Cloning and Functional Analysis of cDNAs with Open Reading Frames for 300 Previously Undefined Genes Expressed in CD34+ Hematopoietic Stem/Progenitor Cells

Qing-Hua Zhang,[1,4] Min Ye,[1,4] Xin-Yan Wu,[1,4] Shuang-Xi Ren,[2,4] Meng Zhao,[1] Chun-Jun Zhao,[1] Gang Fu,[2] Yu Shen,[1] Hui-Yong Fan,[1] Gang Lu,[2] Ming Zhong,[2] Xiang-Ru Xu,[2] Ze-Guang Han,[2] Ji-Wang Zhang,[1] Jiong Tao,[1] Qiu-Hua Huang,[1] Jun Zhou,[1] Geng-Xi Hu,[3] Jian Gu,[1,2] Sai-Juan Chen,[1] and Zhu Chen[1,2,5]

[1]Shanghai Institute of Hematology (SIH), Rui Jin Hospital affiliated with Shanghai Second Medical University, Shanghai 200025, China; [2]Chinese National Human Genome Center (CHGC) at Shanghai, Shanghai 201203, China; [3]Institute of Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China

Three hundred cDNAs containing putatively entire open reading frames (ORFs) for previously undefined genes were obtained from CD34+ hematopoietic stem/progenitor cells (HSPCs), based on EST cataloging, clone sequencing, in silico cloning, and rapid amplification of cDNA ends (RACE). The cDNA sizes ranged from 360 to 3496 bp and their ORFs coded for peptides of 58–752 amino acids. Public database search indicated that 225 cDNAs exhibited sequence similarities to genes identified across a variety of species. Homology analysis led to the recognition of 50 basic structural motifs/domains among these cDNAs. Genomic exon–intron organization could be established in 243 genes by integration of cDNA data with genome sequence information. Interestingly, a new gene named as HSPC070 on 3p was found to share a sequence of 105bp in 3′ UTR with *RAF* gene in reversed transcription orientation. Chromosomal localizations were obtained using electronic mapping for 192 genes and with radiation hybrid (RH) for 38 genes. Macroarray technique was applied to screen the gene expression patterns in five hematopoietic cell lines (NB4, HL60, U937, K562, and Jurkat) and a number of genes with differential expression were found. The resource work has provided a wide range of information useful not only for expression genomics and annotation of genomic DNA sequence, but also for further research on the function of genes involved in hematopoietic development and differentiation.

[The sequence data described in this paper have been submitted to the GenBank data library under the accession nos. listed in Table 1, pp 1548–1552.]

The Human Genome Project now is at a historic turning point, from genomic DNA sequencing to functional genomics. According to the announcement from both public domain and private sector sequencing efforts, a "working draft" of the human genome sequence was just obtained, and the completion of the sequence will be achieved before the end of 2001 (Collins et al. 1998; Venter et al. 1998; Marshall 1999, 2000). The gene discovery and understanding of genetic information will require annotation of the sequence data using bioinformatic tools (Burge and Karlin 1997). Meanwhile, cloning of full-length cDNA has been listed as one of the major tasks of the current phase of genomic science (Collins et al. 1998). The integration of cDNA sequences with the genomic ones will greatly ease the identification of transcriptional units, as well as their mRNA levels and specificities in cells/tissues as a result of a fine regulation of the transcriptional expression at genomic level (Dunham et al. 1999; Hattori et al. 2000). Moreover, the cDNA project links directly to protein structural biology and exerts significant impact on the medical genetics and biotechnology/pharmaceutical industries.

Hematopoietic stem/progenitor cells (HSPCs) possess important roles for the physiological and pathological hematopoiesis, one of the essential areas in biomedicine, and the molecular basis of hematopoiesis remains to be better understood (Morrison et al. 1995, 1997). Over the last 3 years, we have been undertaking to catalog the expressed sequence tags (ESTs) from cDNA libraries of CD34+ HSPC populations from both umbilical cord blood (Mao et al. 1998) and adult bone marrow (Gu et al. 2000). This approach turned out to be very successful in terms of both gene expression profiling and discovery of novel genes in an efficient

[4]These authors contributed equally to this work.
[5]Corresponding author.
E-MAIL zchen@ms.stn.sh.cn; FAX 86-21-6474 3206.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.140200.

10:1546–1560 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

way. More recently, we have been extending this work to the cloning and sequencing of full-length cDNAs for previously undefined genes and to investigate their functions.

In this work, we report on the characterization of structural/functional features, chromosomal localization, and transcriptional expression patterns in different hematopoietic cell lines of 300 cDNAs with putatively entire open reading frames (ORFs) isolated from CD34+ cells. We also tried to integrate these data with the genomic sequence information and to propose some strategies to deal with the major challenges in expression genomics facing the completion of the human genomic sequences in the coming 1 or 2 years.

RESULTS

Primary Gene Expression Profiles of CD34+ HSPCs

RT/PCR-based Capfinder cDNA libraries were constructed using mRNA from highly purified CD34+ HSPCs of cord blood and adult bone marrow, using methods described previously (Mao et al. 1998). In total, 1×10^6 recombinant clones were obtained from CD34+ cell library of cord blood origin (CB) and 0.5×10^6 clones were acquired from that of bone marrow (BM). The average size of inserts in both libraries was 1.2 kb. Among 9866 and 4142 EST sequences obtained from CB and BM CD34+ cell libraries, respectively, the repetitive DNA elements, rRNA, and mitochondrial DNA sequences accounted for 11.7% and 17.3% of the total, respectively. After eliminating these sequences, the meaningful ESTs were classified into known gene, dbEST, and novel EST groups by database search. For useful ESTs from both origins, the known and named gene groups occupied the largest portion (5377 out of 7376 from cord blood and 2265 out of 3424 from bone marrow, respectively); the list of all ESTs corresponding to known genes from both origins is now available at http://www.chgc.sh.cn. The ESTs representing undefined genes (dbEST and novel EST groups) were assembled into 2060 clusters, which then served as candidates for cloning of full-length coding sequences.

Cloning of cDNAs with Putatively Entire ORF for Previously Undefined Genes

Sequences of cDNA clones representing 2060 EST clusters of undefined genes were obtained. Those clones with continuous sequences encoding at least 100 amino acids (with an exception of a few smaller ORFs bearing very high homology to the known small genes) were checked for the presence of putatively entire ORFs using the following criteria. First, when a sequence had high homology to a known gene, its ORFs were compared with each other. If the amino acid sequences of both ORFs initiated by an ATG codon could be reasonably aligned, the ORF contained in the novel gene cDNA was defined as a putatively complete one. Second, those sequences without homology to known genes were searched for in-frame stop codons upstream of an ATG codon-initiated ORF of >100 amino acids. If no such stop codon was found ahead of an ORF, the nucleic acid sequence flanking the first ATG should bear similarity to the well-conserved KOZAK motif (Kozak 1986). The above analysis revealed that 222 of our clones might contain an entire ORF. In 78 EST clusters, an obvious but incomplete reading frame was present. Different methods were employed to prolong the ORF in these 78 clones until complete ones were considered to be reached according to the aforementioned criteria. In silico cloning with dbEST extension allowed us to obtain 69 putative entire ORFs, which were then confirmed by sequencing of material cDNA clones obtained by appropriately designed RT-PCR. Finally, for those sequences that could not be extended properly with an electronic approach, rapid amplification of cDNA ends (RACE) was applied to get the 5′ or 3′ ends with Marathon-ready cDNA libraries from appropriate tissue origins. Another nine entire ORFs were cloned and sequenced this way. In total, 300 cDNAs with putatively entire ORFs were obtained. Their nucleic acid sequences were 360–3496 bp in length and their ORFs coded for peptides of 58–752 amino acids. The major features of each gene are summarized in Table 1. It is worth pointing out that, although a 3′ poly(A) sequence or a polyadenylation signal was found in most (214/300) cDNAs as evidence of containing the complete 3′ UTR, the integrity of the 5′ UTR could not be certain in the majority of the cDNAs.

In the remaining 1760 EST clusters corresponding to previously undefined genes, 512 clusters contained partial reading frames, 806 represented 3′ UTRs as they had no obvious reading frames but presented polyadenylation signal and poly(A) tails, and the remaining 442 contained sequences of which the features should be further analyzed.

Functional Significance Indicated by Homology Comparison with Genomic Sequences through Evolution

It is well accepted that homologous genes often share similarities at sequence and/or functional levels (Henikoff et al. 1997). Hence, sequence similarity acquisition is an efficient method to predict the function of a novel gene. Members belonging to the same gene families could be assumed/determined with this strategy and conserved genes often show conserved sequence elements within the important functional domains or motifs. Based on this consideration, putative ORFs from model organisms with completed genome sequence, including bacteria, *Saccharomyces cerevisiae*,

Table 1. List of the 300 cDNAs from HSPC and Their Analysis and Expression Screening in This Paper

GenBank Accession number	Gene name[a]	cDNA[b] (bp)	Pep (aa)	Motifs and structure features[d]	Exon number	Homology[d] B	Y	C	D	A	M	UniGene libraries[e]	Expression[f] NB4	HL6	Jurkat	K56	U937
AF038950	hABC-7	2384	752	ABC,P-loop,6TM	>12	a	a	a		a	c	U					
AF038952	co-chaperonin	574	108		1		a		a		c	U					
AF038953	E25	1082	263	LA, TM	6				a		c	U					
AF038954	vacuolar H(+)-ATPase subunit	1078	118		3			b	b	a	c	U					
AF038955	G protein gamma 5	520	68		1			a			b	U					
AF038956	GMF beta-h	561	142		7		a		b	a	c	U					
AF038957	IF4e-h	989	236	eIF-4E	8		a	b	a	a	c	U					
AF038958	SC2	1146*	308	LA							c	U					
AF038959	SC2(s)	629	157	LA					a	b		c					
AF038960	SKD1	1978	444	AAA, P-loop	>7	a	b	a	b	b	c	H					
AF038961	SL15	1410	247		7			a	b		c	U					
AF038962	HD-VDAC3	1414	283		10		a	a	b		c	U					
AF038965	proteasome subunit RS-2	1439	418	AAA	11	a	b	c	c	b	c	U					
AF038966	SCAMP	1926	338					a	a		c	H					
AF047432	ARF6	1356	175		>1		b	b	c	b	b	U					
AF047433	b(2)gcn-h	1112	245	LA	7	a	b	b	c	b	c	U					
AF047434	CI-15 hmolog	540*	106		2						c	U					
AF047435	CI-KFYI hmolog	419	76		4						c	U					
AF047436	F1Fo-ATPase synthase f	452	94	LA				a			c	U					
AF047437	FSA-1-h	1027	293		8				a		c	U					
AF047438	GOS28/P28	1012	255		>2				a		c	N					
AF047439	HSPC001	1138	300		8							U					
AF047440	RPL33 like	512	65									U					
AF047441	hRPA40	1103	342		9	a	a	b	b	b	c	U					
AF047442	vt-sec22b	1462	215	Syn, SP,TM			a	a	b	a	a	U					
AF054174	histone macroH2A1.2	1920*	371			a	b	b	b	b	c	U					
AF054175	68MP-h	627	58									U					
AF054176	AII/AVP like (AAR)	2218	514	LZ	>2						a	H					
AF054177	CHD-1 like	1147	220	DEAD/DEAH	>6	a	a	a	a	a	a	H					
AF054178	CI-B14.5a homolog	531	118	RPS5	>2				a		b	U					
AF054179	H beta 58	2669#	327		>3		b	b	b	b		U					
AF054180	ZNF254(HD-ZNF1)	1619	353	C2H2, ZB	>3		a	a	a		a	H					
AF054181	CI-MNLL homolog	437*	58	LA, IF	>3						c	N					
AF054182	MPPB	1771*	489	IF,ZB		a	a	b	b	a	c	U					
AF054183	ras like protein	1148	216				c	a	c	c	c	U					
AF054184	Sec61 gamma	482	68	secE/sec61		a	a	c	c	b		U					
AF054185	proteasom subunit HSPC	969	248		>7	a	b	b	b	b	c	U					
AF054186	P18	840	174						a			U					
AF054187	alpha NAC	1089	215						b	b	c	U					
AF067163	BMH	2790#	681	ABC, LZ	>1	a	a	a	a	b	c	U					
AF067166	CI-AGGG homolog	509	106		4						c	U					
AF067167	CI-B17 homolog	605	128		5						c	U					
AF067168	CI-B22 homolog	739	179					a	a		c	U					
AF067169	CI-PDSW homolog	678	172					a	a		c	U					
AF067170	endosulfine	2520	117		>3				a		c	U					
AF067171	GEF-2h	924#	117		6		b	c		b	b	U					
AF067172	HDRC	1919#	265		>5		a	a	b			N					
AF067173	h-Mago	650	146		>3			c	c	b	c	N					
AF070650	ATP synthase d	628	161		1				a		c	U					
AF070652	CI-B14.5b homolog	615	119	ZP, TM	>2						b	U					
AF070653	CI-B9 homolog	360	84	TM	5						c	U					
AF070654	cornichon	1395*	144	3TM			a	b	b		c	U					
AF070655	F1Fo-ATPase synthase g	501	103		>3			a	b		c	U					
AF070656	fsth-h	1922*	517	AAA,P-loop,SP,TM		b	b	a	b	b	c	U					
AF070657	GSTK1	1017	226	MCTS				a				U					
AF070658	HSPC002	561	163		4						c	N					
AF070659	HSPC003	602	125		5				b			U					
AF070660	HSPC004	790*	162		5				a			U					
AF070661	HSPC005	418	79	MI-SOD,2TM	4				b			U					
AF070662	HSPC006	929*	196		2				a		c	U					
AF070663	HSPC007	700	187		>1			a	a			U					

(Continues on pp 1549–1552.)

Table 1. (*Continued*)

GenBank Accession number	Gene name[a]	cDNA[b] (bp)	Pep (aa)	Motifs and structure features[d]	Exon number	Homology[d] B	Y	C	D	A	M	UniGene libraries[e]	Expression[f] NB4	HL6	Jurkat	K56	U937
AF070664	HSPC008	1709	323	Amid, LZ	>2		a		a		c	U					
AF070665	HSPC009	793	106		2						b	U					
AF070667	N-G,N-G-h	1351*	285	RGD, HRCT	7				a		c	U					
AF070668	RPS27i	500*	84	RPS27, SP	4		b	c	c	c		U					
AF077028	CI-ASHI homolog	679	186		5			a	a		c	U					
AF077029	CI-B8 homolog	600	99		3			b	b		c	U					
AF077034	HSPC010	1327	93		1				b			U					
AF077035	HSPC011	600	130	RPS4	4	a		a	a			N					
AF077036	HSPC012	1636	219									U					
AF077037	HSPC013	923	171	EGF-L	7							U					
AF077038	URP-h	1094	259		>3		a		b		c	U					
AF077039	TIM17-h	824	199				a	a	b	b	c	U					
AF077042	RPS7 like	1046	242	RPS7	5	a		a	a		b	U					
AF077043	RPL36	538	105	RPL36	3		a		b	b	c	U					
AF077044	RPA16	726	133		>1	a	a	a	a	a	c	U					
AF077196	ankyrin like	1084*	260		9	a	a	a	a	a	c	U					
AF077197	VAMP-2 like	609*	116		>3		a	a		a	b	U					
AF077198	lysophospholipase (LYPL-A1)	1348	230		>3	a	a	a	b		c	U					
AF077199	LYPL-A1 (short form)	1300	214			a	a				c	U					
AF077200	HSPC014	634	141		>5		a		b			U					
AF077201	HSPC015	1163*	337		>5		a		a			N					
AF077202	HSPC016	384	64		1			b				U					
AF077203	HSPC017	2100*	368		>10	a			b		c	U					
AF077204	HSPC018	700	167	TM								U					
AF077205	HSPC019	2411*	128		4							U					
AF077206	HSPC020	900	121		7		a				a	N					
AF077207	HSPC021	1897	564		>1			a	b			U					
AF077208	HSPC022	761*	201									U					
AF078848	BUP	933	195		>5							U					
AF078852	HSPC023	616	99		>3							U					
AF078856	P47	1346#	370		1		a		a	a	c	U					
AF083241	HSPC024	581	157		7			a				U					
AF083243	HSPC025	1901	564		13			a	b			U					
AF083244	HSPC026	1567*	275	Heme-b,C_3HC_4	6			b	b			N					
AF083245	HSPC027	1600	377					a	a	a		U					
AF083246	HSPC028	1869	419	P-loop, LZ	>7				b		c	U					
AF083247	MDG1	1929*	223	dnaJ	3	a	a	a	a	a	b	U					
AF083248	RPL26-h	678	145	RPL24, NDPO	1	b	b	b	b	b	c	U					
AF085357	flotillin	1729*	427		13	a			b		c	U					
AF085358	HSPC029	863	218		8			a	a			U					
AF085359	HSPC030	795	91									U					
AF085360	HSPC031	1363	180		5		b	b	b	b		N					
AF085361	HSPC032	1104	303	LA	>5			a	a			U					
AF085362	UbcM2	1294*	207	UbcE	1		b	b	c	b	c	U					
AF092138	HSPC033	798	91	LA, SP, 2TM	4			a	a	a		U					
AF100746	AUP1	1450	410		12							U					
AF100747	HSPC034	598	114		6							U					
AF100748	HSPC035	1840*#	339	SP	>3			a				U					
AF100749	Sec 22 -h	1773#	282	LA			a	a	a		c	U					
AF100750	SLAP-2-h	1568#	452		>5		a				b	N					
AF100763	AAK1	1707#	550	PK	>4	a	b	b	b	a	c	N					
AF125096	HSPC042	949	115		5				a			N					
AF125098	HSPC037	1182	185	ARF, P-loop	>4		a		a			U					
AF125099	HSPC038	719	76		>1			b	b			U					
AF125100	HSPC039	1583	82		>2					a		U					
AF125101	HSPC040	988	109		4		b	b	b		a						
AF125102	HSPC041	1898	104		6				b			U					
AF125103	NSPCh	1798*	199	LZ, SP, 2TM	>5			a			b	H					
AF151017	HSPC183	1441*	258					a	a			U					
AF151018	HSPC184	1013	183				a		b			U					
AF151019	HSPC185	1231*	236		>5							U					
AF151020	HSPC186	1592	183	SP, TM	>6			a	a			U					
AF151021	HSPC187	1152*	306	PK	1	a	a	b	a	a	c						
AF151022	HSPC188	1802	103				a					N					

Table 1. (*Continued*)

GenBank Accession number	Gene name[a]	cDNA[b] (bp)	Pep (aa)	Motifs and structure features[d]	Exon number	Homology[d] B	Y	C	D	A	M	UniGene libraries[e]	Expression[f] NB4	HL6	Jurkat	K56	U937
AF151023	HSPC189	1256*	222	C2H2	>2			a			a	U					
AF151024	HSPC190	988	147														
AF151025	HSPC191	643	116	Syn	4		a	a		a	b	U					
AF151026	HSPC192	722	160	A-tRNAs	>2	a	b	b	b	b							
AF151027	HSPC193	1513	118		2							U					
AF151028	HSPC194	1011*	112		>4				b			U					
AF151029	HSPC195	1108*	198		2							U					
AF151030	HSPC196	1193	162		>3							U					
AF151031	HSPC197	1906	290	LA, LZ								N					
AF151032	HSPC198	1187	162	LA	>4							U					
AF151033	HSPC199	2027	422	Amid, LZ	>2	b	a		b			N					
AF151034	HSPC200	862	175	NLS	>3	a	a	a	a	a	a	N					
AF151035	HSPC201	1072	153		6			b	b			N					
AF151036	HSPC202	1017	153	MCTS, 3TM			a		a			U					
AF151037	HSPC203	710*	106	GCR								U					
AF151038	HSPC204	2010	136		>1			a	a			N					
AF151042	HSPC208	984*	139		1						a	N					
AF151043	HSPC209	795	144	MCTS	>1							U					
AF151044	HSPC210	1152*	155		4			a				U					
AF151045	HSPC211	621	110									U					
AF151046	HSPC212	1614	297		>9							N					
AF151048	HSPC214	641	132				a	b	c	a		U					
AF151049	HSPC215	512	119		>2							U					
AF151050	HSPC216	1702	342		>3							U					
AF151051	HSPC217	1760	170	MDB	>4							U					
AF151052	HSPC218	1461	127		>2							U					
AF151053	HSPC219	933*	157		>3	b			a			N					
AF151054	HSPC220	1818	176		>3			a	b			U					
AF151055	HSPC221	1100*	180		>3							N					
AF151056	HSPC222	1120*	117	LA							c	U					
AF151057	HSPC223	1208	309		>4	a		a				U					
AF151058	HSPC224	692	152	LZ	>4												
AF151059	HSPC225	1954*	205		>5	a	a				c	N					
AF151060	HSPC226	1349*	149		>8							U					
AF151061	HSPC227	1006	140	PPPT	>1			a	a		a	U					
AF151062	HSPC228	1919	305		8			a	a			U					
AF151063	HSPC229	1475	203									U					
AF151064	HSPC230	923	240		4				a			N					
AF151065	HSPC231	604	106						b		c	U					
AF151066	HSPC232	1104	319		>5							N					
AF151067	HSPC233	1548*	286		>2			a	a			U					
AF151068	HSPC234	873*	198	LA	>3							H					
AF151069	HSPC235	1466*	352		>2							U					
AF151070	HSPC236	855	185		>5		a					U					
AF151071	HSPC237	908*	169	GFR	1							U					
AF151072	HSPC238	601	153	C3HC4	>1		a	a	a	a	a	U					
AF151073	HSPC239	1147*	293						a			N					
AF151074	HSPC240	1614*	246	Heme-b				a	a			N					
AF151075	HSPC241	684	128		3			a	a			U					
AF151076	HSPC242	1294*	206		4			a	a			U					
AF151077	HSPC243	1582	241		>3	a					a	H					
AF151078	HSPC244	1060	87		5			a				U					
AF151079	HSPC245	497	124	Tyrosinase	5		a	a	b		c	N					
AF151080	HSPC246	799*	112									U					
AF151081	HSPC247	998	103		1							U					
AF151083	HSPC249	694	151		3							U					
AF151084	HSPC250	712	148		>1	a	a					N					
AF151085	HSPC251	1141*	212		>3	a					a	U					
AF151086	HSPC252	1289*	323		>3							U					
AF161458	HSPC108	2266	342		>9	b		b	b	a	a	U					
AF161459	HSPC109	1261	389	NLS	12				a			N					
AF161460	HSPC111	910	178		>5							N					
AF161461	HSPC112	664	56	LA	>3						b	U					
AF161462	HSPC113	841	178	SP, TM	>7						a	U					

Table 1. (*Continued*)

GenBank Accession number	Gene name[a]	cDNA[b] (bp)	Pep (aa)	Motifs and structure features[d]	Exon number	Homology[d] B	Y	C	D	A	M	UniGene libraries[e]	Expression[f] NB4	HL6	Jurkat	K56	U937
AF161463	HSPC114	744	174	Amid, RNP-1	1	a		b	b	a	a	U					
AF161464	HSPC115	976	219	mutT		a	a	a				U					
AF161465	HSPC116	1710	314		>9			a	a		a	U					
AF161466	HSPC117	2049	505		>12	b		b	c			U					
AF161467	HSPC118	677	163		5				b			N					
AF161469	HSPC120	1696*	465		>8	a	a	b	b	b	c	N					
AF161470	HSPC121	1693*	373		>6			a	a		c	U					
AF161471	HSPC122	1174	269								c	U					
AF161472	HSPC123	1524	213						a			U					
AF161473	HSPC124	1211	318	IPP	>5	a	b	b	b	a	b	U					
AF161474	HSPC125	725*	175		>1				a			U					
AF161475	HSPC126	1424	270		>5			a	a			U					
AF161476	HSPC127	2116*	188		>3			a	a			U					
AF161477	HSPC128	1066*	241		1							U					
AF161478	HSPC129	3041*	466				b	a	a	a		U					
AF161479	HSPC130	2130	473		16				b			U					
AF161480	HSPC131	1765*	468		>14			a	b	a		N					
AF161481	HSPC132	1171	76		>2		a					N					
AF161482	HSPC133	963	132		6	a		a	b			U					
AF161483	HSPC134	1009	223		6		a	b	b								
AF161484	HSPC135	995	307	LZ, RDG	4	a						H					
AF161485	HSPC136	492	121		5				a								
AF161486	HSPC137	2588	237	P-loop, CAAX	7		a	a	b	a	c	N					
AF161488	HSPC139	1155	101	Ua-E	>3			a	b	a	c	U					
AF161489	HSPC140	2071	346		>3	a	a	a	a	a	a	U					
AF161490	HSPC141	615	188		3							U					
AF161491	HSPC142	1432*	367		10							U					
AF161492	HSPC143	1234	234		6	a	b	b	b	a		U					
AF161493	HSPC144	1323	225		>8	a						U					
AF161494	HSPC145	1147	296		>4		a		b			U					
AF161495	HSPC146	1443	204		>5		b	b	b			U					
AF161496	HSPC147	1860	350	LA, WD-40		a	a	a	b	a	c	U					
AF161497	HSPC148	1050	229		>3			b									
AF161498	HSPC149	1116	324	LA, LZ, SP, 2TM	>6	a	a	a	b	a		U					
AF161499	HSPC150	928*	197	UbcE	7		a	a	a	a	a	U					
AF161500	HSPC151	525	83	TM	2							U					
AF161501	HSPC152	612*	125		4		a	a	b			U					
AF161502	HSPC153	1070*	318	NLS	8		a	b	a	a	a	N					
AF161503	HSPC154	1347	248		>3			a	a			U					
AF161504	HSPC155	1137	167						c			U					
AF161505	HSPC156	708	108	C-5, Syn	2							U					
AF161506	HSPC157	1074	154	LA,TM	2							U					
AF161507	HSPC158	724	228	RPL22	7	a			b			U					
AF161508	HSPC159	801	172		5			a			a	H					
AF161509	HSPC160	609	153	MCTS, 2TM	6		a		b	a		U					
AF161510	HSPC161	1416	296		2							U					
AF161511	HSPC162	687	96					a	b		c	U					
AF161512	HSPC163	652	139	3TM			a	a	a		a	U					
AF161515	HSPC166	1318	198		>5							U					
AF161516	HSPC167	2091	586		>6	a		a	a	b		U					
AF161517	HSPC168	1199	292		4							U					
AF161518	HSPC169	1605	306		9			a	a	a		U					
AF161519	HSPC171	850	145	SP	>6				a								
AF161520	HSPC172	1002	219		>5			a	b			U					
AF161521	HSPC173	1950	589			a	a	b	b		c	U					
AF161522	HSPC174	1592*	253	LA, 4TM	5							U					
AF161523	HSPC175	681	125	RNP-1	4		a	b	c	b	a	N					
AF161524	HSPC176	624	139		>4		a		b			U					
AF161525	HSPC177	1028*	219		>1		a	b	b	b		N					
AF161526	HSPC178	852*	211	2TM	>3	a		a			a	U					
AF161527	HSPC179	883	197	TM	>2				a			N					
AF161528	HSPC180	1914	133		>4		b	b	b	b		N					
AF161529	HSPC181	942	177	CAC	>1		a	b	b		c	U					
AF161530	HSPC182	1059*	194	CAT	>3		a	b	b			U					

Table 1. (*Continued*)

GenBank Accession number	Gene name[a]	cDNA[b] (bp)	Pep (aa)	Motifs and structure features[c]	Exon number	Homology[d] B	Y	C	D	A	M	UniGene libraries[e]	Expression[f] NB4	HL6	Jurkat	K56	U937
AF161531	HSPC046	1743	291	RGD, SP	13							H					
AF161532	HSPC047	918	136		>2							N					
AF161533	HSPC048	710	110		1												
AF161534	HSPC049	2610	741					a	a			N					
AF161535	HSPC050	1139	110	MCTS	>2												
AF161536	HSPC051	450	132	LA	2							N					
AF161537	HSPC052	928	127		>3												
AF161538	HSPC053	531	154		4												
AF161539	HSPC054	1173	96		1												
AF161540	HSPC055	2629	167	C2H2	5												
AF161541	HSPC056	2879	275		>6					a		N					
AF161542	HSPC057	3496	515		17				a			N					
AF161544	HSPC059	2299	599	C2H2	3		a	a	a		a						
AF161545	HSPC060	2489	341														
AF161546	HSPC061	1661	320	Amid, LZ	2		a		a		c						
AF161547	HSPC062	1834	176	TM	>2							H					
AF161548	HSPC063	2210	203		>1												
AF161549	HSPC064	2032	96	MCTS	>2		b	b	b								
AF161550	HSPC065	893	190	Amid, LZ	>3												
AF161551	HSPC066	1425	280		9						a						
AF161552	HSPC067	1495	283					a	a								
AF161553	HSPC068	2801	641		>15			a	a	a	a						
AF161554	HSPC069	2081	591		>2		a	a		a	a						
AF161555	HSPC070	3050*	350	C3HC4, RGD	8				a								
AF161556	HSPC071	968	179		>2				a			U					
AF161557	HSPC072	961	95	RPL33	5												
AF161558	HSPC073	705	123		>1												
AF229065	HSPC043	1001	104		8							H					
AF229066	HSPC045	1236	213		1							N					
AF229067	PADI-h	862	214		>2						b						
AF229068	HSPC170	612	125		8		a	a	b			N					
AF229069	KIAA0220-h	496	153		>5							U					
AF208847	BM005[$]	1947*	645	C2H2	5						a	N					
AF208849	BM007[$]	780	566	CAAX	>5							U					
AF208854	BM012[$]	1999*	350					a	b			U					
AF208858	BM016[$]	1821	117									U					
AF208859	BM017[$]	2764*	1028		>3							U					
AF208862	BM020[$]	1556*	326		5							U					
AF208863	BM021[$]	2113*	467	ACBD, LA	>4							U					
AF208865	hEDRF[$]	501	102		12						b	U					
AF212220	hTara[$]	2986	221		4				b		c	U					
AF212222	BM024[$]	1720*	211		>2						a	U					
AF212223	BM025[$]	881	114						a			U					
AF212225	BM022[$]	1283	304		>6	a						U					
AF217508	BM031[$]	1581	233	SCO	4		a	a	a		a	U					
AF217514	BM038[$]	1146*	206		>2							U					
AF217522	BM046[$]	1721	251								a	U					

[a]Genes marked with $ were cloned from bone marrow library; others were from cord blood library.
[b](*) Full-length cDNA was obtained with dbEST assemblying; (*) obtained with RACE.
[c]Abbreviations of the motifs or structure features were listed in Table 2.
[d]Based on percentage sequence identity, homologous genes were divided into 3 groups: (a) 25-50%; (b) 50-75%; (c) >75%.
(B) Bacteria; (Y) yeast; (C) *C. elegans*; (D) *drosophilia*; (A) *arabidopsis*; (M) mammals (excluding primates).
[e]Tissue resources from UniGene. (U) >10; (N) <10 tissues; (H) 1 or 2 besides hematopoiesis related tissues; (blank) no UniGene entries hit.
(f)Array screening in 5 hematopoietic cell lines. Gray densities stand for the relative signal intensity comparing with GAPDH on same membrane (detailed primary results could be obtained from website: http://www.chgc.sh.cn).
<0.1 0.1–0.5 0.5–2.0 >2.0

Caenorhabditis elegans, and *Drosophila*, and ORFs of identified genes from *Arabidopsis* and mammals (excluding primates) were retrieved to compare the amino acid sequence similarities with those of ours (Table 1; Fig. 1). Sequences with similarity >25% over a region of 50–100 amino acids were considered here to have some homology (Russell et al. 1997). Among our 300 cDNA sequences, 21 share similarity to the coding sequences in all species examined, indicating that they are well-conserved genes and important for cell life. In fact, 16 of these 21 genes have assigned functions. A total of 204 cDNAs contained ORFs with >25% similarity to the sequences in at least one species. Functional clues have been available in 105 of these 204 genes. Taken as a whole, at least 225 genes identified in the current work are evolutionarily conserved. Interestingly, as shown in Figure 1, an increased gradient of similarity in terms of both number of related genes and the degree of homology is present from bacteria to *Drosophila*. In the case of *Arabidopsis*, only part of the genomic sequence is available in the public database. However, 66 of our cDNAs found their homologs in this plant. As expected, the number of genes with high homology (>50%) was great in mammals. The fact that 75 cDNAs had so far no obvious similarity to any genes across different species implied that they might be functionally specific genes acquired relatively late during evolution.

Structural and Functional Assignment with Bioinformatic Prediction

Basic structural motifs predicted by some algorithms on the primary structure in the ORFs are listed in Tables 1 and 2, including leucine zipper, C2H2 zinc finger, and C3HC4 ring finger. Some consensus patterns of protein kinase, growth factor, and cytokine receptor-associated protein were also found by such methods. However, caution should be taken in interpreting these data. For instance, leucine zipper motif was predicted on primary structure in 12 ORFs using the Motifs software in the GCG package. Further analysis with Coilscan and Peptidestructure programs also provided by the GCG package revealed, nevertheless, that only 1 of these 12 leucine zippers was located in a coiled-coil structure. Because a typical leucine zipper should be included in a coiled-coil domain, this result indicates the importance of integration of information generated by different prediction methods, including those for conserved motifs at primary sequence level and those for secondary or higher structures. In analyzing the signal peptide, two different approaches, Spscan (in GCG package) and signalP (http://www.cbs.dtu.dk/services/Signalp/) were applied to our ORFs. The former algorithm is based on the weight matrix method in concert with McGeoch's discrimination of a minimum signal peptide, whereas the latter is based on two neural network methods for recognition of signal peptides and their cleavage sites. Of note, only cleavable signal peptides, but not the uncleavable ones like signal anchor and internal signal, can be detected with these algorithms. Interestingly, the two approaches gave quite coherent results in predicting putative amino-terminal potential signal peptides in 11 ORFs, including 8 with α-helix transmembrane domains outside the signal peptide region. One such example was an ORF with both signal peptide and 6-transmembrane domains (HABC7, GenBank accession no. AF038950), which contains an ABC transporter family signature. We therefore speculated this ORF encodes a putative transmembrane transporter protein.

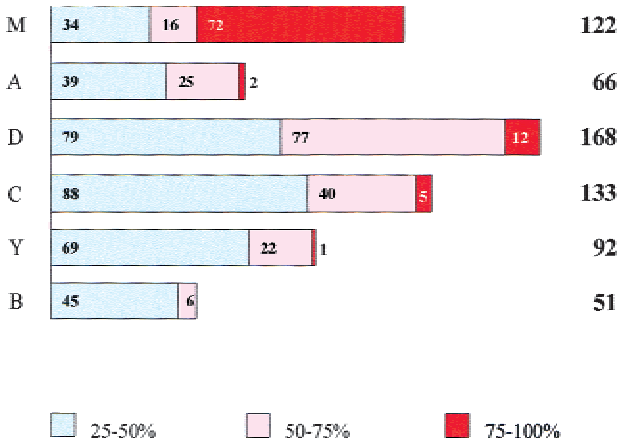

Figure 1 Homology comparison of ORFs contained in our cDNAs to known genes from different model organisms. The horizonal blocks represent numbers of the ORFs bearing homology to genes in a given species; different colors indicate the degree of homology. Each number listed at *right* indicates the total number of our ORFs having homologous genes in that organism. (B) Bacteria; (Y) yeast; (C) *C. elegans*; (D) *Drosophila*; (A) *Arabidopsis*; (M) mammals not including primates.

Genomic Organization and Alternative Splicing Identification

Of our genes, 243 were preliminarily characterized in terms of exon-intron organization after comparison of cDNA sequences with the genomic sequences in the database (Table 1). The estimated genomic sizes of these genes spanned 384 bp to 144 kb, containing 1 to >17 exons, and correspondingly 0 to >16 introns. The size distribution of the exons was from 20 bp to 2023 bp, whereas that of characterized introns ranged from 77 bp to 86 kb. Of note, 17 genes composed of only 1 exon varied in sizes from 384 bp (HSPC016, accession no. AF077202) to 1346 bp (P47, accession no. AF078856). On the other hand, cDNAs of short length could contain multiple exons. For example, HSPC245 (accession no. AF151079), consisting of 5 exons, and HSPC024 (accession no. AF083241), consisting of 7 ex-

Table 2. List of the Abbreviations of Motifs and Structure Features in Table 1

Abbreviation	Motifs/Structures
AAA	AAA-protein family signature
ABC	ABC transporters family signatures
ACBD	Actinin-type actin-binding domain signatures
amid	Amidation site
ARF	ADP-ribosylation factors family signature
A-tRNAs	Aminoacyl-transfer RNA synthetases class-I signature
C2H2	Zinc finger, C2H2 type, domain
C3HC4	Zinc finger, C3HC4 type (RING finger)
C-5	C-5 cytosine-specific DNA methylases signatures
CAAX	Prenyl group binding site (CAAX box)
CAC	Clathrin adaptor complexes small chain signature
CAT	Aspartate and ornithine carbamoyltransferases signature
DEAD/DEAH	DEAD and DEAH box families ATP-dependent helicases signatures
dnaJ	dnaJ domains signatures and profile
EGF-L	EGF-like domain signatures
eIF-4E	Eukaryotic initiation factor 4E signature
GCR	G-protein coupled receptors signature
GFR	Growth factor and cytokines receptors family signatures
Heme-b	Cytochrome c family heme-binding site signature
HRCT	Hexapeptide-repeat containing-transferases signature
IF	Insulinase family
IPP	Inorganic pyrophosphatase signature
LA	Prokaryotic membrane lipoprotein lipid attachment site
LZ	Leucine zipper pattern
MCTS	Microbodies C-terminal targeting signal
MDB	Myb DNA-binding domain repeat signatures
MI-SOD	Manganese and iron superoxide dismutases signature
mutT	mutT domain signature
NDPO	Pyridine nucleotide-disulphide oxidoreductases class-I active site
NLS	Nuclear localization signal
PK	Protein kinases signatures and profile
P-loop	ATP/GTP-binding site motif A (P-loop)
PPPT	Purine/pyrimidine phosphoribosyl transferases signature
RGD	Cell attachment sequence RGD tripeptides
RNP-1	Eukaryotic putative RNA-binding region RNP-1 signature
RPL22	Ribosomal protein L22 signature
RPL24	Ribosomal protein L24 signature
RPL33	Ribosomal protein L33 signature
RPL36	Ribosomal protein L36e signature
RPS27	Ribosomal protein S27e signature
RPS4	Ribosomal protein S4e signature
RPS5	Ribosomal protein S5 signature
RPS7	Ribosomal protein S7 signature
SCO	Serine carboxypeptidases active sites
secE/sec61	Protein secE/sec61-gamma signature
SP	Signal peptide
Syn	Synaptobrevin signature
TM	transmembrane region
Tyrosinase	Tyrosinase signatures
Ua-E	Ubiquitin-activating enzyme signatures
UbcE	Ubiquitin-conjugating enzymes active site
WD-40	Trp-Asp (WD-40) repeats signature
ZB	Zinc-binding region signature
ZP	ZP domain signature

ons, were only 497 bp and 581 bp in length, respectively.

During the characterization of the genome organization of our genes, some alternative splicings were determined. A 453-bp sequence in hSC2 (accession no. AF038958) was deleted in an isoform (accession no. AF038959), which was only found in CD34+ cells so far, whereas *LYPL-A1* (accession no. AF077198) used a 48-bp stretch that did not exist in the short form transcript (accession no. AF077199) (Fig. 2A). The fact that these alternatively used sequences are located in ORFs in an in-frame way supports the idea that these are physiologically existing isoforms and not artifacts in cDNA library construction. Indeed, the isoforms of the

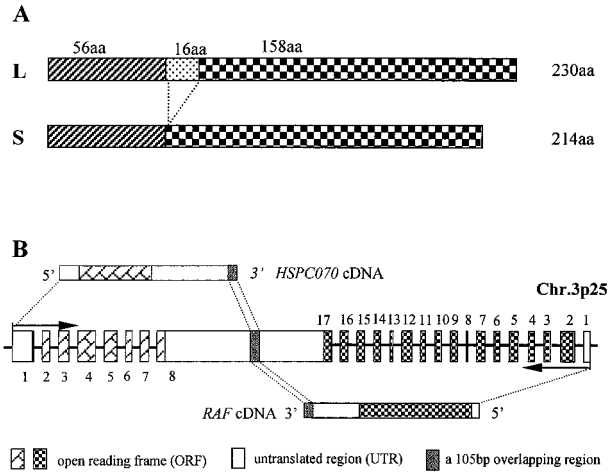

Figure 2 (*A*) Alternative splicing present in lysophospholipase gene transcripts as long (L, accession no. AF077198) and short (*S*, accession no. AF077199) forms. The numbers indicate the amino acid positions of deduced proteins. Note that the ORF is maintained in the alternatively spliced S isoform. (*B*) Overlapping of HSPC070 (accession no. AF161555) and *RAF* genes located on opposite DNA strands at the same locus. Both genes are mapped to the same region on chromosome 3p25. The comparison of sequences between cDNAs and genomic DNA has allowed the exon–intron structure of both genes to be established, with exons represented by boxes and their numbers indicated. Note that a stretch of 105 bp is shared by the 3′ UTRs of both genes. Arrows indicate the orientations of transcription.

two genes were further confirmed by RT/PCR assay (data not shown). Interestingly, the cDNA sequence of *HSPC070* (accession no. AF161555) located on chromosome 3p25 was found to share a 105-bp stretch in the 3′ UTR including the polyadenylation signal with that of RAF oncogene (accession no. X03484) (Bonner et al. 1986) in reversed orientation (Fig. 2B). This was further confirmed by the draft genome sequence from GenBank (AC018494, AC018500, AC026153, and AC026170) (see legend for Fig. 2B).

Chromosomal Mapping

Chromosome localization is an important aspect of a gene's general information. Combining strategies of bioinformatics acquisition from both UniGene and other databases, and radiation hybrid (RH), a total of 230 genes were mapped to proper chromosome positions (Fig. 3). Among 55 genes mapped with G3 or GeneBridge 4 RH panels, 38 had not been mapped previously, whereas the remaining 20 RH results showed good concordance with those by electronic mapping. The detailed mapping results are available at http://www.chgc.sh.cn. Of note, the 5 C2H2 zinc finger genes are all located on chromosome 19.

Expression Patterns in Different Tissues and in Distinct Hematopoietic Cell Lines

Among the 300 cDNAs, 270 could be analyzed using electronic Northern because their dbEST hits were available from UniGene resource. As shown in Table 1, most (207/270) genes showed ubiquitous transcriptional expression patterns as their corresponding ESTs were found in >10 tissues. The expression was found in a more selective way (<10 tissues) in 63. Only 13 showed relatively restricted expression in hematopoietic organs/tissues (bone marrow, foetal liver, spleen, lymph nodes, etc.). To explore the biological meanings of our genes in hematopoiesis, 285 cDNAs from the CB CD34+ cell library were also examined using cDNA macroarray for their expression levels in hematopoietic cell lines (the array membrane used in this work did not include the 15 cDNAs from the BM CD34+ cell library). The cDNA probes were prepared with mRNAs isolated from NB4 (granulocytic), HL60 (granulocytic), U937 (monocytic), K562 (erythro-megakaryocytic), and Jurkat (T lymphocytic) cell lines representing distinct lineages of hematopoietic cells. The RNA quality was ensured with appropriate ratio between 18S and 28S rRNA bands on agarose gel electrophoresis, and the labeling efficiencies of cDNA probe were confirmed to be >50%. To evaluate the expression levels, the membranes were exposed to Phosphor screen and the relative intensity of each gene was quantified with FLA-300 detection system. Hybridization signals in separate experiments with different membranes and/or probes were calibrated using housekeeping genes including *GAPDH* and total amount of signals on the membrane as reference. The feasibility of the technology system was confirmed by reproducible results of the paralleled duplicate spots on the same membrane (Fig. 4A) and with independent tests on different membranes (Fig. 4B). The comparison of expression levels in different cell lines for 285 genes examined is shown on Table 1 (normalization with different references revealed similar results though only those based on *GAPDH* control are shown). Although most genes exhibited expression in all five cell lines, 35 of them displayed restricted expression in only one or two lineages. Northern blot analysis was performed for three genes, HSPC070, ZNF254, and HSPC135. According to the UniGene data, HSPC070 has a ubiquitous expression pattern, whereas the expression of ZNF254 and HSPC135 could be restricted to hematopoietic system (Table 1). Indeed, Northern blot analysis showed that HSPC070 was expressed in a variety of tissues (Fig. 4C) whereas no obvious transcriptional expression of ZNF254 and HSPC135 was detected in these tissues (data not shown). However, the three genes were all found expressed in most of the hematopoietic cell lines examined in this work.

DISCUSSION

Because tissue- or development stage-related differential expression exists for many genes, cloning of full-

Figure 3 Chromosome localization of 230 previously undefined genes by applying both STS searching and radiation hybrid (those marked with #). Detailed mapping information can be obtained from http://www.chgc.sh.cn

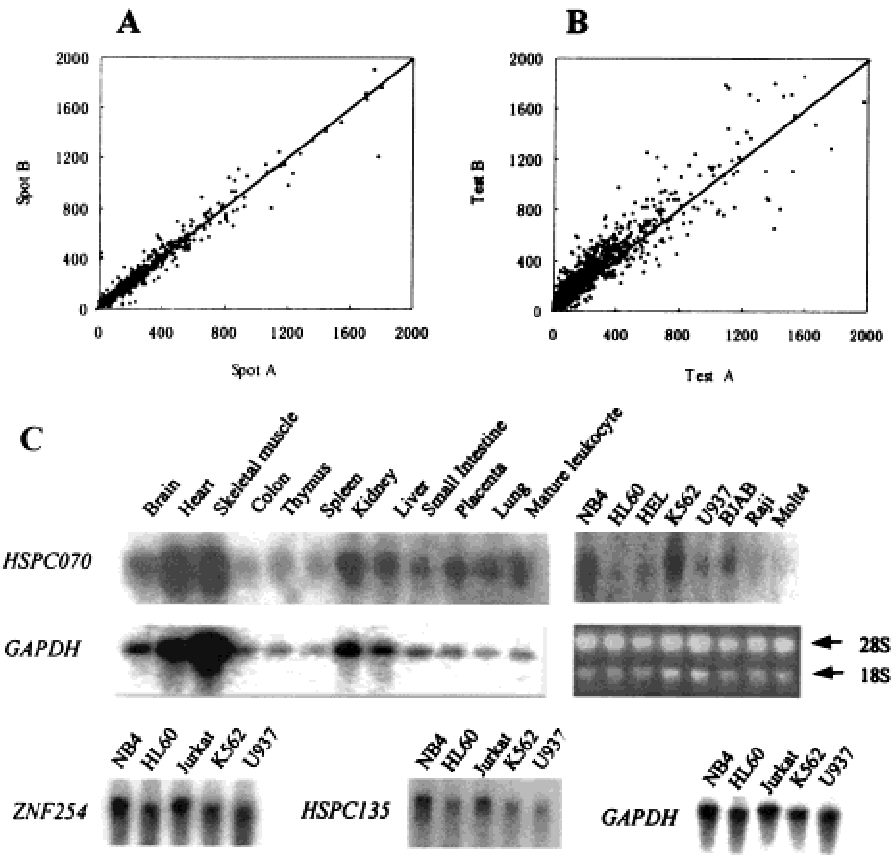

Figure 4 Regression analysis of the cDNA array results (*A*,*B*) and Northern blot analysis of three cDNAs (*C*). (*A*) The scatterplot of detected signal intensity for duplicate spots on the same membrane. (*B*) Scatterplot of detected signal intensity for the corresponding dots in two membranes with independent tests from RNA of same cell origin. All signals are normalized by using *GAPDH* gene as internal control. The figures were made with Microsoft Excel spread sheet and the correlation line was indicated. (*C*) Northern blot analysis of *HSPC070*, *ZNF254*, and *HSPC135*. (*Top*) *HSPC070* with a ubiquitous tissue expression pattern. *GAPDH* or 28S/18S ribosomal RNAs were used as sample loading control. (*Bottom*) Expression of *ZNF254* and *HSPC135* in hematopoietic cell lines, with *GAPDH* as control.

length cDNA based on EST analysis in different tissues represents a useful approach for gene identification, especially for those subject to temporo-spatial regulation. In strict sense, a full-length cDNA should cover both the ORF and the complete 5′ and 3′ UTR. Although a number of methods have been used to surmount the technical obstacles for getting the 5′ end of cDNA (Carninci et al. 1996), it is still difficult to reach the transcription start site in many cases. However, as the most important functional information of an mRNA is contained in the ORF, cDNAs containing entire ORFs are often considered as being full-length. By combining several technologies including construction of full-length cDNA enriched libraries, in silico cloning, and RACE, a relatively efficient working system has been established to obtain full-length cDNAs, or more precisely, cDNAs including entire ORFs, in a cost-effective way. This system has enabled the first resource of cDNAs with putatively entire ORFs to be generated for previously undefined genes whose expression is found in human CD34+ HSPCs.

One strong challenge to genomic science presently is to elucidate the functions of the newly discovered huge amount of genes. In this work, we tried to apply the currently available bioinformatic tools to the analysis of the structural and functional characteristics of each ORF. Using BLAST search, 121 out of 300 ORFs were found to share homology to genes with functional information, offering important clues for the choice of appropriate functional assays in further study. The difficulty was how to deal with the majority of the ORFs without obvious functional information. We therefore attempted to evaluate the conservation of the sequences through evolution. As a result, 225 ORFs show >25% similarity at amino acid level to those identified in organisms including bacteria, *S. cerevisiae*, *C. elegans*, *Drosophila*, *Arabidopsis*, and nonprimate mammals, whereas 75 have so far no similarity. It is quite possible that the 21 ORFs well-conserved across a wide range of species may be derived from the "essential genes." Although a large proportion of these evolutionarily conserved genes are of unknown function, this analysis can provide at least the following information: On the one hand, they are most likely to exert important biological functions; and on the other, the lower organisms containing homologous sequences can be used as models in the functional study with gene knockout or other methods. Moreover, efforts have been made to approach the gene function by search of distinct motifs and domains with combined use of algorithms based on different methods and taking into consideration not only the primary sequence but also the secondary structure of the proteins. Of note, in addition to those well-known functional motifs such as zinc finger and leucine zipper, a putative signal peptide was found in 11 ORFs with or without transmembrane motif in proper location. This information may lead to future works to identify possible secreted proteins and transmembrane proteins, and hence may allow recog-

nition of new regulatory pathways involved in the self-renewal and/or differentiation of HPSCs.

Characterization of gene expression with regard to tissue distribution is another way to approach the gene function. Genes with ubiquitous expression are more likely "housekeeper" genes, whereas genes whose expression shows tissue specificity may exert functions related to the development and differentiation of a given tissue or cell population. In this work, both electronic Northern and macroarray screening were carried out to study gene expression patterns. Because the majority of the genes presented in this work had been already hit by dbESTs and relevant information was available in UniGene (Boguskiand Schuler 1995; Shi et al. 1999), the electronic Northern could give an approximate estimation of the tissue distribution patterns. Of note, among 270 genes thus analyzed, 207 were hit by ESTs from >10 tissues while only 13 were mainly hit by ESTs of hematopoietic tissues. On the other hand, the macroarray system with relatively high efficiency and throughput was used in this work to study gene expression within the hematopoietic systems. Probes prepared from five hematopoietic cell lines were applied to cover granulocytic, monocytic, erythro-megakaryocytic, and lymphoid lineages. Of 285 genes expressed in CD34+ cells of cord blood origin, 35 were picked that showed relatively restricted or preferential expression along with a given orientation of differentiation. Therefore, combination of the two methods allowed us to find genes which may play a role in hematopoiesis-related functions.

In this work, we have also tried to take the opportunity of ever-increasing genomic mapping and sequence data to promote the understanding of structural organization of our genes discovered by cDNA approach. Application of bioinformatic information from public database, including sequence tag sites (STS) map (Stewart et al. 1997) and UniGene database (Boguski and Schuler 1995), allowed us to assign the chromosomal localizations for 192 novel genes. Retrieving genomic sequences from the "working draft" corresponding to our cDNAs obtained the exon–intron organizations in 243 genes, and the characterization of genomic structure of all genes can be expected in the near future with the accelerated schedule of the Human Genome Project. Although our work is only a small part in the international effort to establish a detailed whole genome transcription map, it may give some suggestions to the future study. Now, the gene discovery in genomic DNA sequencing depends largely on annotation but the successful rate based on theoretical prediction is not high enough. Hence, full-length cDNA cloning projects will provide the definitive evidence to the predicted transcription units. In contrast, genomic DNA sequences can also offer unique information for the full-length cDNA cloning. For instance, obtaining the 5′ ends of genes with large coding sequence is often difficult. Exon prediction may lead experimental work to help their cloning. Besides, genes with very low expression levels or extremely narrow expression windows may be absent or poorly represented in most of the cDNA libraries. Annotation of genomic sequences may facilitate the identification of these genes. Moreover, comparison of cDNA and genomic sequences can reveal some complex mechanisms of genomic organization and expression. To this end, it is interesting to note the overlapping in reversed orientation of our HSPC070 gene and the known *RAF* gene located on chromosome 3p25, as well as the alternative splicing patterns in some genes. According to the comparative analysis between the whole genome sequence data from *C. elegans* (The C. elegans Sequencing Consortium 1998) and *Drosophila* (Adams et al. 2000), the functional complexity of a genome is determined not only by the number of the genes, but even more importantly by the alternative splicing as well as complex regulatory mechanisms of the genome at transcriptional level. Finally, the chromosomal distribution of genes bears not only evolutionary meaning, such as the mapping of all five C2H2 zinc finger genes on chromosome 19 suggestive of recent duplication events, but also indicates candidate genes in disease-related loci.

Methods

EST Sequencing and Data Analysis

Mononucleated cells were harvested from cord blood and bone marrow with gradient centrifugation and CD34+ populations were separated with anti-CD34 MAb-conjugated MACS system (Miltenyi Biotec, Germany). After two rounds of separation, CD34+ cells were of 96%–99% purity according to flow cytometry analysis (Gu et al. 2000). RNA extraction, ZAPII cDNA libraries construction, Bluescript phagemid templates preparation, sequencing strategy, and data management were manipulated as before (Mao et al. 1998; Gu et al. 2000). The sequencing primers were universal primers including M13 Reverse and/or Forward, T3 and/orT7 primers, and sequencing mix was BigDye Terminator (Perkin Elmer). 5′ or 3′ end ESTs generated were categorized into known gene, dbEST, and novel EST groups by searching against GenBank database with BLAST and FASTA programs in GCG package.

Cloning of Full-Length cDNA

The EST clones corresponding to previously undefined genes were candidates for full-length cDNA cloning. The clone inserts were sequenced with end sequencing, primer extension, and sequencing after partial deletion/subcloning. AutoAssembler (Perkin Elmer) was applied to assemble the sequences into contigs. DNA Strider (Version 1.0) was employed to analyze the ORF. For those clones containing partial reading frames, in silico EST assembly and RACE were performed. Proper Marathon-ready cDNA libraries (Clontech) were chosen as RACE template, and the gene-specific primers were generated according to the clone sequence. The ORFs thus obtained were confirmed with RT-PCR.

Structure and Function Analysis with Bioinformatics

Sequence Similarity Comparison

The GCG package contains the release versions of EMBL and GenBank databases where the known genes and predicted ORFs were deposited. All amino acid sequences encoded by our cDNAs were searched against the nucleic acid sequence sub-databases of some important model organisms such as bacteria, *S. cerevisiae*, *C. elegans*, *Drosophila*, *Arabidopsis*, and mammals (excluding primates) with the tfasta program in the GCG package. There were two reasons to choose this strategy for homology search: First, there were many more nucleic acid sequences than amino acid sequences in the databases; second, through evolution, the amino acid sequences are more conserved than those of nucleic acid ones. In this study, two amino acid sequences were considered as homologs when they shared a similarity >25% over a region of 50–100 amino acids and the Z-score value was >200. Based on the percentages of sequence identity, these homologs were divided into 3 groups: 25%–50%, 50%–75%, and 75%–100%.

Genomic Organization Determination

The human genome sequences in GenBank (release 113) and htgs database hit by our cDNAs were retrieved, and the exon–intron organization was obtained by sequence comparison with the sim4 program (Yan et al. 1998).

Fundamental Structural and Functional Elements Searching

Programs including Motifs, Profilescan in GCG package, and Prosite at the Expacy website (http://www.expacy.ch/tools/scnpsite.html) were employed to scan for the motifs on primary structure of the peptides (Hofmann et al. 1999). Programs including Peptidestructure, Plotstructure, Pepplot, Coilscan, and Hthscan in the GCG package were applied to analyze the secondary structure of the proteins, and Spscan (GCG package) and signalP (http://www.cbs.dtu.dk/services/SignalP/), as well as TMHMM (http://www.cbs.dtu.dk/services/TMHMM-1.0/), were used to predict the signal peptide and the α-helix transmembrane domains in those novel ORFs so as to explore the secreted or membrane anchored proteins.

Chromosomal Mapping

Electronic Mapping

dbESTs were searched to find the corresponding sequences, then UniGene database (http://www.ncbi.nlm.nih.gov/UniGene) was applied to determine the tissue expression pattern and chromosomal mapping of these novel genes (Schuler et al. 1996). The cDNA-matched genomic DNA sequence data can also provide mapping information.

Radiation Hybrid

In addition to the electronic mapping results, Stanford G3 and GeneBridge 4 Radiation Hybrid (RH) panels (Research Genetics Inc.) were applied to map the novel genes according to procedures described previously (He et al. 1998). The results were submitted to the RH Mapping Server at Stanford Human Genome Center (SHGC; http://www-shgc.stanford.edu) and Whitehead Institute/MIT Center for Genome Research (http://www-genome.wi.mit.edu/cgi-bin/contig/rhmapper.pl). SHGC or MIT framework markers linked to the subjected genes with a LOD score >6.0 were returned from the autoservers. Framework maps from SHGC, MIT, and Genethon (http://www.ceph.fr/quickmap.html) were used to infer the cytogenetic band locations corresponding to the RH mapping results.

Gene Expression in Different Tissues

In silico Northern Blot

For each entry in UniGene database (http://www.ncbi.nlm.nih.gov/ UniGene), beside the STS mapping information, cDNA source could also provide expression information.

Northern Blot

The MTN membranes used were from Clontech and the homemade membranes for hematopoietic cell lines were prepared according to the standard protocols (Sambrook et al. 1989). Probes were ^{32}P[dCTP] (DuPont) labeled with T7 quick primer (Amersham Pharmacia Biotech). Prehybridization and hybridization were performed with Expresshyb solution (Clontech). Membrane washing and autoradiography were carried out according to the standard protocol.

Screening of Gene Expression in Different Hematopoietic Cell Lines with Macroarray

Membrane Preparation

A total of 2430 unique cDNA clones corresponding to EST clusters identified in cord blood CD34+ HSPCs were PCR-amplified. The reactions were carried out using T3/T7 universal primer pairs in 50μl volume including rTaq and dNTPs (TaKaRa, Dalian, China) and on 9600 GeneAmp PCR system (Perkin Elmer) under the following conditions: 1 min at 94°C, 1 min at 54°C, and 2 min and 20 sec at 72°C for 30 cycles and finished by an extra 10 min at 72°C. The PCR products were quantitated, precipitated with 35μl isopropanol, washed with 70% ethanol, and redissolved in 10μl 1N NaOH. BioGrid 0.4-mm 384-pins total array system (TAS) arrayer (Bio-robotics) was used to spot cDNA PCR products onto 8 × 12 cm^2 nylon membranes (Amersham Pharmacia Biotech) with duplicate spots. The cDNA samples were immobilized with UV crosslinker after drying.

Preparation of the Probes

Total RNAs were isolated with TRIzol (Life Technologies) from hematopoietic cell lines NB4, HL60, U937, K562, and Jurkat cultured under conditions described previously (Zhu et al. 1995). mRNAs were then purified from 200 μg of total RNAs with Oligotex column (Qiagen). Probes were labeled while first-strand cDNA was synthesized. A mixture containing 2 μg mRNA, 3 μl oligo(dT) primer (0.5 μg/μl), and 2 μl random primers (0.5 μg/μl) was incubated at 68°C for 5 min. Then the following items were added: 10 μl of 5× RT buffer, 1 μl of 200 mmole/l NaPP, 33 mmole dNTPs (without dATP), 15 μl [α-^{33}P]dATP (DuPont) (10 mCi/ml), 1 unit of RNase inhibitor, 60 units of AMV Reverse transcriptase (Promega), and ddH_2O to a final volume of 50 μl. The reaction was performed at 42°C for 2 hr and terminated with 100°C water bath for 5 min.

Hybridization

The spotted membranes were rinsed with 6× SSC at room temperature for 5 min, and prehybridized in 20 ml of ExpressHyb hybridization solution added with sheared salmon sperm DNA to 100 μg/μl at 68°C for 3 hr in a roller bottle. Then hybridization was carried out overnight in 5 ml of solution (ExpressHyb hybridization solution, 100 μg/μl ssDNA) mixed with the denatured cDNA probes. Washing was performed under stringent conditions (Sambrook et al. 1989):

solution I (2× SSC, 0.1% SDS) at 65°C for 30 min twice and solution II (1× SSC, 0.5% SDS) at 65°C for 30 min once.

Signal Detection and Gene Expression Quantification

After stringent wash, the membranes were exposed to FLA-3000 system phosphor screens overnight, and measured with the attached ImageGauge program (Fuji). Fifteen no-sample areas were circled as background. The relative intensity for each gene was quantified after position and background correction. Only those signals with intensity value >10 could be considered as positive ones. The expression was considered as negative in the case where a negative value was recorded. The signal of housekeeping genes such as *GAPDH* or β-actin was chosen as reference for normalization, and the total signal amount of the membranes were also applied as reference. The ratio of each gene's signal to that of *GAPDH* on the same filter was chosen to compare the relative expression levels between cell lines (Pietu et al. 1999; Rhee et al. 1999).

ACKNOWLEDGMENTS

This work was supported in part by the Chinese High Tech Program (863), the Chinese National Key Program for Basic Research (973), the National Natural Science Foundation of China, Shanghai Commission for Science and Technology, and the Clyde Wu Foundation of SIH. The authors thank Dr. Charels Auffray in ERS 1984 CNRS of France and all members of SIH and of CHGC for their constructive discussion and encouragement.

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

Bonner, T.I., Oppermann, H., Seeburg, P., Kerby, S.B., Gunnell, M.A., Young, A.C. and Rapp, U.R. 1986. The complete coding sequence of the human raf oncogene and the corresponding structure of the c-raf-1 gene. *Nucleic Acids Res.* **14:** 1009–1015.

Boguski, M.S. and Schuler, G.D. 1995. ESTablishing a human transcript map. *Nat. Genet.* **10:** 369–371.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., et al. 1996. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37:** 327–336.

Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. 1998. New goals for the U.S. Human Genome Project: 1998–2003. *Science* **282:** 682–689.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–496.

Gu, J., Zhang, Q.H., Huang, Q.H., Ren, S.X., Wu, X.Y., Ye, M., Huang, C.H., Fu, G., Zhou, J., Niu, C., et al. 2000. Gene expression in CD34+ cells from normal bone marrow and leukemic origins. *Hematol. J.* **1:** 206–217.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K., et al. 2000. The DNA sequence of human chromosome 21. *Nature* **405:** 311–319.

He, K.L., Gu, B.W., Zhang, Q.H., Fu, G., Wu, J.S., Han, Z.G., Cao, W.J., Zhou, J., Mao, M., Liu, J.X., Chen, Z., and Chen, S.J. 1998. Application of radiation hybrid in gene mapping. *Sci. China (Ser. C)* **41:** 644–649.

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K., and Hood, L. 1997. Gene families: The taxonomy of protein paralogs and chimeras. *Science* **278:** 609–614.

Hofmann, K.., Bucher, P., Falquet, L., and Bairoch, A. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27:** 215–219.

Kozak, M. 1986. Point mutations define a sequence flanking the AUG initiator that modulates translation by eukaryotic ribosomes. *Cell* **44:** 283–292.

Mao, M., Fu, G., Wu, J.S., Zhang, Q.H., Zhou, J., Kan, L.X., Huang, Q.H., He, K.L., Gu, B.W., Han, Z.G., et al. 1998. Identification of genes expressed in human CD34+ hematopoietic stem/progenitor cells by expressed sequence tags and efficient full-length cDNA cloning. *Proc. Natl. Acad. Sci.* **95:** 8175–8180.

Marshall, E. 1999. Sequencers endorse plan for a draft in 1 year. *Science* **284:** 1439–1441.

———. 2000. Rival genome sequncers celebrate a milestone together. *Science* **288:** 2294–2295.

Morrison, S.J., Uchida, N., and Weissman, I.L. 1995. The biology of hematopoietic stem cells. *Annu. Rev. Cell Dev. Biol.* **11:** 35–71.

Morrison, S.J., Wright, D.E., Cheshier, S.H. and Weissman, I.L. 1997. Hematopoietic stem cells: Challenges to expectations. *Curr. Opin. Immunol.* **9:** 216–221.

Pietu, G., Mariage-Samson, R., Fayein, N.A., Matingou, C., Eveno, E., Houlgatte, R., Decraene, C., Vandenbrouck, Y., Tahi, F., Devignes, M.D., et al. 1999. The Genexpress IMAGE knowledge base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics. *Genome Res.* **9:** 195–209.

Rhee, C.H., Hess, K., Jabbur, J., Ruiz, M., Yang, Y., Chen, S., Chenchik, A., Fuller, G.N., and Zhang, W. 1999. cDNA expression array reveals heterogeneous gene expression profiles in three glioblastoma cell lines. *Oncogene* **18:** 2711–2717.

Russell, R.B., Saqi, M., Sayle, R.A., Bates, P.A., and Sternberg, M.J. 1997. Recognition of analogous protein folds: Analysis of sequence and structure conservation. *J. Mol. Biol.* **269:** 423–439.

Sambrook, J., Fritsch, E.F., and Maniatis, T. 1989. *Molecular cloning: A laboratory manual*. 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, H., White, E.R., Rodriguez-Tom, P., Aggarwal, A., Bajorek, E., et al. 1996. A gene map of the human genome. *Science* **274:** 540–546.

Shi, Y., Wang, W., Yourey, P.A., Gohari, S., Zukauskas, D., Zhang, J., Ruben, S., and Alderson, R.F. 1999. Computational EST database analysis identifies a novel member of the neuropoietic cytokine family. *Biochem. Biophys. Res. Commun.* **262:** 132–138.

Stewart, E.A., McKusick, K.B., Aggarwal, A., Bajorek, E., Brady, S., Chu, A., Fang, N., Hadley, D., Harris, M., Hussain, S., et al. 1997. An STS-based radiation hybrid map of the human genome. *Genome Res.* **7:** 422–433.

The C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode C. elegans: A platform for investigating biology. *Science* **282:** 2012–2018.

Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O., and Hunkapiller, M. 1998. Shotgun sequencing of the human genome. *Science* **280:** 1540–1542.

Winzeler, E.A. and Davis, R.W. 1997. Functional analysis of the yeast genome. *Curr. Opin. Genet. Dev.* **7:** 771–776.

Yan, Y., Smant, G., Stokkermans, J., Qin, L., Helder, J., Baum, T., Schots, A., and Davis, E. 1998. Genomic organization of four beta-1,4-endoglucanase genes in plant-parasitic cyst nematodes and its evolutionary implications. *Gene* **220:** 61–70.

Zhu, J., Shi, X.G., Zhu, H.Y., Tong, J.H., Wang, Z.Y., Naoe, T., Waxman, S., Chen, S.J., and Chen, Z. 1995. Effect of retinoic acid isomers on proliferation, differentiation and PML relocalization in the APL cell line NB4. *Leukemia* **9:** 302–309.

Received March 9, 2000; accepted in revised form July 19, 2000.