1991; Parimoo et al. 1991), have been proposed. Exon amplification, an example of the latter category, relies on selection for functional splice sites flanking exons and thereby, avoids problematic issues such as tissue specificity or relative mRNA abundance that are inherent to other gene identification approaches. Recently, we have modified the exon amplification technique to make it applicable to the isolation of gene sequences from very complex sources of genomic DNA (Church et al. 1994). As an initial test, we have applied this method to the isolation of large numbers of gene sequences from a single human chromosome.

[3]These authors contributed equally to this work.
[4]Corresponding author.
E-MAIL buckler@helix.mgh.harvard.edu; FAX (617)726-5736.

RESULTS

Construction of a Chromosome-specific Exon Library

Human chromosome 22 was chosen as a model for construction of exon libraries because of intensive mapping and disease gene identification efforts in this region of the genome. Plasmid DNA was prepared from pooled clones of each of the 130-microtiter plates in an arrayed cosmid library constructed from flow-sorted human chromosome 22 (LL22NC03), which represents approximately five equivalents of chromosome

5:214–224 ©1995 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/95 $5.00

22. An additional four pools were generated from a human-specific subset of cosmids derived from a human–hamster hybrid cell line, GM10888, containing chromosome 22 as its sole human component (Lichter et al. 1990). A total of 134 plate-pool cosmid DNAs were prepared, and each sample was digested and shotgun-cloned into the in vivo splicing vector, pSPL3 (Church et al. 1994). Plasmid DNA from pSPL3 subclones (pools of 500–2000 insert-containing clones) was transfected transiently into COS7 cells, which facilitated SV40 large T-mediated plasmid amplification and transcription. Cytoplasmic RNA derived from these transfectants was used in RNA–PCR amplifications, and the resulting products were cloned directionally as described in Methods. Exon clones (48 from each pool, or ~6400 clones) were arrayed, grown, and stored in 96-well microtiter plates. Approximately 24 clones from each of the first 100 pools were sequenced (a total of 2304 sequences generated) in a single pass, yielding 709 unique sequences, or an average of 7.1 unique sequences per pool. To determine the number of unique sequences in the rest of the exon library, the remaining 24 exon clones from 10 pools were sequenced and compared to all of the 709 initial sequences. An additional 40 unique sequences were produced, yielding an average of four remaining unique clones per pool. Therefore, we estimate that an average of 11.1 unique clones exist in each pool, and that sequence has been produced for ~64% (7.1 ÷ 11.1) of the unique clones in the first 100 pools. Because there are 134 exon pools, our upper estimate of the total number of unique sequences in the entire exon library is 1487 (134 × 11.1). Thus, the sequences that we have generated to date represent approximately half (47%) of those present in the library.

Complete sequence was produced for 91% of the clones analyzed, yielding a minimum average length of ~125 bp per clone. This is close to the value of 135 bp that we have reported previously for completely sequenced sets of exons (Church et al. 1994), and likely will be similar when all the sequences are complete. We have estimated the accuracy of the sequences produced to be ~99.5%, based on alignments to previously known, well-characterized gene sequences (Table 1).

An average threefold redundancy of clones corresponding to each unique sequence was observed, and is likely to be higher than this value when sequencing is complete; this may be attributable to biases introduced during certain steps of the procedure. Although cultured in separate wells, growth differences among individual cosmid clones within each pool may have introduced bias during the shotgun subcloning step. However, we have designed our approach to minimize the probability that more than one exon will be isolated from the same cosmid (to maximize the uniformity of exon distribution across the chromosome). By maintaining a high complexity of target DNA (i.e., large numbers of cosmids), the growth bias is likely to be distributed across several cosmids in each pool, thus increasing the likelihood that an exon will be derived preferentially from several of the more prevalent genomic clones. Bias in the library may also be attributable to differential splicing efficiency of particular exons, as well as preferential RNA–PCR amplification or exon cloning. These aspects may be more difficult to control and are likely to be dependent on the sequence composition of exons or splice site sequences. As a result, specific exons may have a reduced probability of being trapped, but it is likely that other exons from the same gene will be identified.

Sequence Data Base Comparisons

The sequences were compared to those in public data bases (Altschul et al. 1990; BLAST comparison with Genbank and EMBL versions and updates that were available 7/23/95), and a summary of these results is presented in Tables 2 and 3. One hundred ninety-nine of the 709 sequences (28%) analyzed are highly similar to known genes from a number of species. Included in these are 101 sequences that are identical (⩾97% nucleotide identity) to segments of previously identified human genes. These can be subdivided into 48 sequences from 24 different genes that were mapped previously to chromosome 22 (in some cases, multiple exons were isolated from the same gene), and 53 other sequences corresponding to heretofore unmapped genes and expressed sequence tags (EST). Included in the latter group were sequences from RanGTPase-activating protein 1 (Bischoff et al. 1995), phosphatidlyinositol 4-kinase (Wong and Cantley 1994), small nuclear ribonucleoprotein Sm D3 (Lehmeier et al. 1994), glutathionine *S*-transferase T1 (Pemble et al. 1994), and cadherin-13 (Tanihara et al. 1994), which are now localized provisionally to chromosome 22. In a few cases, the sequence identity was found with the com-

Table 1. Summary of sequence data base search results

Sequence number	Length	Representative clone	BLASTN	Species	Accession	% Sim	BLASTX	Species	Accession	% Sim
Identity, near identity, or homologue										
10	60	10H4	RanGTPase activating protein 1	HU	X82260	100	RanGAP1	HU	X82260	100
12	164	10H6	EST07064	HU	T09171	95				
13	206	10H7	merlin (NF2)	HU	L11353	100	merlin (NF2)	HU	S33809	100
14	84	10H8	yg84a03.r1 Homo sapiens cDNA	HU	R53636	96				
21	110	11C10	Ig lambda light chain*	HU	X51754	94				
25	138	11D7	EST03946	HU	T06057	90				
28	88	12F2	casein kinase I-alpha	BO	M76543	69	casein kinase I delta	RA	A46002	100
34	66	12G3	TUP1-like enhancer of split (TUPLE1)	HU	X75296	98	TUP1-like enhancer of split (TUPLE1)	HU	X75296	100
36	71	12G5	yg53a05.r1 Homo sapiens cDNA clone	HU	R25258	100	glutathione reductase	Pf	X83603	78
37	69	12G10	catechol methyltransferase, upstream*	RA	Z12652	85				
40	284	39E7	yd18f06.s1 Homo sapiens cDNA clone	HU	T69835	92				
42	144	13A7	fibulin A,B,C,D (FBLN1)	HU	X53743	100	fibulin A,B,C,D (FBLN1)	HU	X53743	100
48	102	13H4	fibulin A,B,C,D (FBLN1)	HU	X53743	100	fibulin A,B,C,D (FBLN1)	HU	X53743	100
59	67	14A9	ye81c04.r1 Homo sapiens cDNA	HU	R01350	100				
79	36	15F11	yh30c02.r1 Homo sapiens cDNA	HU	R24354	100				
80	153	16A1	DGCR2	HU	X84076	100	DGCR2	HU	X84076	100
86	111	16B2	phosphatidylinositol transfer protein	HU	M73704	77	phosphatidylinositol transfer protein	HU	Q00169	86
100	39	19E1	peroxisome proliferator activated recept.	HU	L02932	100				
101	130	19E2	Ewing sarcoma (EWSR1)	HU	X66899	100	Ewing sarcoma (EWSR1)	HU	X66899	100
107	58	19F7	catechol-O-methyltransferase (COMT)	HU	M65212	100	catechol-O-methyltransferase	HU	M65212	100
117	117	20A2	yg59e09.r1 Homo sapiens cDNA	HU	R35102	100				
128	123	21E 3	merlin (NF2)	HU	L11353	100	merlin (NF2)	HU	S33809	100
137	136	21F3	GM-CSF receptor beta chain (CSF2RB)	HU	M59941	98	GM-CSF receptor beta chain (CSF2RB)	HU	M59941	99
141	56	21F10	cDNA clone seq1298	HU	T10083	98				
145	249	22A4	ym20f07.r1 Homo sapiens cDNA	HU	H15984	95	aconitase	PI	P16276	96
158	75	24A7	yh83a06.s1 Homo sapiens cDNA	HU	R33473	100				
169	136	26A5	fibulin A,B,C,D (FBLN1)	HU	X53743	100	fibulin A,B,C,D (FBLN1)	HU	X53743	100
171	137	26B1	ya53d06.r1 Homo sapiens cDNA	HU	T67209	100	probable glutathione reductase	Ce	P30635	66
172	206	26B2	partial cDNA sequence; clone c-2lb05	HU	Z45264	99				
174	154	26B9	yf53a05.r1 Homo sapiens cDNA	HU	R11814	96				
175	75	28A1	EST02272	HU	M85753	100				
177	204	28A6	EST02272	HU	M85753	98				
180	141	29E3	Ig lambda chain variable	HU	M94113	100				
182	127	29E8	ye98f07.r1 Homo sapiens cDNA	HU	R07694	100				
188	112	2E4	Human homologue of yeast sec7	HU	M85169	77	Human homologue of yeast sec7	HU	S24168	97
192	116	2F6	chloride channel	HU	X77197	89	chloride channel	HU	X77197	97
211	193	32A1	partial cDNA sequence; clone c-2ka09	HU	F07861	96				
212	193	32A2	fetal-lung cDNA 5'-end sequence	HU	D31239	100				
220	123	32B8					period clock protein	MU	P08399	85
228	94	33F4	yt02e10.r1 Homo sapiens cDNA	HU	R53774	96				
235	111	36A1	gamma-glutamyltransferase related	HU	M64099	100	gamma-glutamyltransferase related	HU	A41125	100
251	100	39E2	breakpoint cluster region (BCR)	HU	U07000	100				
258	165	3A4	yf13g05.r1 Homo sapiens cDNA	HU	R07096	100	alpha-actinin	CH	S45673	62
265	118	3B12	partial cDNA sequence; clone c-10d08	HU	Z39110	92				
273	184	40A10	T10	MU	X74504	86	T10	MU	X74504	100
282	63	41E8	yi93a07.r1 Homo sapiens cDNA	HU	R80434	87				
283	96	41E9	TUP1-like enhancer of split (TUPLE1)	HU	X75296	100	TUP1-like enhancer of split (TUPLE1)	HU	X75296	100
284	96	41E11	ye83a05.r1 Homo sapiens cDNA	HU	R02199	100	R01H2.6 gene product	Ce	U00035	71
285	230	41F1	Ewing sarcoma (EWSR1)*	HU	X66899	99				
287	60	41F3	TUP1-like enhancer of split (TUPLE1)	HU	X75296	100				
291	70	42A7	cellular myosin heavy chain (MYH9)	HU	M81105	100	cellular myosin heavy chain	HU	A34876	100
300	113	43E4	GM-CSF receptor beta chain (CSF2RB)	HU	M59941	100	GM-CSF receptor beta chain (CSF2RB)	HU	P32927	100
301	144	9H4	Hk594-f Homo sapiens cDNA	HU	R41006	96				
305	105	43F3	D4S2463 homeobox-like gene	HU	U18977	90				
307	220	43F6	phosphatidylinositol 4-kinase	HU	L36151	100	phosphatidylinositol 4-kinase	HU	L36151	100
310	191	43F12	asparagine synthetase	HU	M27396	96	asparagine synthetase	HU	A27443	97
312	128	44A3	UK-HGMP sequence ID AAAASTE	HU	Z19815	88				
326	149	45E6	EST05647	HU	T07757	99	casein kinase I delta	RA	Q06486	100
329	251	45F3	beta-crystallin subunit beta B1	BO	X01808	86	beta-crystallin subunit beta B1	BO	S07264	93
332	204	45F12	fibulin D (FBLN1)	HU	U01244	99	fibulin D (FBLN1)	HU	U01244	100
333	126	46A1	STS UT6047	HU	L30702	99				
338	150	47A8	yg02g12.r1 Homo sapiens cDNA	HU	R17513	98				
343	136	47B9	yd21f03.r1 Homo sapiens cDNA	HU	T78985	100				
352	76	48F7	EST06311	HU	T08420	100				
360	218	4F3	thyroxine deiodinase	RA	M21476	89	thyroxine deiodinase	RA	P13700	95
364	53	4F8	ym43e06.r1 Homo sapiens cDNA	HU	H18565	87				
368	125	50A1	ye64b05.s1 Homo sapiens cDNA	HU	T99352	100	T10	MU	X74504	100
375	114	50B12					collagen alpha 3(VI) chain	HU	S13679	100
383	106	5D8	IL-2R-beta	HU	A07797	100	IL-2R-beta	HU	A07795	100
390	184	6B6	yg73f08.r1 Homo sapiens cDNA	HU	R51689	97	E46	MU	P28658	94
395	82	7G7	beta-alanine synthase	RA	M97662	88	beta-alanine synthase	RA	Q03248	100
396	140	7G10	break point cluster gene (BCR)	HU	Y00661	95	break point cluster gene (BCR)	HU	U07000	95
400	176	8D12	cDNA clone B270	HU	T20010	100				
409	174	9A10	SnRNP core protein Sm D3	HU	U15009	100	SnRNP core protein Sm D3	HU	U15009	100
415	145	9H12	c-sis (PDGFB)	HU	K01916	100	platelet-derived growth factor B	HU	X00561	100
417	168	100E1	yo77d12.r1 Homo sapiens cDNA	HU	H30621	99				
419	95	100E3	MDB1128R Mus musculus cDNA	MU	R75186	92				
421	111	100E7	rhoGAP protein	HU	Z23024	73	rhoGAP protein	HU	S34296	91
430	206	100F8	EST00943	HU	M78795	100				
440	50	27F8	NMDA receptor NR2C subunit*	RA	M91563	91				
466	190	52A12	UK-HGMP sequence ID AAADFQK	HU	Z20962	99				
470	122	52B11	immunoglobulin lambda-like (IGLL8)	HU	X52204	100				
471	96	52B12	EST04544	HU	T06655	89				
473	141	53E3	ras inhibitor	HU	M37191	90	Ras inhibitor	HU	C38637	100
474	101	53E4	ye81c04.r1 Homo sapiens cDNA	HU	R01350	100				
491	154	54A5	partial cDNA sequence; clone c-1va09	HU	F07091	100				
495	138	54B1	partial cDNA sequence; clone c-0gc09	HU	Z42345	97				

Table 1. *(continued)*

497	136	54B4	NIB1246 Homo sapiens cDNA 3'end	HU	T16398	97					
501	179	54B12	ym15f07.r1 Homo sapiens cDNA	HU	H11873	100					
509	235	55F7	IL-2R-beta (IL2RB)	HU	M26062	98	IL-2R-beta (IL2RB)	HU	A07795	95
515	70	55A6	GTP-binding protein	MU	D10715	97	GTP-binding protein	MU	P32233	100
518	94	56A11	AR1 protein (AR) mRNA	HU	U19345	100	PDGF responsive element bind. protein	MU	U20282	94
520	50	56B7	AR1 protein (AR) mRNA	HU	U19345	100				
523	159	57E1	partial cDNA sequence; clone c-zqe07	HU	Z45647	91				
531	68	57E11	p300 protein	HU	U01877	100	transcriptional adaptor protein p300	HU	A54277	100
533	112	57F11	3BP-1, an SH3 domain binding protein	MU	X87671	88	SH3 domain binding protein	MU	X87671	91
543	105	58B4	Na-dependent glucose cotransporter	PI	L02900	88	Na+/amino acid cotransporter	PI	P31636	94
545	141	58B11	yf39d07.r1 Homo sapiens cDNA	HU	R11100	100				
562	211	60A12	ym51d05.r1 Homo sapiens cDNA	HU	H23042	100				
575	177	61F4	HSL13041 Human fetal-lung cDNA	HU	D31239	96				
590	127	63E3	gamma-glutamyl transpeptidase (GGT)	HU	M24903	100				
592	153	63E6	partial cDNA sequence; clone c-zpd06	HU	Z45600	98				
593	70	63E7	ym51d05.r1 Homo sapiens cDNA	HU	H23042	100				
616	160	65F3	B melanoma antigen (BAGE) mRNA	HU	U19180	100	predicted trithorax protein	Dm	Z31725	57
620	76	81A4	partial cDNA sequence; clone c-2lf03	HU	P07808	97				
625	77	66A5	partial cDNA sequence; clone c-20d07	HU	Z44432	96				
630	75	66B2	breakpoint cluster region (BCR)	HU	M25947	100	breakpoint cluster region (BCR)	HU	B24847	100
635	149	66B10	brain tubulin, alpha chain	CH	J00912	85	alpha-tubulin	CH	P02552	95
636	60	67E1	sodium/glucose cotransporter	RA	D16101	72	Na+/amino acid cotransporter	PI	P31636	94
637	177	67E4	modifier 1	MU	X56690	68	modifier 3	MU	P30658	85
650	194	68A5	break point cluster gene (BCR)	HU	M24603	100	break point cluster gene (BCR)	HU	M24603	100
651	123	68A6	voltage-gated sodium channel	HU	M94055	75	voltage-gated sodium channel	HU	A38195	100
652	181	68A9	breakpoint cluster region (BCR)	HU	Y00661	100	break point cluster gene (BCR)	HU	A28765	100
653	92	68A10	endogenous retrovirus	HU	M14123	91	retrovirus env polyprotein	HU	P10267	83
672	71	70A5	neonatal glycine receptor*	RA	X57281	89				
687	148	71F11	protein phosphatase T	RA	X77237	91	protein phosphatase T	RA	U12203	100
696	199	73E6	anonymous gene	HU	L18972	100	anonymous gene	HU	L18972	100
699	86	74A9	S-lac lectin L-14-II (LGALS2)	HU	M87857	100				
704	96	75E4	ye98f07.r1 Homo sapiens cDNA	HU	R07694	100				
707	239	75E8	EST07064	HU	T09171	100				
712	200	75F5	LZTR-1 mRNA	HU	D38496	100	ORF 882 gene product	Sc	X87941	60
713	94	75F8	Ig lambda chain variable*	HU	X56178	88	Ig lambda chain variable*	HU	S13726	83
733	222	78F6	yc27d07.r1 Homo sapiens cDNA	HU	T67791	88	hypothetical protein YKL207w	Sc	P36039	68
737	129	79A12	glutathione S-transferase T1	HU	X79389	98	glutathione S-transferase T1	HU	X79389	97
739	177	79B5	glutathione S-transferase T1	HU	X79389	98	glutathione S-transferase T1	HU	X79389	98
744	79	80F3	beta-adrenergic receptor kinase 2	HU	X69117	100	beta-adrenergic receptor kinase 2	HU	P35626	100
747	103	81A9	STS UT5309	HU	L30836	85				
751	179	81B9	partial cDNA sequence; clone c-zqe06	HU	Z41320	100				
753	143	82E1	breakpoint cluster region (BCR)	HU	M15025	100	breakpoint cluster region (BCR)	HU	A26664	100
757	98	82E12	N-type calcium channel alpha1	MU	U04999	81	voltage-dependent sodium channel	HU	A38195	88
760	158	82F6	cytochrome P450 IID6 (CYP2D6)*	HU	M33388	100				
766	166	83B5	partial cDNA sequence; clone c-34a04	HU	F11966	99	Ufd1p	Sc	U22153	60
768	53	84E 2	partial cDNA sequence; clone c-2zd01	HU	F11617	96				
799	196	87B4	yd71e10.r1 Homo sapiens cDNA	Hu	T79705	99				
802	135	88E1	breakpoint cluster region (BCR)	HU	X02596	97	breakpoint cluster region (BCR)	HU	A28765	97
809	109	89A4	E46	MU	X61506	80	E46	MU	P28658	94
814	132	89B8	NADH-cytochrome b5 reductase (DIA1)	HU	M28706	100	NADH-cytochrome b5 reductase (DIA1)	HU	M16461	100
821	172	91A7	Tbx2	MU	U15566	72	Tbx1	MU	U15565	97
826	125	92E10	ORF	HU	D29677	87	ORF	HU	D29677	92
833	89	93A4	TUP1-like enhancer of split (TUPLE1)	HU	X75296	100	TUP1-like enhancer of split (TUPLE1)	HU	X75296	100
838	280	95A4	cadherin-13*	HU	L34058	98				
840	161	93B3	TUP1-like enhancer of split (TUPLE1)	HU	X75296	100	TUP1-like enhancer of split (TUPLE1)	HU	X75296	100
848	136	94E11	beta adaptin (ADTB1)	HU	L13939	100	beta adaptin (ADTB1)	HU	L13939	100
859	106	95B7	Ig kappa light chain variable	HU	X72444	84	Ig kappa light chain variable	HU	Z27177	96
861	100	96E3	NADH-cytochrome b5 reductase (DIA1)	HU	M28712	100	NADH-cytochrome b5 reductase (DIA1)	HU	B26616	100
872	151	97A2	ye89b04.r1 Homo sapiens cDNA	HU	R06059	100				
879	120	97B5	breakpoint cluster region (BCR)	HU	Y00661	100	break point cluster gene (BCR)	HU	A28765	100
880	103	97B9	synapsin 2b	RA	M27926	80	synapsin 2b	RA	D30411	94
883	116	98E4	ym51d01.r1 Homo sapiens cDNA	HU	H23039	100				
890	188	98F10	mRNA for randomly sequenced product	HU	D21260	80	clathrin heavy chain	RA	P11442	96
894	121	99A5	leukemia inhibitory factor (LIF)	HU	M63420	100				
899	83	18A4	IL-2R-beta (IL2RB)	HU	M26062	100	IL-2R-beta (IL2RB)	HU	P14784	100
903	199	18A11	LZTR-1 mRNA	HU	D38496	100				
913	182	25F12	breakpoint cluster region (BCR)	HU	U07000	99	breakpoint cluster region (BCR)	HU	M24603	98
914	149	37A1	IL-2R-beta (IL2RB)	HU	M26062	100	IL-2R-beta (IL2RB)	HU	P14784	100
917	60	37A4	Homo sapiens cDNA clone G3622	HU	R58430	100				
936	116	76E4	partial cDNA sequence; clone c-zpd06	HU	Z45600	100				
Strong similarity										
4	177	10G5	cDNA clone b12056	HU	T18862	76				
25	138	11D7	EST03946	HU	T06057	82				
41	147	13A5					hypothetical protein 332	Cp	JQ0367	65
44	215	13A11					malonyl coenzyme A-acyl carrier	Ec	P25715	61
66	157	14B6	peregrin	HU	M91585	67	peregrin	HU	M91585	80
113	40	1B1	cDNA clone hbc787	HU	T11161	85				
114	164	1B2					F47A4.5	Ce	Z49888	73
118	256	20A5	oxysterol-binding protein	RB	J05056	71	oxysterol-binding protein	HU	P22059	83
135	114	21F1					DNA binding protein (CDC10 partner)	Sp	Z32838	71
143	246	22A1	neuronal pentraxin II	HU	U29193	73	apexin	GP	U13236	84
148	108	22A10					orf gene product	Sc	Z48149	64
151	92	23E1					growth factor receptor-bound protein 3	HU	L29511	84
201	205	30B2					GTPase-activating protein rhoGAP	HU	S34296	76
216	77	32A10	Ig lambda light chain	HU	X51754	80				
258	165	3A4					alpha-actinin	HU	P35609	63
262	189	3B2	CD10 neutral endopeptidase	MU	M81591	70	endothelin-converting-enzyme 1	HU	Z35307	81
266	166	40A1					fringe protein	Dm	L35770	54
275	96	40B8	HUMXQG Human mRNA	HU	D16473	75				
278	73	41E2	hb protein	MU	X81634	81	hb gene product	MU	X81634	90
311	101	44A1					semaphorin C	MU	X85992	64
344	114	47B10					transforming protein vav	HU	S05382	71
350	177	48F3	theta glutathione-S-transferase	CH	U13676	66	theta glutathione-S-transferase	CH	U13676	74
371	116	50A7	collagen alpha 3(VI) chain	HU	X52022	84	collagen alpha 3(VI) chain	HU	S13679	82
461	59	51F12	yi75a12.s1 Homo sapiens cDNA	HU	R77368	71				
526	136	57E4	LIMK	CH	D26310	86	LIM motif-containing protein kinase	CH	JP0079	91
536	126	58A3	partial cDNA sequence; clone c-2zg03	HU	F09301	69				

Table 1. *(continued)*

537	190	58A6	phosphatase 1M 110 kda reg. subunit	RA	S74907	65	130 kDa myosin-binding subunit	CH	D37986	86
556	51	60A4	expressed sequence tag ml650	MU	L12133	88				
576	137	61F7	cDNA clone FB10G10	HU	T02824	83				
618	166	65F6	yd30c11.s1 Homo sapiens cDNA	HU	T81609	66				
627	201	66A8	vasoactive intestinal peptide receptor	HU	L20295	66	vasoactive intestinal peptide receptor	RA	P30083	63
665	91	69F1					glycine dehydrogenase	HU	P23378	75
685	186	71F3	D. melanogaster STS	Dm	Z32012	64	alkaline extracellular protease	Yl	P09379	65
742	94	80E10					GTPase-activating protein rhoGAP	HU	S34296	84
787	180	86E1					polyadenylate-binding protein	Dm	S30877	53
804	198	88E4	EST06944	HU	T09052	75				
806	152	88F7	glutathione S-transferase	RA	X67654	73	glutathione S-transferase	RA	Q01579	76
825	84	92E9	yg96d09.s1 Homo sapiens cDNA	HU	R59138	80				
832	213	93A3					hypothetical 58.3 Kd protein F42H10.7	Ce	P34420	59
853	172	95A2					cell division cycle gene CDC25	Sc	X03579	69
869	129	96F5	growth-arrest GAS 6 mRNA	MU	X59846	63	fibrillin-2	HU	P35556	70
926	119	37B9	yi98b11.s1 Homo sapiens cDNA clone	HU	R81941	73				

The criteria for categorizing similarities were based on BLASTN or BLASTX results (Altschul et al. 1990) and were as follows: Identity, near identity, or homolog, nucleic acid, or protein similarities ≥85% and *P* value $\leq 10^{-5}$; strong similarity, nucleic acid, or protein similarity ≥50% and *P* value $\leq 10^{-3}$. Asterisks denote sequence similarity to the complementary strand of the data base "hit." Species names: (BO) bovine; (Ce) *Caenorhabditis elegans;* (CH) chicken; (Cp) *Clostridium perfringens;* (Dd) *Dictyostelium discoideum;* (Dm) *Drosophila melanogaster;* (Ec) *Escherichia coli;* (GP) *Cavia porcellus* (guinea pig); (HU) human; (MU) mouse; (PI) pig; (RA) rat; (RB) rabbit; (Sc) *Saccahromyces cerevesiae;* (Sp) *Schizosaccharomyces pombe;* (Yl) *Yarrowia lipolytica.*

plementary strand of known genes. The most notable of these were the matches of sequences 285 and 760, which were identical to the complementary strands of Ewing sarcoma and cytochrome P450 IID6 gene sequences, respectively. In both cases, the match was found near the 3′ end of the mRNA sequences of these genes, and the sequences flanking the aligned regions closely match consensus splice sites. Whether these sequences represent artifacts or genes encoded on the overlapping DNA strand opposite to the known genes remains unclear.

The remaining 98 sequences represent human homologs of genes from other species, members of gene families, or genes sharing strong similarities with known genes. Among the more interesting sequences with similarities are those that represent novel human genes that are related to others with known or putative functions. For example, the predicted amino acid sequence of exon 637 is highly similar to part of a common domain, termed the chromodomain, found in genes whose products associate with heterochromatin (James and Elgin 1986; Paro and Hogness 1991; Singh et al. 1991; Delmas et al. 1993). The best studied of these genes are *Drosophila* heterochromatin-associated protein, HP1, and Polycomb. Both of these genes have been shown to control, by repression, developmental regulators such as homeotic genes. The chromodomain appears to be essential for assembly of these proteins into chromatin as part of a multiple protein complex, as mutations or deletions in this domain in the Polycomb protein abolish its ability to associate with heterochromatin (Messmer et al. 1992). Thus, the sequence represented by exon 637 may represent a novel regulator of homeotic function in human development. Figure 1 is an amino acid alignment of exon 637 with the chromodomains of other pro-

Table 2. Summary of sequence data base comparisons

Category	Number of sequences	Percent of total
Identity to known human genes/ESTs	101	14.2
Near identity or strong similarity	98	13.8
Weak or no similarity	404	57.0
Repetitive sequence/artifact	106	15.0
Total unique sequences	709	100.0

Sequences have been placed into each category based on the results of BLASTN and BLASTX comparisons (Altschul et al. 1990). Criteria for similarity categories are described in the legend to Table 1.

Table 3. Chromosome 22 gene sequences identified

Locus name	Gene name	Chromosomal position	Number of matching sequences
ADRBK2	β-adrenergic receptor kinase	q11	1
ADTB1	β-adaptin	q12	1
BCR	breakpoint cluster region	q11.2	7
COMT	catechol-*O*-methyltransferase	q11.2	1
CSF2RB	GM–CSF receptor, β-chain	q13	2
CYP2D6[a]	cytochrome P450 IID6	q13	1
DIA1	NADH–cytochrome b5 reductase	q13.3	2
EWSR[a]	Ewing sarcoma	q12.1–q12.2	2
FBLN1	fibulin-1 (isoforms A, B, C, D)	q13.2–q13.3	4
GGT	γ-glutamyl transpeptidase	q11.2	1
GGT-rel	GGT-related	q11.2	1
IGLL8	immunoglobulin λ-like	q11.2	1
IGLV	immunoglobulin light-chain variable	q11.2	1
IL2RB	interleukin 2 receptor, β-subunit	q13	2
LGALS2	S-lac lectin L-14-II	q12.2–q13	1
LIF	leukemia inhibotoory factory	q12	1
MYH9	cellular myosin heavy chain	q12.3–q13.1	1
NF2	merlin, neurofibromatosis type 2	q12.1–q12.2	2
PDGFB	c-*sis,* platelet-derived growth factor	q13	1
PPAR	peroxisome proliferation activated receptor	q12–q13.1	1
TUPLE1	TUP1-like enhancer of split	q11	5
	anonymous gene	q12	1
	p300 transcription adaptor	q13.2	1
	LZTR-1	q11	2
	AR1	?	2

The previously localized genes identified by data base comparisons and the number of different sequences matching them are listed.
[a] One of the sequences matching EWSR1 and the matching sequence for CYP2D6 were identical to the mRNA complementary strands of these genes.

teins. Exon 637, however, does not contain the complete chromodomain, but begins several amino acids downstream and continues beyond the carboxyl end of the motif. Interestingly, the genomic structure of the Polycomb locus of *Drosophila* has been determined, and the 5′ end of exon 637 is at the precise location of an intron–exon boundary within this gene (Paro and Hogness 1991). This suggests that some phylogenetic conservation of genomic structure exists for the 637 gene or that the 637 gene may represent the human homolog of Polycomb.

In addition to exon 637, several of the sequences appear to be closely related to genes involved in growth regulatory, developmental, or cell type-specific-processes. Isolation of these types of genes, many of which are likely to be representative of low-abundance or tissue-specific mRNA species exemplifies an advantage of the exon amplification approach: expression-independent gene identification. The overwhelming majority of human genes are expressed at low levels, producing low-abundance mRNA (Hastie and Bishop 1976). Many other approaches that require significant levels of gene expression, or knowledge of tissue specificity, may fail to identify such genes with any efficiency. This includes most large-scale random cDNA sequencing strategies (Adams et al. 1991, 1992; Khan et al. 1992; Okubo et al. 1992), which are biased toward identification of mRNAs that are highly expressed in the tissue (or cells) from which the library was generated, as well as approaches using RNA derived from monochromsomal- or region-specific human–rodent hybrid cell lines (Liu et al. 1989; Corbo et al. 1990; Jones

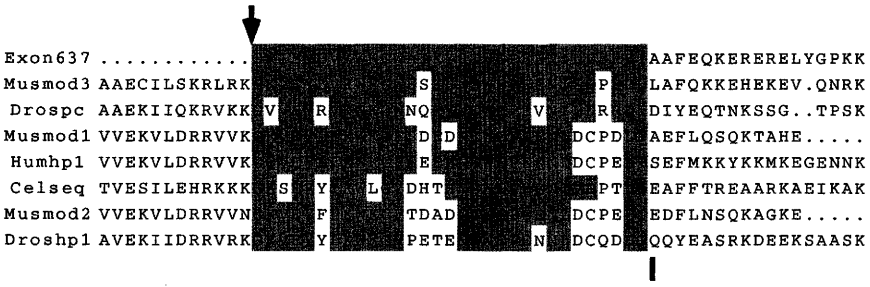


Figure 1 Alignment of the predicted peptide sequence of 637 with other chromodomain-containing proteins. Identities or conservative amino acid differences between 637 and each protein are shaded. The arrow indicates the location of an intron in the *Drosophila melanogaster* Polycomb gene. (Musmod3) Mouse modifier-3 (accession no. P30658); (Drospc) *Drosophila melanogaster* Polycomb (accession no. P26017); (Musmod1) mouse modifier-1 (accession no. P23197); (Humhp1) human HP1 homolog (accession no. S62077); (Celseq) *Caenorhabditis elegans*, hypothetical 33.8-kD protein (accession no. P34618); (Musmod2) mouse modifier-2 (accession no. P23198); (Droshp1) *Drosophila melanogaster* heterochromatin protein HP1 (accession no. P05205).

et al. 1992). Direct or cDNA selection procedures (Parimoo et al. 1991; Lovett et al. 1991) have been designed to minimize this problem of representation by enriching for rare mRNAs and complement exon amplification in that they can enrich for these low abundance species, if they are present. These approaches are also adaptable for en masse, region-specific gene identification, as Del Mastro et al. (1995) demonstrate using human chromosome 5 as a target.

One hundred six of the 709 sequences (~15%) were artifacts or similar to repetitive elements. The artifacts were derived from a number of sources but originate primarily from pSPL3 and the pLawrist16 cosmid vector. The higher prevalence of these clones as artifacts in these assays, as compared to assays of single cosmid clones, may be attributable to the vast molar excess of these sequences relative to the genomic sequences targeted for exon isolation. It should be noted that sequences from pSPL3 or pLawrist comprise ~20% (10% each) of all clones in the library, based on sequencing of ~2300 clones. These can be eliminated readily by hybridization detection, thereby significantly reducing the effort required for sequencing of similar libraries.

Localization of Exons to Chromosome 22

The starting genomic DNA for these experiments was derived from flow-sorted chromosomes derived from a human–hamster hybrid that contains preferentially human chromosome 22, but also retains chromosomes 9 and Y at a low frequency. Thus, the possibility exists that some of the exons originate from nonchromosome 22 genomic DNA, including hamster. This estimate has been confirmed by mapping these sequences to Southern blots of human, hamster, and GM10888 (monochromosome 22 human–hamster hybrid) DNAs (data not shown). Of 21 randomly chosen exons that hybridized to the blots, 17 (81%) mapped to chromsome 22, 3 were of hamster origin, and 1 was human, but apparently not originating from chromosome 22. These numbers are consistent with the estimated nonchomosome 22 content of the starting cosmid library. The percentage of chromosome 22-specific sequences is likely to be higher, as we excluded from our analysis exons from previously mapped genes.

DISCUSSION

The collection of clones described above represents one of the first large-scale, chromosome-specific isolations of human gene sequences, and will serve as an inroad to the development of an integrated physical and transcription map of human chromosome 22. It is likely to be an invaluable tool for creating the high resolution maps that are needed for sequencing of the human genome. Extrapolation of our results leads to a prediction that ~1500 sequences will be generated after identification of all unique sequences in this collection of clones (of which >1200 will be nonrepetitive and nonartifact), and an estimation that nearly half of these sequences have been produced to date. In this study 24 of 59 (41%) of the fully sequenced genes (nonpseudogenes) currently known to map to chromosome 22 were identified, suggesting that a similar percentage of all genes on this chromosome are represented by the sequences that have been generated thus far. Therefore, it is anticipated that exons from as many as 80% of the genes on this chromosome

will be represented in this library after completion of library sequencing. To further increase the representation of genes in this library and eliminate much of the bias in gene identification, subsequent isolation of exons will be performed on cosmids that stochastically failed to produce exons in the initial library construction.

Several aspects of this approach make it ideal for construction of detailed physical/transcription maps. First, the use of genomic DNA from a specific human chromosome allows positional information to be associated provisionally to each exon, circumventing the need to localize cloned gene fragments that have been isolated and sequenced as random cDNA, and provides a more direct approach to saturate specific regions of the genome with such sequences. It should be noted, however, that sources such as flow-sorted chromosomes, or libraries constructed from them, frequently originate from human–rodent somatic cell hybrids; a small fraction of genomic DNA, hence exons, may originate from the rodent parent or from other human chromosomes present within the hybrid. The chromosome 22 cosmid library used in this study was derived from a hybrid cell line that also retains human chromosomes 9 and Y at a low frequency. We have estimated that 10%–20% of the exon library contains sequences from genomic DNA not originating from human chromosome 22, the majority of which is from hamster. Thus, although the exon library is highly enriched for chromosome 22 gene sequences, it is not pure, and the sequences are annotated as such. It should be noted, however, that no exons from known genes that map to any other human chromosome were identified to date in these studies, whereas 48 exons from chromosome 22-specific genes were isolated. In addition, the recent availability of high-quality monochromosomal hybrids for most human chromosomes, coupled with improved flow-sorting will reduce dramatically the problem for future exon library constructions.

Second, the exons produced by this procedure are consummate multi-purpose mapping reagents. Because the vast majority of exons are single copy sequences, they can be used as hybridization probes in filter-based mapping procedures. Moreover, we have found that exon sequences are easily converted to sequence-tagged sites (STS) for use in PCR-based mapping schemes (Green and Olson 1990). With an average spacing of 60–70 kb, the chromosome 22 exons identified thus far could be used to complete the Human Genome Initiative's goal of one STS per 100 kb for each chromosome. The conversion of exons to cDNAs would then provide both ordering and orientation across groups of yeast artificial chromosomes (YACs) or cosmids as well as confirm any existing contig information. Also, the cross-species conservation of many exons allows for effective comparative mapping of genes and for direct comparison of emerging physical maps in humans, mice, and other model genomes.

Third, because exon amplification is not dependent on the level or pattern of expression of the gene that is isolated, representational biases inherent in tissues or cells from which cDNA libraries are constructed, are eliminated. Exons can be used as a DNA sequence source for determining the specific expression pattern of the gene from which it originated by using quantitative assays of RNA expression, such as Northern blotting, S1 nuclease, or RNase protection, in situ hybridization, or by using PCR to detect the presence of exon sequences in a cDNA library (Church et al. 1993; Munroe et al. 1995). The resulting information allows for effective screening of appropriate cDNA libraries to saturate quickly a given genomic region with genes. This property of the technique has already been applied successfully to positional cloning efforts in Huntington's disease and neurofibromatosis 2 (Huntington's Disease Collaborative Research Group 1993; Trofatter et al. 1993), resulting in identification of the disease genes by isolation of 28 and 8 of their exons, respectively, as well as successful effort to identify several other human and mouse disease genes (Vidal et al. 1993; Vulpe et al. 1993; Walker et al. 1993; Cachon-Gonzalez et al. 1994; Hästbacka et al. 1994).

The large-scale exon isolation approach that we have applied here is currently being transferred to several other human chromosomes. Our data suggest that this method could identify segments from the majority of human genes before to the generation of the human genome's sequence. Moreover, the strategy would help to achieve this goal by facilitating the necessary construction and comparison of fine structure physical maps while simultaneously integrating them into transcription maps of greater utility to a wide range of researchers in genetics and biology. Continued application of this strategy represents a most effective and cost-efficient means of substantiating the most prominent rationale for pursuing the Human Genome Initiative, cre-

ation of the infrastructure needed to support the rapid and efficient discovery of genes causing human disease.

METHODS

Exon Library Construction

Exon amplification was performed as described (Church et al. 1994), with some modification. Cosmid-containing clones were propagated in 96-well microtiter plates, pooled, and cosmid DNA purified using the alkaline lysis method. Before propagation, cosmids containing ribosomal gene DNA (rDNA) were identified by hybridization and removed from each plate. This was done to insure against overrepresentation of chromosome 22 rDNA sequences in the amplified and cloned products. Shotgun cloning into pSPL3, transfections, and RNA isolations were performed as described (Church et al. 1994). Initially, RNA-based–PCR amplification (RT–PCR) and cloning of the resulting products was performed as described (Church et al. 1994), but was replaced by ligation-independent cloning using uracil DNA glycosylase (UDG; Rashtchian et al. 1991). This entailed the replacement of oligodeoxynucleotide primers SD2 and SA4 in the second PCR amplification, with SDDU and SADU. The sequences of these primers are as follows:

SDDU: 5′-AUAAGCUUGAUCUCACAAGCTG CACGCTCTAG-3′,

SADU: 5′-UUCGAGUAGUACUTTCTATTCCT TCGGGCCTGT-3′.

Complementary primers were also designed for the cloning vector pBluescript IIKS + (Stratagene), surrounding the *Eco*RV site; these are as follows:

BSDU: 5′-GAUCAAGCUUAUCGATACCGT CGACC-3′,

BSAU: 5′-AGUACUACUCGAAUTCCTGCA GCC-3′.

Ten nanograms of *Eco*RV-digested pBluescript IIKS + was amplified with BSDU and BSAU. Fifty nanograms of the amplified, linearized plasmid was mixed with 50–100 ng of RT–PCR product, and the mixture was digested with 1 unit of UDG (GIBCO-BRL) at 37°C in a 10-μl volume of 1 × PCR buffer. The digested and annealed products were immediately transformed into *Escherichia coli* DH5α host. UDG cloning streamlined the procedure and completely eliminated a significant frequency of clone chimerism. Clones from each pool were picked, propagated, frozen, and stored in 96-well microtiter plates. Sequencing was performed using the method of Sanger et al. (1977). Sequences were automatically read using a Millipore Bioimage DNA sequence film reader operating on a Sun Sparc Station. Sequence data base comparisons were performed using the BLAST network service of the National Center for Biotechnology Information (Altschul et al. 1990). The sequences have been deposited in Genbank with the following accession numbers: H55062–H55737.

ACKNOWLEDGMENTS

We thank Drs. M. Duyao, M.K. McCormick, D.J. Munroe, and M. Lovett for helpful comments, discussion, and critical reading of the manuscript. We also thank the Lawrence Livermore National Laboratory for providing the flow-sorted chromosome 22 cosmid library. This work was supported by grants HG00672 (A.J.B.) and HG00169 (J.F.G.) from the National Institutes of Health.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno, A.R. Kerlavage, W.R. McCombie, and J.C. Venter. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651–1656.

Adams, M.D., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M Kelley, T.R. Utterback, J.W. Nagle, C. Fields, and J.C. Venter. 1992. Sequence identification of 2,375 human brain genes. *Nature* **355:** 632–634.

Altschul, S.F., W. Gish, W. Miller, E. Myers, and D.J. Lippman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Auch, D. and M. Reth. 1990. Exon trap cloning: Using PCR to rapidly detect and clone exons from genomic DNA fragments. *Nucleic Acids Res.* **18:** 6743–6744.

Bischoff, F.R., H. Krebber, T. Kempf, I. Hermes, and H. Ponstingl. 1995. Human RanGTPase-activating protein RanGAP1 is a homologue of yeast Rna1p involved in mRNA processing and transport. *Proc. Natl. Acad. Sci.* **92:** 1749–1753.

Buckler, A.J., D.D. Chang, S.L. Graw, J.D. Brook, D.A. Haber, P.A. Sharp, and D.E. Housman. 1991. Exon amplification: A strategy to isolate mammalian genes based on RNA splicing. *Proc. Natl. Acad. Sci.* **88:** 4005–4009.

Cachon-Gonzalez, M.B., S. Fenner, J.M. Coffin, C. Moran, S. Best., and J.P. Stoye. 1994. Structure and expression of the hairless gene of mice. *Proc. Natl. Acad. Sci.* **91:** 7717–7721.

Church, D.M., A.C. Rogers, S.L. Graw, D.E. Housman, J.F. Gusella, and A.J. Buckler. 1993. Identification of human chromosome 9 specific genes using exon amplification. *Human Mol. Genet.* **2:** 1915–1920.

Church, D.M., C.J. Stotler, J.L. Rutter, J.R. Murrell, J.A. Trofatter, and A.J. Buckler. 1994. Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nature Genet.* **6:** 98–105.

Cohen, D., I. Chumakov, and J. Weissenbach. 1993. A first-generation physical map of the human genome. *Nature* **366:** 698–701.

Corbo, L., J.A. Maley, D.L. Nelson, and C.T. Caskey. 1990. Direct cloning of human transcripts with HnRNA from hybrid cell lines. *Science* **249:** 652–655.

Delmas, V., D.G. Stokes, and R.P. Perry. 1993. A mammalian DNA-binding protein that contains a chromodomain and an SNF2/SWI2-like helicase domain. *Proc. Natl. Acad. Sci.* **90:** 2414–2418.

Del Mastro, R., L. Wang, A.D. Simmons, T.D. Gallardo, G.A. Clines, J.A. Ashley, C.J. Hilliard, J.J. Wasmuth, J.D. McPherson, and M. Lovett. 1995. Human chromosome-specific cDNA libraries: New tools for gene identification and genome annotation. *Genome Res.* **5:** 185–194.

Duyk, G.M., S. Kim, R.M. Myers, and D.R. Cox. 1990. Exon trapping: A genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. *Proc. Natl. Acad. Sci.* **87:** 8995–8999.

Green, E. D. and M.V. Olson. 1990. Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. *Proc. Natl. Acad. Sci.* **87:** 1213–1217.

Hästbacka, J., A. de la Chapelle, M. Mahtani, G. Clines, M.P. Reeve, M. Daly, B. Hamilton, K. Kusumi, B. Trivedi, A. Weaver, M. Lovett, A. Buckler, I. Kaitila, and E.S. Lander. 1994. The diastrophic dysplasia gene encodes a novel sulfate transporter: Positional cloning by fine-structure linkage disequilibrium mapping. *Cell* **78:** 1073–1087.

Hastie, N.D. and J.O. Bishop. 1976. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **9:** 761–774.

The Huntington's Disease Collaborative Research Group. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72:** 971–983.

James, T.C. and S.C.R. Elgin. 1986. Identification of a nonhistone chromosomal protein associated with heterochromatin in Drosophila melanogaster and its gene. *Mol. Cell. Biol.* **6:** 3862–3872.

Jones, K.W., M. Chevrette, M.H. Shapero, and R.E.K. Fournier. 1992. Generation of region and species-specific expressed gene probes from somatic cell hybrids. *Nature Genet.* **1:** 278–283.

Khan, A.S., A.S. Wilcox, M.H. Polymeropoulos, J.A. Hopkins, T.J. Stevens, M. Robinson, A.K. Orpana, and J.M. Sikela. 1992. Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nature Genet.* **2:** 180–185.

Lehmeier, T., V.A. Raker, H. Hermann, and R. Luehrmann. 1994. cDNA cloning of the Sm proteins D2 and D3 from human small nuclear ribonucleoproteins: Evidence for a direct D1-D2 interaction. *Proc. Natl. Acad. Sci.* **91:** 12317–12321.

Lichter, P., S.A. Ledbetter, D.H. Ledbetter, and D.C. Ward. 1990. Fluorescence in situ hybridization with Alu and L1 polymerase chain reaction probes for rapid characterization of human chromosomes in hybrid cell lines. *Proc. Natl. Acad. Sci.* **87:** 6634–6638.

Liu, P., R. Legerski, and M.J. Siciliano. 1989. Isolation of human transcribed sequences from human–rodent somatic cell hybrids. *Science* **246:** 813–815.

Lovett, M., J. Kere, and L.M. Hinton. 1991. Direct selection: A method for the isolation of cDNAs encoded by large genomic fragments. *Proc. Natl. Acad Sci* **88:** 9628–9632.

Messmer, S., A. Franke, and R. Paro. 1992. Analysis of the functional role of the Polycomb chromo domain in Drosophila melanogaster.*Genes & Dev.* **6:** 1241–1254.

Munroe, D.J., R. Loebbert, E. Bric, T. Whitton, D. Prawitt, D. Vu, A. Buckler, A. Winterpracht, B. Zabel, and D.E. Housman. 1995. Systematic screening of an arrayed cDNA library by PCR. *Proc. Natl. Acad. Sci.* **92:** 2209–2213.

Okubo, K., N. Hori, R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima, and K. Matsubara. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* **2:** 173–179.

Paro, R. and D.S. Hogness. 1991. The Polycomb protein shares a homologous domain with a heterochromatin-associated protein of Drosophila. *Proc. Natl. Acad. Sci.* **88:** 263–267.

Parimoo, S., S.R. Patanjali, H. Shulka, D.D. Chaplin, and S.M. Weissman. 1991. cDNA selection: Efficient PCR approach for the selection of cDNAs in large genomic DNA fragments. *Proc. Natl. Acad. Sci.* **88:** 9623–9627.

Pemble, S., K.R. Schroeder, S.R. Spencer, D.J. Meyer, E. Hallier, H.M. Bolt, B. Ketterer, and J.B. Taylor. 1994. Human glutathione S-transferase theta (GSTT1): cDNA cloning and the characterization of a genetic polymorphism. *Biochem. J.* **300:** 271–276.

Rashtchian, A., G.W. Buchman, D.M. Schuster, and M. Berninger. 1991. Uracil DNA glycosylase-mediated cloning of polymerase chain reaction-amplified DNA: Application to genomic and cDNA cloning. *Anal. Biochem.* **206:** 91–97.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci.* **74:** 5463–5467.

Singh, P.B., J.R. Miller, J. Pearce, R. Kothary, R.D. Burton, R. Paro, T.C. James, and S.J. Gaunt. 1991. A. sequence motif found in a Drosophila heterochromatin protein is

conserved in animals and plants. *Nucleic Acids Res.* **19:** 789–794.

Tanihara, H., K. Sano, R.L. Heimark, T. St.John, and S. Suzuki. 1994. Cloning of five cadherins clarifies characteristic features of cadherin extracellular domain and provides further evidence for two structurally different types of cadherin. *Cell Adhesion Commun.* **2:** 15–26.

Trofatter, J.A., M.M. MacCollin, J.L. Rutter, J.R. Murrell, M.P. Duyao, D.M. Parry, R. Eldridge, N. Kley, A.G. Menon, K. Pulaski, V. Haase, C. Ambrose, D. Munroe, C. Bove, J.L. Haines , R.L. Martuza, M.E. MacDonald, B.R. Seizinger, M.P. Short, A.J. Buckler, and J.F. Gusella. 1993. A novel moesin-, ezrin-, radixin-like gene is a candidate for the neurofibromatosis 2 tumor suppressor. *Cell* **72:** 791–800.

Vidal, S.M., D. Malo, K. Vogan, E. Skamene, and P. Gros. 1993. Natural resistance to infection with intracellular parasites: Isolation of a candidate for Bcg. *Cell* **73:** 469–485.

Vulpe, C., B. Levinson, S. Whitney, S. Packman, and J. Gitschier. 1993. Isolation of a candidate gene for Menkes disease and evidence that it encodes a copper-transporting ATPase. *Nature Genet.* **3:** 7–13.

Walker, A.P., F. Muscatelli, and A.P. Monaco. 1993. Isolation of the human Xp21 glycerol kinase gene by positional cloning. *Hum. Mol. Genet.* **2:** 107–114.

Wong, K. and L. Cantley. 1994. Cloning and characterization of a human phosphatidylinositol 4-kinase. *J. Biol. Chem.* **269:** 28878–28884.

Received June 15, 1995; accepted in revised form September 22, 1995.