The pattern of gene expression in human CD34$^+$ stem/progenitor cells

Guolin Zhou*, Jianjun Chen*, Sanggyu Lee*, Terry Clark†, Janet D. Rowley*, and San Ming Wang*‡

Departments of *Medicine and †Computer Science, University of Chicago Medical Center, 5841 South Maryland Avenue, MC2115, Chicago, IL 60637

Contributed by Janet D. Rowley, October 4, 2001

We have analyzed the pattern of gene expression in human primary CD34$^+$ stem/progenitor cells. We identified 42,399 unique serial analysis of gene expression (SAGE) tags among 106,021 SAGE tags collected from 2.5 × 10^6 CD34$^+$ cells purified from bone marrow. Of these unique SAGE tags, 21,546 matched known expressed sequences, including 3,687 known genes, and 20,854 were novel without a match. The SAGE tags that matched known sequences tended to be at higher levels, whereas the novel SAGE tags tended to be at lower levels. By using the generation of longer sequences from SAGE tags for gene identification (GLGI) method, we identified the correct gene for 385 of 440 high-copy SAGE tags that matched multiple genes and we generated 198 novel 3′ expressed sequence tags from 138 high-copy novel SAGE tags. We observed that many different SAGE tags were derived from the same genes, reflecting the high heterogeneity of the 3′ untranslated region in the expressed genes. We compared the quantitative relationship for genes known to be important in hematopoiesis. The qualitative identification and quantitative measure for each known gene, expressed sequence tag, and novel SAGE tag provide a base for studying normal gene expression in hematopoietic stem/progenitor cells and for studying abnormal gene expression in hematopoietic diseases.

Hematopoietic stem cells have self-renewal ability. They can differentiate into different hematopoietic lineages, including myelomonocytic, megakaryocytic, lymphoid, and erythroid cells (1, 2). Hematopoietic stem cells have been widely used in the treatment of hematopoietic disorders such as leukemia (3). Recent data show that hematopoietic stem cells are highly plastic. Under certain conditions, they can differentiate into nonhematopoietic cells such as brain, liver, and cardiac cells (4–6). These features suggest that hematopoietic stem cells can potentially be used for the treatment of nonhematopoietic disorders such as neural and cardiac diseases.

Although much knowledge about hematopoietic stem cells has been gained, we still do not know much about the genetic mechanisms determining their development. We initiated a genome-scale analysis to characterize the pattern of gene expression in human hematopoietic stem cells. Our goal in this study was to answer the following questions: (*i*) How many genes are expressed in hematopoietic stem cells? (*ii*) Which genes are expressed in these cells? (*iii*) What is the level of expression of each gene in these cells? (*iv*) What is the quantitative relationship among the genes expressed in these cells? In this study, we used the serial analysis of gene expression (SAGE; ref. 7) technique as the tool for the analysis to provide the broadest coverage for expressed genes and to provide quantitative information for each identified gene. We also used the generation of longer sequences from SAGE tags for gene identification (GLGI) technique (8, 9) to confirm the genes corresponding to the SAGE tags. In this report, we present the results of this analysis.

Materials and Methods

Cell Purification. The CD34$^+$ cells were purchased from Poietics (Gaithersburg, MD) with Institutional Review Board approval and donor consent. Cells were isolated from mononuclear cells of human bone marrow through positive immunomagnetic selection (CD34 Progenitor Cell Isolation kit, Miltenyi Biotec, Auburn, CA). The purity of the isolated cells was determined by fluorescence-activated cell sorter analysis. The CD34$^+$ cells from three donors were pooled for the analysis.

SAGE Performance. SAGE was performed with our modified SAGE protocol (10), and the data were processed by use of our procedure (11).

GLGI Performance. The GLGI method was designed for two purposes. One is to identify the correct sequence from multiple sequences matched by a single SAGE tag, and the other is to generate a longer 3′ expressed sequence tag (EST) for a SAGE tag that does not match to known expressed sequences for further analysis (8). In the GLGI process, a SAGE tag sequence is used as the sense primer, and a universal antisense primer located at the 3′ end of cDNA is used as the antisense primer to amplify the original cDNA template from which the SAGE tag was derived. We developed the original GLGI method into a high-throughput GLGI procedure for large-scale conversion of SAGE tags into 3′ ESTs (9). In this study, we used the high-throughput procedure to identify the correct sequences for the multimatched SAGE tags with more than nine copies, and to convert novel SAGE tag sequences into 3′ ESTs for SAGE tags with more than four copies.

Results and Discussion

Experimental Design. It is a challenge to perform a genome-level analysis of gene expression in human hematopoietic stem cells because of the rarity of stem cells in human bone marrow. We developed two strategies to overcome this obstacle. The first strategy was to use CD34$^+$ cells for the analysis. CD34$^+$ cells represent hematopoietic stem cells and progenitors for myeloid, erythroid, megakaryoid, and lymphoid cells. Normal bone marrow contains about 1–2% CD34$^+$ cells (12, 13). This amount will provide a minimal number of cells for analysis. The second strategy is to modify the standard SAGE protocol to decrease the initial amount of mRNA required for SAGE analysis. Our modified SAGE protocol needs 100-fold less mRNA compared with the standard SAGE protocol. This modification enabled us to perform the entire SAGE analysis by using 2.5 × 10^6 CD34$^+$ cells. The purity of the CD34$^+$ cells was 96.4, 98.7, and 97.3% in the three samples as measured by fluorescence-activated cell sorter analysis (Fig. 1). We used the CD34$^+$ cells directly for analysis without any *in vitro* culture, to reflect the original pattern of gene expression in the cells.

Distribution of SAGE Tags in CD34$^+$ Cells. We collected a total of 106,021 SAGE tags, from which we identified 42,399 unique SAGE tags. These unique SAGE tags were matched to the SAGE database for gene identification (Tables 1 and 6–9, which are published as supporting information on the PNAS web site,

Abbreviations: SAGE, serial analysis of gene expression; GLGI, generation of longer cDNA fragments from SAGE tags for gene identification; EST, expressed sequence tag; UTR, untranslated region.

‡To whom reprint requests should be addressed. E-mail: swang1@midway.uchicago.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

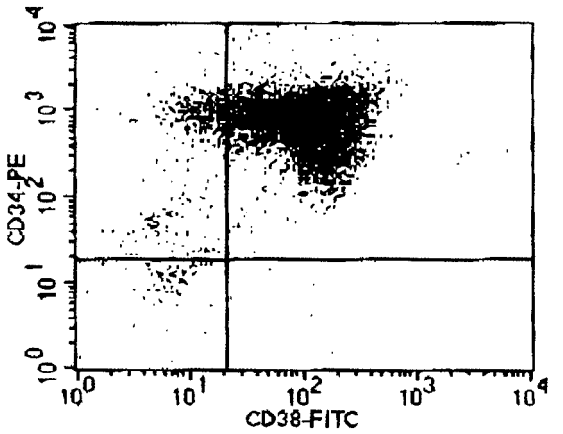

Fig. 1. Purity of $CD34^+$ cells. The $CD34^+$ cells were isolated from bone marrow with CD34 immunomagnetic beads. The purity of isolated cells was above 96%. This figure was obtained from one of the three samples used for the analysis.

www.pnas.org). We observed three features of the SAGE tag distribution.

(i) *The quantitative distribution of SAGE tags.* In $CD34^+$ cells, a few genes are expressed at high levels, and most genes are expressed at low levels. Only 91 (0.2%) SAGE tags were present in more than 100 copies; 9,085 (21.4%) SAGE tags had between 9 and 2 copies, whereas 32,453 (76.5%) SAGE tags were present as a single copy. This distribution is consistent with that observed in other mature somatic cell types (14), indicating the universal pattern of quantitative distribution of expressed genes between stem/progenitor cells and mature cells.

(ii) *The quantitative distribution of matched and novel SAGE tags.* Of the unique SAGE tags, 21,546 (50.9%) matched existing known expressed sequences and 20,854 (49.1%) were novel SAGE tags without matches. The distribution of matched tags and novel SAGE tags shows a reciprocal relationship with their copy numbers. The matched tags tended to be the ones with more copies, whereas the novel SAGE tags tended to be the ones with fewer copies. This pattern supports our previous observation that a large number of genes in the human genome have not been identified (11), particularly for the genes expressed at low levels, and at different stages of development such as those in $CD34^+$ cells.

(iii) *The rate of multiple matches.* By using the matched UniGene cluster as the measure, we observed that 34% of matched SAGE tags matched more than one sequence located in different UniGene clusters. The low specificity of a SAGE tag representing a gene is largely caused by the short length of the SAGE tag sequence (15, 16). The distribution of these multiple-matched SAGE tags also paralleled their copy numbers. For example, 61 of 85 SAGE tags (71.8%) with more than 99 copies matched multiple sequences. Thus, it is highly unreliable to identify the correct genes for these SAGE tags based solely on a database search.

Identification of Correct Genes for SAGE Tags with Multiple Matches. Because the SAGE tags with multiple matches tend to be the ones with more copies, we tried to identify the correct genes for 440 SAGE tags with more than 9 copies with multiple matches. By using the high-throughput GLGI method, we converted these 440 SAGE tags into 3′ ESTs and we used these longer sequences to search databases to identify their corresponding genes. We identified the correct gene for 385 (88%) of these 440 SAGE tags (Table 10, which is published as supporting information on the PNAS web site). Similar to other cell types, many housekeeping genes were among the highly expressed genes, including 55 ribosomal proteins. However, many genes with specified function were also expressed at high levels, such as *v-fos*, *TNF*, and *Myeloperoxidase*. Interestingly, there were many functional unknown genes among these highly expressed genes, including 77 ESTs, 13 hypothetical genes, and 10 KIAA protein genes. The high-level expression of these genes suggests their functional importance in the development of hematopoietic stem cell. Table 2 shows the top 60 genes after subtraction of 41 ribosomal protein genes and removal of 4 SAGE tags with poly(A) nucleotide. All of these 60 genes were identified from multimatched SAGE tags.

Conversion of Novel SAGE Tags into Novel 3′ ESTs. The novel SAGE tags account for half of the unique SAGE tags detected in this analysis. The question arises whether these novel SAGE tags represent unidentified novel genes expressed in $CD34^+$ cells. We used the GLGI method to convert 219 novel SAGE tags present in more than 4 copies into 3′ ESTs, and we matched these longer 3′ ESTs to the database. With 85% of sequence homology as the cut-off value to distinguish the known or novel sequences, a total of 198 sequences generated from 138 novel SAGE tags was confirmed to be novel 3′ ESTs. These ESTs range in length from 8 to 454 bp (mean = 152 bp) (Tables 3 and 11, which is published as supporting information on the PNAS web site). Sequences from 32 novel tags matched known expressed sequences, and reactions for 49 novel tags did not generate qualified sequences. This result, as well as our earlier study on $CD15^+$ cells (11), indicates that the novel SAGE tags we have identified truly represent a large number of novel genes. The conversion of novel SAGE tags into 3′ ESTs provides an efficient way to identify novel genes on a large scale. Our strategy becomes especially important for novel gene identification, in view of the reports that the number of genes in the human genome has been seriously underestimated (17, 18).

Known Genes Identified in $CD34^+$ Cells. Based on the UniGene clusters single-matched by SAGE tags, we identified 3,687 known genes expressed in $CD34^+$ cells (Table 12, which is published as supporting information on the PNAS web site). This is the largest number of known genes identified in human hematopoietic $CD34^+$

Table 1. Distribution of 42,400 unique SAGE tags in $CD34^+$ stem cells

	Copy number of SAGE tags					
	>100	99 to 10	9 to 5	4 to 2	1	Total
Total unique tags*	91 (0.2)	771 (1.8)	1,328 (3.1)	7,757 (18.3)	32,453 (76.5)	42,399 (100)
Total novel tags†	**6 (6.5)**§	**85 (11.0)**	**128 (9.6)**	2,123 (27.3)	1,851 (57.0)	20,853 (49.1)
Total matched tags	85 (93.4)	686 (89.0)	1,200 (90.4)	5,634 (72.6)	13,941 (43.0)	21,546 (50.9)
Single-matched tags‡	24 (28.2)	307 (44.8)	657 (54.8)	3,360 (59.6)	9,864 (70.8)	14,212 (66.0)
Multiple-matched tags	**61 (71.8)**	**379 (55.2)**	543 (45.3)	2,274 (40.4)	4,077 (29.2)	7,334 (34.0)

*The number in parentheses is the percentage of tags in the total unique tag set.
†The number in parentheses is the percentage of the tags in the total tags of each subgroup.
‡The number in parentheses is the percentage of tags within the total matched tags.
§The number in bold represents the tags analyzed by GLGI.

Table 2. The top 60 genes expressed in CD34$^+$ cells

No.	SAGE tag	Copy	Matched UniGene cluster	GLGI confirmation	Gene
1	TGTGTTGAGA	1711	6	Hs.181165(X03558)	Translation elongation factor 1 alpha 1
2	CCTGTAATCC	516	430	Hs.106004(AK025503)	Hypothetical protein FLJ22347
3	GTGAAACCCC	471	325	Hs.225030(AI246594)	EST
4	TGGGCAAAGC	377	2	Hs.2186(BG222874)	Eukaryotic translation elongation factor 1 gamma
5	CCACTGCACT	331	175	Hs.293521(BE205895)	EST
6	GCCTCAGTTC	256	3	AL571622	EST
7	AATGGATGAA	253	3	BG223362	EST
8	AGCCCTACAA	253	2	BG214720	EST
9	GTGAAACCCT	239	155	Hs.184376(BF692242)	Synaptosomal-associated protein, 23kD
10	CCAGAGAACT	165	2	Hs.6975(AW582859)	PRO1073 protein
11	TTTTTGATAA	156	2	Hs.181165(BG770933)	Eukaryotic translation elongation factor 1 alpha 1
12	TAGGTTGTCT	148	5	Hs.279860(AW167109)	Tumor protein, translationally-controlled 1
13	TTCATACACC	147	3	Hs.297184(AW193452)	EST
14	CACCTAATTG	140	2	AL583021	EST
15	TGATTTCACT	138	4	AL583322	EST
16	GCAAGCCAAC	132	2	BG687451	EST
17	CATTTGTAAT	131	2	BE890421	EST
18	ATTGTTTATG	127	2	Hs.181163(BF593317)	High-mobility group protein 17
19	CTCATAAGGA	126	3	BG385900	EST
20	GCTCCCCTTT	126	2	Hs.1817(AV736453)	Myeloperoxidase
21	AGGTCAGGAG	122	47	H77590	EST
22	ACCCTTGGCC	122	3	AI880722	EST
23	AAGGTGGAGG	118	3	BG056715	EST
24	TGTAATCAAT	117	3	Hs.249495(BG655713)	Heterogeneous nuclear ribonucleoprotein A1
25	TTGGCCAGGC	109	96	AI078409	EST
26	TCACAAGCAA	109	3	Hs.32916 (BG271651)	Nascent-polypeptide-associated complex a polypeptide
27	GGGCATCTCT	102	2	Hs.76807(AU157203)	HLA-DRA
28	CCTGTAGTCC	100	116	Hs.314307(AA749235)	EST
29	TACCCTAAAA	100	49	BG271479	EST
30	GTGTTAACCA	98	2	Hs.85301(BF941019)	Calcium binding protein P22
31	GTTCGTGCCA	96	3	Hs.179666(BG236685)	Uncharacterized hypothalamus protein HSMNP1
32	CCATTGCACT	91	59	AU117661	EST
33	CTCATAGCAG	84	2	Hs.279860(BG654607)	Tumor protein, translationally-controlled 1
34	ACTTTTTCAA	76	46	BG099326	EST
35	AAAAGAAACT	76	4	Hs.172182(BG744897)	Poly(A)-binding protein, cytoplasmic 1
36	GCTTTATTTG	75	2	Hs.288061(AA554747)	Actin, beta
37	GCCTTCCAAT	74	3	Hs.76053(BF941985)	DEAD/H box polypeptide 5
38	TACCATCAAT	72	6	Hs.169476(BG370213)	Glyceraldehyde-3-phosphate dehydrogenase
39	ACTCCAAAAA	72	5	BG236559	EST
40	GCAAAACCCC	70	44	BF931620	EST
41	GCGAAACCCC	68	39	Hs.269899(AI079278)	EST
42	GCATTTAAAT	64	4	Hs.275959(BG655489)	Eukaryotic translation elongation factor 1 beta 2
43	GTCTGGGGCT	63	3	Hs.75725(BF591438)	Transgelin 2
44	ATTTGTCCCA	61	2	Hs.139800(AU160425)	High-mobility group protein isoforms I and Y
45	TCTGCTAAAG	58	4	Hs.274472(BF434300)	High-mobility group protein 1
46	TGTACCTGTA	58	3	Hs.278242(BG222897)	Hypothetical protein MGC12992
47	AACCCGGGAG	56	19	AA428792	EST
48	GTTCCCTGGC	55	3	Hs.177415(BG271519)	FBR-MuSV
49	CCTGTAATCT	53	44	Hs.35088(BF515942)	EST
50	GGCTTTACCC	51	2	Hs.119140(BF432256)	Eukaryotic translation initiation factor 5A
51	CCTAGCTGGA	50	4	Hs.182937(BG655492)	Peptidylprolyl isomerase A (cyclophilin A)
52	TTGGCTTTTC	50	2	Hs.41569(BF475411)	Phosphatidic acid phosphatase type 2A
53	CCTATAATCC	49	73	Hs.25328(AL048825)	EST
54	GATGCTGCCA	49	3	Hs.129914(L21756)	AML1 oncogene
55	ATTTGAGAAG	48	5	Hs.169921(BE621880)	General transcription factor II
56	TAGAAAGGCA	48	2	Hs.78909(W37407)	Butyrate response factor 2
57	TGTGTTAAGA	47	4	Hs.288036(AI804500)	TRNA isopentenylpyrophosphate transferase
58	GTGGCTCACA	46	44	Hs.120769(AI982685)	Homo sapiens cDNA FLJ20463 fis
59	TTGGTCAGGC	45	34	Hs.12094(AA650333)	Hypothetical protein
60	GTGCACTGAG	45	5	Hs.181244(D32129)	Major histocompatibility complex, class I, A

The list was generated from the top expressed genes after removing 40 ribosomal protein genes and 4 SAGE tags with poly(A) nucleotides.

Table 3. Examples of novel 3′ ESTs from novel SAGE tags

Copy	SAGE tag	GenBank accession no.	Sequence length, bp
114	GAGCGGCGCT	BI094690	38
		BI094691	257
		BI388635	226
		BI388636	47
104	GTGCCACGGG	BI094692	127
97	TGGCGTACGG	BI094693	204
83	TAAGCGGCGC	BI388635	28
79	AGAGCGGCGC	BI388636	59
75	AAAGCGGCGC	BI094694	154
59	GTGCCCACGG	BI388637	413
41	CCGACGGGCG	BI094695	344
		BI094696	344
34	GATGCCCCCC	BI388638	258
32	GTGCCCCGGG	BI094697	300
		BI388639	90
30	CTGAGCGGCG	BI094698	133
		BI388640	104
29	CCACTTCTGG	BI094699	132
27	GTGACCAAGG	BI094700	129
27	GTGAAGCAGT	BI094701	359
26	GGGACCACCG	BI094702	113
24	CAACACCACA	BI388641	409
		BI388642	215
23	AGCTCTGTAG	BI094703	233
23	GGTCAGTCGG	BI094704	211
		BI094705	137
22	AGCGGCGCTC	BI388643	37
21	ACCCCGCCGG	BI388644	130
20	GTGCCCAGGG	BI388645	163

Table 3. (continued)

Copy	SAGE tag	GenBank accession no.	Sequence length, bp
19	TAGAGCGGCG	BI388646	181
19	CCAGAGGCTG	BI388647	114
18	TGCACCGTTT	BI388648	143
		BI388649	157
17	GGACCACGGG	BI094706	127
16	ACTCCTGAAC	BI094707	301
16	CTATAGCGGC	BI094819	281
15	AGCTCTTCCT	BI094820	109
15	TTACCCACAC	BI094821	65
15	CAAGCGGCGC	BI094822	341
14	GGGAAGCAGA	BI094823	90
		BI094824	35
14	ATCAAAGGTG	BI094825	428
14	AGCACCTTCA	BI094826	57
		BI094827	210
14	GTTCCAGCCG	BI094828	16
14	CTAAGCGGGG	BI094829	45
		BI094830	119
13	GGTGACCACG	BI094831	129
13	TAGGTTGCTA	BI094832	172
		BI094833	82
13	GTTGTCCTAC	BI094834	49
12	ATGGCGCCTC	BI094835	85
12	ACCCGCCCGG	BI388650	168
12	TCGCCGCGAC	BI388651	98
11	ACCCAGGGAG	BI094836	82
11	GCACGTGTCT	BI094837	329
11	AACGAAAAAA	BI094838	16
11	GTAAAGCCAG	BI094708	68

cells (19–23). Considering the identification of 42,399 unique SAGE tags and for the matches of 21,546 unique SAGE tags to known expressed sequences, the 3,687 known genes account for only 9% of the total unique SAGE tags and 17.6% of the matched SAGE tags. These data indicate that most of the genes expressed in $CD34^+$ cells have not been identified or studied. Clarifying the function of such a large number of genes expressed in stem/progenitor cells is a serious challenge for stem cell biology.

Heterogeneity of 3′ Untranslated Region (UTR) Sequences. When matching SAGE tags with expressed sequences, we frequently observed that different SAGE tags matched different sequences located within the same UniGene cluster (Table 4). One assumption would be that, if a SAGE tag matched sequences upstream of the most 3′ tag, this tag was most likely to be derived from partially digested cDNA templates in the process of SAGE library construction. Two lines of evidence fail to support this assumption. (*i*) The SAGE tag database was constructed through extracting the SAGE tag sequences after the last CATG site from expressed transcripts in the database (24). The match between an experimental SAGE tag and a tag in the SAGE database indicates the existence of an expressed transcript for this SAGE tag. (*ii*) We converted more than 1,000 SAGE tags into 3′ ESTs through the GLGI method. Examination of these 3′ ESTs shows that 97% of them do not have internal CATG, which is the restriction sequence of *Nla*III used for SAGE library construction. This result strongly indicates that *Nla*III restriction digestion of cDNA templates is very efficient. Because SAGE tags are located in the 3′ part of transcripts, the identification of different SAGE tags that match different sequences located in one UniGene cluster reflects the presence of transcripts from the same gene with different 3′ UTRs. The 3′ UTR plays important roles in regulating the function of expressed genes, such as mRNA stability and translational efficiency (25–29). Therefore, analyzing the heterogeneity of the 3′ UTR through SAGE tags will provide information for understanding the relationship between 3′ UTR structure and the function of the genes expressed during hematopoiesis.

Genes Known to Be Important for Hematopoiesis. We analyzed the genes known to be important for hematopoiesis (Tables 5 and 13, which is published as supporting information on the PNAS web site). Many of these genes were identified in this study. Because SAGE provides a simultaneous quantitation for the expressed genes in a sample, the levels of expression among different genes can be compared directly. For example, seven *HOX* genes were identified from the *HOX-A*, *B*, and *C* clusters, including *A3*, *A5*, *A9*, *B2*, *B7*, *B13*, and *C9* (30). All of these *HOX* genes were expressed at low levels. Some genes known to play roles in hematopoiesis were not detected in this analysis, such as *PU.1* (31) and *SCF* (stem cell factor; ref. 32). This discrepancy may arise because of the difference in methodologies used for the analyses. SAGE analyzes gene expression in a "horizontal" way. That is, all of the transcripts expressed from different genes were simultaneously identified and quantified. In other studies, genes were identified based on the reverse-transcription (RT)-PCR method. RT-PCR analyzes gene expression in a "vertical" manner. That is, it can identify genes expressed at very low levels through millionfold amplification. Other explanations may be related to differences in the cell types used or differences in culture conditions. For example, a gene can be expressed in transformed cell lines or cells treated with various cytokines in *in vitro* culture conditions but may not be expressed in *in vivo* physiologic conditions.

In summary, the data generated from this study provide an overview of the pattern of gene expression in normal human $CD34^+$ stem/progenitor cells. With this information in hand, we are in a position to identify the genes important in hematopoietic

Table 4. Examples of transcripts with different 3′ UTRs detected by SAGE tags

Gene	Tag	Gene	Tag
Transcriptional/translational factors		DNA structure	
C-fos (Hs.25647)	ACATCAAAAA	Chromatin assembly factor-I p150 subunit (Hs.79018)	GACGGCTTCC
	CTTTCAGACT		ACCCCTGAGA
	ACTGTAATTG	High mobility group box (SSRP1) (Hs.79162)	TAAGGCCAGG
C-jun (Hs.78465)	AGGAACCGCA		AAAGAATATG
	GCGAGGGGAG	Histone acetyltransferase (MORF) (Hs.27590)	TGGGCGGGTC
	GGGGGAGGGC		CTAAAGTTTT
	CCTTTGTAAG		TAGTACAATG
	CTAACGCAGC	Nucleosome assembly protein 2 (Hs.78103)	CTTTCTCAGT
CREB (Hs.79194)	CCCGCTTCGC		AGGACGGGCT
	ATTCCTGGCG		ACGCAGGCGC
Forkhead protein (FKHR) (Hs.170133)	GATGGAGATA	Scaffold attachment factor A (Hs.103804)	TTTAGTAACC
	AAGTCTAACA		AGTTTTATAG
	AGGTCAGTAG		AGGTCAGCAT
Methyl-CpG binding protein MBD4 (Hs.35947)	TTTTGAGATT		CAACTGTGAG
	TGGTGAGACC		GAGCCCGGGA
NF-kappa-B (Hs.83428)	GTTACAATCA		TTTGGACCGG
	TATTACTTTA		GAGTCAGCAT
	ATTATGGGCA		AATTGCATTA
Poly(A) binding protein II (Hs.117176)	TGTCACTAAG		AATCTGGTTG
	AATAAAGTTG	Telomeric repeat binding factor (TRF1) (Hs.194562)	ACAGTATTTT
	GTATTCCCCT		GGGCAAAATT
	ATCACCGGGC		GAACCCAGCA
Serum response factor (Hs.155321)	GCGCTCGATG		
	GTCACAGTCC	Signal transduction	
	TTTCAGCCAG	Calmodulin 1 (CALM1) (Hs.177656)	GAGGCTACGA
	AATTATACAC		TAGGTCTAGT
Splicing factor SF3a60 (Hs.77897)	GTAGAGTTGG		TACACTGGCC
	CTGGCAGATT		ACTTGGAGCC
SWI/SNF complex 155-kDa subunit (Hs.172280)	TCTCTCAGTG		ACAAACTTAG
	TCTTTGATCT	CDC2L5 protein kinase (Hs.59498)	TAGAAGTGTT
	TTTCTCAGTG		CTGCCTGAAG
TFIID subunit TAFII55 (TAFII55) (Hs.155188)	AATGTTAGGT	ERK activator kinase (MEK1) (Hs.3446)	TGTTCTTTAT
	ATGAGCTTCG		ATAGCTGGGG
TIF-IA (Hs.110103)	GACCATTTTA	G protein gamma-10 subunit (Hs.79126)	AGAAAAGAAC
	CCATATGAAT		AAGAAAATAT
		IL-1 receptor-associated kinase (Hs.182018)	GGAACGATTA
Cell cycle regulators			CCCCCGTGAA
CDC2 (Hs.184572)	CCAAAATTTG	K-ras (Hs.184050)	AACTGTACTA
	GAAAGAAACT		CTATCCAGTA
CDC21 (Hs.154443)	GCTACTATTA	Protein tyrosine kinase 6 (PTK6) (Hs.51133)	AGTGTGCGTG
	CTAAAGTAAG		GTCATTTTAA
Checkpoint suppressor 1 (Hs.211773)	GGCGACAGGG	Rab11a GTPase (Hs.75618)	CTATGGCTTC
	GAAGAAGTTC		TTCCACCAAC
	GAGAAGAAAT		CAGTAGAGTC
Cyclin protein (Hs.78996)	GAAACACAGA		CTTGTGGGCA
	TCGATAAAGA		CTCTGTAATT
Cyclin I (CYC1) (Hs.79933)	CAAAATGAGG	Serine/threonine kinase (KDS) (Hs.12040)	CAAAAAAATA
	CCACTTTTAA		ATAACTGTCA
	AATTTTAATT		CCAACAACTC
	CAAAAATCAG		GAGAGCCTCA
	ATGACCCCCG	Serine/threonine protein kinase MASK (Hs.23643)	GTAATAAATT
	CAAAATCAGG		AGTTGCAAAA
	TTATAACTGA		TTATGGGTTA
	ACCTTAAGAG		

stem cells, to understand the regulatory network of self-renewal and differentiation of hematopoietic stem cells into different lineages, and to identify the genes whose expression are abnormal in hematopoietic diseases such as leukemia.

This work was supported by National Institutes of Health Grants CA78862-01 (to J.D.R. and S.M.W.) and CA42557 (to J.D.R.), American Cancer Society Grant IRG-41-40 (to S.M.W.), and by the G. Harold and Lelia Y. Mathers Foundation (S.M.W.).

1. Morrison, S. J., Wright, D. E., Cheshier, S. H., Weissman, I. L. (1997) *Curr. Opin. Immunol.* **9,** 216–221.
2. Orkin, S. H. (2001) *Stem Cell Biology*, eds. Marshak, D. R., Gardner, R. & Gottlieb, D. (Cold Spring Harbor Lab. Press, Plainview, New York), pp. 289–301.
3. Weissman, I. L. (2000) *Science* **287,** 1442–1446.
4. Brazelton, T. R., Rossi, F. M., Keshet, G. I. & Blau, H. M. (2000) *Science* **290,** 1775–1779.
5. McDonnell, W. M. (2000) *Hepatology* **32,** 1181.
6. Orlic, D., Kajstura, J., Chimenti, S., Jakoniuk, I., Anderson, S. M., Li, B., Pickel, J., McKay, R., Nadal-Ginard, B., Bodine, D. M., *et al.* (2001) *Nature (London)* **410,** 701–705.
7. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995) *Science* **270,** 484–487.
8. Chen, J., Rowley, J. D. & Wang, S. M. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 349–353.
9. Chen, J., Lee, S., Zhou, G. & Wang, S. M. (2001) *Genes Chromosomes Cancer*, in press.
10. Lee, S., Chen, J., Zhou, G. & Wang, S. M. (2001) *BioTechniques* **31,** 348–354.

Table 5. Examples of genes known to be important for hematopoiesis

Gene	UniGene ID	SAGE tag	Copy no.
Cathepsin B	Hs.249982	GAAAAGGACA	4
		GGGGTAACCA	2
		ATTCTTTAAT	1
		TGGGTAAGCC	1
		AGGGGAAGGG	1
		ATTAGCAGAG	1
Cathepsin C	Hs.10029	CAAAATGCAA	3
		CTGGCAACCT	2
		CTATATTTTT	2
		CACCCACCCA	1
Cathepsin D (lysosomal aspartyl protease)	Hs.79572	TATTGGCCTG	2
		GCAGCTCAGG	2
		ATCTCAAAGA	2
		TTAAGCATAA	1
Cathepsin S	Hs.181301	ACCAGTGAAG	2
		GTGGAGCCCC	1
C/EBP alpha	Hs.76171	GGGGGTGAAG	1
C/EBP delta	Hs.76722	CTCACTTTTT	2
		CTCCCTTTTT	1
CD33 antigen (gp67)	Hs.83731	GAAAACACCA	1
CD34 antigen	Hs.85289	GCTTCCTCCT	2
		GGACCAGGGT	1
		GTCCTGCCTA	1
CD8 antigen, alpha polypeptide (p32)	Hs.85258	ATTATTATTT	1
M-CSF1	Hs.173894	GTATCCAGCT	2
G-CSF3 receptor	Hs.2175	CTCCATCCAG	7
Friend of GATA2	Hs.106309	GCTTCTATTT	2
GATA-binding protein 2	Hs.760	GACAGTTGTT	1
Hemoglobin, beta	Hs.155376	GCAAAGAAAG	1
		GCAAGAAAGT	1
Homeo box A3	Hs.248074	GACTATGGGG	1
Homeo box A5	Hs.37034	TGCGTGGAAG	2
Homeo box A9	Hs.127428	TACCTCACCA	1
Homeo box B13	Hs.66731	ACTCCCTGTT	1
Homeo box B2	Hs.2733	AAGCACAAGC	1
Homeo box B7	Hs.819	CTTGCAGCCT	1
Homeo box C9	Hs.40408	CCGCGGGCTG	1
Jun D proto-oncogene	Hs.2780	ACCCCCCGGC	1
Macrophage migration inhibitory factor	Hs.73798	AACGCCGGCA	1
		AACGCGGTCA	1
Mesenchymal stem cell protein DSCD75	Hs.25237	GGAAAGCTGC	1
		GGAAAGCTTG	1
Myeloid differentiation primary response gene	Hs.82116	TTTTGTACGC	12
Myeloid leukemia factor 2	Hs.79026	CATTGAAGGG	12
		GCAGGAGTAG	3
		ACAGCTGGAG	1
Myeloperoxidase	Hs.1817	TATGTGCGAA	2
Retinoid X receptor, alpha	Hs.288688	CAGATGGACA	1
		CCCGGCCGGC	1
Retinoid X receptor, beta	Hs.79372	ATTTTTGCCC	6

11. Lee, S., Zhou, G., Clark, T., Chen, J., Rowley, J. D. & Wang, S. M. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 3340–3345.
12. Krause, D. S., Fackler, M. J., Civin, C. I. & May, W. S. (1996) *Blood* **87,** 1–13.
13. D'Arena, G., Musto, P., Cascavilla, N., Di, Giorgio. G., Zendoli, F. & Carotenuto, M. (1996) *Haematologica* **81,** 404–409.
14. Velculescu, V. E., Madden, S. L., Zhang, L., Lash, A. E., Yu, J., Rago, C., Lal, A., Wang, C. J., Beaudry, G. A., Ciriello, K. M., *et al.* (1999) *Nat. Genet.* **23,** 387–388.
15. Lee, S., Clark, T., Chen, J., Zhou, G., Scott, L. R., Rowley, J. D. & Wang, S. M. (2001) *Genomics*, in press.
16. Clark, T., Lee, S., Scott, L. R. & Wang, S. M. (2001) *J. Comput. Biol.,* in press.
17. Wright, F. A., Lemon, W. J., Zhao, W. D., Sears, R., Zhuo, D., Wang, J. P., Yang, H. Y., Baer, T., Stredney, D., Spitzner, J., *et al.* (2001) *Genome Biol.* **2,** RESEARCH 0025.
18. Hogenesch, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G., Cooke, M. P., *et al.* (2001) *Cell* **106,** 413–415.
19. Yang, Y., Peterson, K. R., Stamatoyannopoulos, G. & Papayannopoulou, T. (1996) *Exp. Hematol.* **24,** 605–612.
20. Claudio, J. O., Liew, C. C., Dempsey, A. A., Cukerman, E., Stewart, A. K. & Na, E. (1998) *Genomics* **50,** 44–52.
21. 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) *Proc. Natl. Acad. Sci. USA* **95,** 8175–8180.
22. Zhang, Q. H., Ye, M., Wu, X. Y., Ren, S. X., Zhao, M., Zhao, C. J., Fu, G., Shen, Y., Fan, H. Y., Lu, G., *et al.* (2000) *Genome Res.* **10,** 1546–1560.
23. 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) *Hematol. J.* **1,** 206–217.
24. Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., *et al.* (1999) *Cancer Res.* **59,** 5403–5407.
25. Sonenberg, N. (1994) *Curr. Opin. Genet. Dev.* **4,** 310–315.
26. Decker, C. J. & Parker, R. (1995) *Curr. Opin. Cell Biol.* **7,** 386–392.
27. Wickens, M., Anderson, P. & Jackson, R. J. (1997) *Curr. Opin. Genet. Dev.* **7,** 220–232.
28. Gallie, D. R. (1998) *Gene* **216,** 1–11.
29. Varani, G. (2001) *Proc. Natl. Acad. Sci. USA.* **98,** 4288–4289.
30. Buske, C. & Humphries, R. K. (2000) *Int. J. Hematol.* **71,** 301–308.
31. Klemsz, M. J., McKercher, S. R., Celada, A., Van Beveren, C. & Maki, R. A. (1990) *Cell* **61,** 113–124.
32. Martin, F. H., Suggs, S. V., Langley, K. E., Lu, H. S., Ting, J., Okino, K. H., Morris, C. F., McNiece, I. K., Jacobsen, F. W., Mendiaz, E. A., *et al.* (1990) *Cell* **63,** 203–211.