The pattern of gene expression in human CD15$^+$ myeloid progenitor cells

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

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

Contributed by Janet D. Rowley, January 8, 2001

We performed a genome-wide analysis of gene expression in primary human CD15$^+$ myeloid progenitor cells. By using the serial analysis of gene expression (SAGE) technique, we obtained quantitative information for the expression of 37,519 unique SAGE-tag sequences. Of these unique tags, (*i*) 25% were detected at high and intermediate levels, whereas 75% were present as single copies, (*ii*) 53% of the tags matched known expressed sequences, 34% of which were matched to more than one known expressed sequence, and (*iii*) 47% of the tags had no matches and represent potentially novel genes. The correct genes were confirmed by application of the generation of longer cDNA fragments from SAGE tags for gene identification (GLGI) technique for high-copy tags with multiple matches. A set of genes known to be important in myeloid differentiation were expressed at various levels and used different spliced forms. This study provides a normal baseline for comparison of gene expression in myeloid diseases. The strategy of using SAGE and GLGI techniques in this study has broad applications to the genome-wide identification of expressed genes.

Hematopoietic cells undergo an extensive process of differentiation starting from stem cells and going through commitment to a particular lineage and maturation. A number of genes expressed in a time-restricted manner in the genome control the process of differentiation. Alterations of regulatory pathways in pathologic conditions can change normal gene expression resulting in cellular transformation. Knowledge of the pattern of gene expression in normal hematopoietic cells is necessary for an understanding of the genetic regulation of normal hematopoietic differentiation and thus leads to insights regarding the consequences of gene alteration in hematopoietic diseases.

Myeloid cells originate from stem cells, become committed granulocyte-monocyte stem cells, and differentiate to myeloblasts, promyelocytes, myelocytes, metamyelocytes, and segmented neutrophils (1). Deregulation of myeloid differentiation is associated with many hematologic diseases such as myeloid leukemia (2). To understand gene expression during myeloid differentiation, we performed a systematic survey for characterizing the pattern of normal gene expression in primary human CD15$^+$ myeloid progenitor cells through the serial analysis of gene expression (SAGE) (3) and generation of longer cDNA fragments from SAGE tags for gene identification (GLGI) technique (4). The data reveal many particular features of gene expression in myeloid progenitor cells, and they provide a reference for further studies of various diseases within this lineage including myeloid leukemia.

Materials and Methods

Isolation of Myeloid Progenitor Cells. Human bone-marrow mononuclear cells were obtained from the Poietics Company (Gaithersburg, MD). These cells were isolated from bone marrow with Ficoll/Paque solution and stored in liquid nitrogen. Cells from three donors were thawed at 37°C, pooled, and used immediately for the isolation of myeloid progenitor cells with CD15 magnetic beads (Dynal, Oslo, Norway) according to the manufacturer's protocol. The purity of isolated cells was greater than 95% as confirmed through fluorescence-activated cell sorter analysis.

Synthesis of cDNA. The cells isolated by CD15 magnetic beads were lysed directly with TRIzol reagent (Life Technologies, Rockville, MD) for isolation of total RNA according to the manufacturer's instructions. mRNA was purified from 5 μg of total RNA with oligo(dT)$_{25}$ beads (Dynal) following the manufacturer's protocol. cDNA was synthesized with a cDNA synthesis kit (Life Technologies) according to the manufacturer's instructions with the following exceptions. (*i*) To prevent the inclusion of poly(dA/dT) sequences in the cDNA templates, 5′-biotinylated and 3′-anchored oligo(dT) primers were used for reverse transcription (5): 5′-biotin-ATCTAGAGCGGCCGC-T$_{16}$A-3′; 5′-biotin-ATCTAGAGCGGCCGCT$_{16}$G-3′; 5′-biotin-ATCTAGAGCGGCCGCT$_{16}$CA-3′; 5′-biotin-ATCTAGAG-CGGCCGCT$_{16}$CG-3′; 5′-biotin-ATCTAGAGCGGCCG-CT$_{16}$CC-3′. (*ii*) To increase the cDNA yield, the reaction for first-strand synthesis was repeated three times under the following conditions: the primary reaction was run at 37°C for 30 min, heated at 60°C for 3 min, and run again at 37°C with the addition of 2 μl of Maloney murine leukemia virus reverse transcriptase.

SAGE Procedures. SAGE was performed according to the SAGE protocol (3) with the following modifications. (*i*) We performed a low-cycle PCR to amplify the 3′ cDNA to generate a sufficient amount of templates for SAGE analysis from a limited RNA sample. The sense primer used was SAGE primer 1 (GGATT-TGCTGGTGCAGTACA) or SAGE primer 2 (CTGCTC-GAATTCAAGCTTCT); the antisense primer used was 5′-ACTATCTAGAGCGGCCGCTT-3′, which was located in the 3′ end of all cDNAs generated from the anchored oligo(dT) primers. (*ii*) The *Bsm*FI-released fragments containing the SAGE tags were gel purified before being used for ditag formation and concatenation to provide high-quality tags for SAGE analysis. SAGE-tag sequences were collected with the Big-Dye sequencing kit and ABI377 sequencer (Perkin–Elmer Applied Biosystems), and tag sequences were extracted with SAGE 300 software.

Bioinformatic Analysis. A reference SAGE-tag database was constructed from the UniGene Human database (release 127) representing most of the known human expressed sequences in GenBank. The conditions used for determining SAGE tags in sequences included (*i*) the orientation of each transcript, (*ii*) the presence of a poly(A) signal (AATAAA or ATTAAA), (*iii*) the presence of a poly(A) tail, and (*iv*) the presence of the last CATG cleavage site in the sequence (6). All SAGE tags extracted from the reference sequences were used for building the reference SAGE database. A computational program, GIST (Gene Identification and Sequence Topography), was developed for matching experimental SAGE tags against the reference SAGE database (www.hpcl.cs.uchicago.edu/gist/) for identifying

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

‡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.

www.pnas.org/cgi/doi/10.1073/pnas.051013798

Table 1. Distribution of SAGE tags from CD15$^+$ myeloid progenitor cells

	Copies of SAGE tags					
	≥100	99 to 10	9 to 5	4 to 2	1	Total
Total tags	90 (0.2)	918 (2)	1,338 (4)	7,131 (19)	28,042 (75)	37,519 (100)
Novel tags	0 (0)	59 (6)	159 (12)	1,795 (25)	15,507 (55)	17,520 (47)
Matched tags	90 (100)	859 (94)	1,179 (88)	5,236 (73)	12,535 (45)	19,899 (53)

The number within parentheses is the percentage of the tags among the total tags in each group.

potential corresponding genes for each SAGE tag. Sample tags that matched reference tags exactly were called "matched tags"; unmatched tags were called "novel tags" including from single-base mismatch till no match for all 10 bases.

Gene Confirmation Through GLGI Technique. A high-throughput GLGI procedure was used for converting SAGE tags into their corresponding 3′ cDNA (ref. 4; J.J.C., S.L., G.Z., and S.M.W., unpublished data). Briefly, we used a SAGE-tag sequence to design the sense primer (5′-GGATCCCATGxxxxxxxxxx-3′), and the sequence (5′-ACTATCTAGAGCGGCCGCTT-3′) at the 3′ end of all of the cDNAs, which was incorporated during the reverse transcription from the anchored oligo(dT) primers, was used as the antisense primer. The same 3′ poly(dA/dT)$^-$ cDNA sample used for SAGE analysis was used as the template. Platinum *Taq* polymerase (Life Technologies) was used for GLGI amplification. The amplified fragments were cloned into TOPO TA vector (Invitrogen) and sequenced with M13 reverse or forward primers. All steps were performed in 96-well format. Each sequence was matched through BLAST to the GenBank NR (nonredundant) or human expressed sequence tag (EST) databases (http://www.ncbi.nlm.nih/BLAST/). We used the accession numbers of the matched sequences to search the UniGene database for identification of the corresponding UniGene cluster number.

Results

Distribution of the SAGE Tags and Match of SAGE Tags to Known Expressed Sequences. A total of 37,519 unique SAGE tags were identified from 99,369 individual SAGE tags. Among the 37,519 identified unique tags, 75% were single copies, 19% were between five and two copies, 4% were between nine and five copies, 2% were between 99 and 10 copies, and only 0.2% of tags were present in more than 100 copies (Table 1). Comparison of these 37,519 SAGE tags to the SAGE database showed that 53% of the tags matched known expressed sequences including known genes and ESTs, and 47% of the tags had no match. Analysis of the matched and novel tags showed that tags present in high copies had a high percentage of matches to known expressed sequences, whereas the majority of the novel tags were concentrated in the low-abundance class, especially those with a single copy (Table 1).

The Problem of Multiple Matches for SAGE Tags. Among the 19,899 matched tags, 6,776 (34%) were matched to multiple sequences. The distribution of tags with multiple matches paralleled the abundance of SAGE tags (Table 2). The total number of tags with matches (19,899) was considerably less than the total number of clusters matched by these tags (36,156) because of the multiple matches. This difference primarily was due to the number of matched EST clusters, which was 2.2-fold higher than the number of tags matched to ESTs (Table 3).

Genes Highly Expressed in Myeloid Progenitor Cells. There were 90 SAGE tags with more than 99 copies. All 59 highly expressed SAGE tags with multiple matches were analyzed with the GLGI technique for identification of the corresponding genes (Table 4). A total of 35 ribosomal protein genes were present among these highly expressed genes. Among the 49 genes with various functions, some were highly expressed ubiquitously in various cell types including eukaryotic translational elongation factor 1, ferritin heavy chain and light chain, and translational-controlled tumor protein (6). Some may play a role in myeloid differentiation. For example, *MRP14* (779 copies) has been identified as a regulator of differentiation of human myelocytes and monocytes (7). Thymosin-β4 (301 copies) is involved in the differentiation of lymphocytes, macrophages, and granulocytes (8). Dual-specificity phosphatase 1 (254 copies) is a protein-tyrosine phosphatase involved in the cellular response to environmental stress (9). *FOS* (207 copies), an important transcriptional factor, is expressed highly in the myeloid progenitor cells. Together with *JUN*, *FOS* plays an important role in regulating the expression of many genes (10, 11).

Genes Important for Myeloid Differentiation. We analyzed the level of expression of a set of genes considered to be important for myeloid differentiation. We used the UniGene clusters for these genes to search for the SAGE tags potentially derived from these genes in our SAGE-tag collection. Because most of these genes have been cloned, our searches focused on the mRNA sequences rather than ESTs. Several patterns of matches were found: a tag can match a unique cluster for a gene or match several clusters representing several genes. For example, tag AGAAGCCGTG for *CD11b* matched a single cluster (Hs.172631), and tag TGGAAAGTGA for *FOS* matched four clusters (Hs.25647, Hs.187890, Hs.18127, and Hs.214906). In addition, several tags matched the same cluster, e.g., four different tags matched the *MLL* gene cluster. The power of GLGI is illustrated by our analysis of the SAGE tags matched to the *FOS* and *MLL* genes. Of the four possible unique clusters in the database matched by TGGAAAGTGA, *FOS* was identified by GLGI to be the correct gene. In contrast, for tags that matched *MLL*, the tag TG-

Table 2. Distribution of matched SAGE tags from CD15$^+$ myeloid progenitor cells

	Copies of SAGE tags					
	≥100	99 to 10	9 to 5	4 to 2	1	Total
Total matched tags	90	859	1,179	5,236	12,535	19,899
Multi-matches	59 (66)	449 (52)	494 (42)	2,071 (40)	3,703 (30)	6,776 (34)
Single match	31 (34)	410 (48)	686 (58)	3,165 (60)	8,832 (70)	13,121 (66)

The number within parentheses is the percentage of the tags among the total matched tags in each group.

Table 3. Summary of matched tags and matched clusters

Total tags with matches	19,899
Tags matched to genes	6,193
Tags matched to ESTs	13,706
Total clusters matched by tags	36,156
Matched clusters with genes	7,525
Matched clusters with ESTs	28,631

CACGTTTT present in 110 copies was not from *MLL* but rather it represented ribosomal protein L32 (Hs.169793), and the tag TCAAGTTTAT represented an EST (Hs.116468). For the tag GCTCCCCTTT, which could represent myeloperoxidase, GLGI showed that this tag represented an EST (Hs.292231). Among 30 tags from 17 genes in the analysis, 19 tags representing 13 genes were confirmed by GLGI (Table 5). Of great importance is the fact that four SAGE tags were shown not to be the genes initially assumed from the cluster analysis. These genes were expressed at various levels ranging from high to single copies. Several of these expressed genes existed as different splice variants, e.g., *FOS* had three, and *JUN* had two different splicing variants. Among the seven tags that could not be confirmed by GLGI, some were artifacts of PCR amplification as confirmed by the lack of CATG in the 5′ end of the matched sequences such as *JUNB*. Some genes considered to be involved in myeloid differentiation were not detected in this study, including *PU.1* and *HOX A9* (12).

Discussion

CD15$^+$ Cells Represent the Myeloid Progenitor Cells in Bone Marrow. CD15 is expressed on myeloid precursor cells, neutrophils, eosinophils, monocytes, myeloid leukemia, and myeloid cell lines (13). The following features identify the purified cells as being myeloid progenitor cells. (*i*) The cells used for this study were bone-marrow mononuclear cells isolated through Ficoll/Paque solution; therefore, mature polymorphonuclear cells such as neutrophils, eosinophils, and basophils were virtually eliminated. (*ii*) Although monocytic cells are also CD15$^+$, these cells constitute only 0.3% of bone-marrow cells. The potential contamination by these cells in the purified myeloid progenitors is minimal. In contrast, more than 50% of bone-marrow cells are myeloid cells (13). (*iii*) CD15 is not expressed on erythrocytes, platelets, or T and B cells (13). Therefore, these cells were not included in the purified cells.

Distribution of Matched and Novel SAGE Tags in the Collection. The distribution of SAGE tags indicates that in myeloid progenitor cells as in other cell types, a small number of genes is expressed at high levels, and the majority of genes are expressed at lower levels (6, 14). Matching the unique tags with known expressed sequences shows that most of the matched tags are present at higher copy levels, whereas a large proportion of tags at low copy levels have no match to known expressed sequences. This finding indicates that the known expressed sequences identified thus far largely represent the genes expressed at higher levels, and a large number of genes expressed at low levels remain to be identified.

Specificity of SAGE Tags for Gene Identification. The match of SAGE tags to known expressed sequences shows that more than one third of these tags have multiple matches with many sequences. This problem is especially serious for high-copy tags, on which many target genes identified through SAGE analysis were located. This result raises the question of the correctness of SAGE tags said to represent a unique gene. We have observed that both the length of SAGE tags and the redundancy of the reference database contribute to this problem (S.L., T.C., J.-J.C., G.Z., L. R. Scott, J.D.R., S.M.W., unpublished results). For SAGE tags with multiple matches, a gene assignment based on a tag can be reliable only after confirmation. A proven way to enhance the specificity of SAGE tags is to convert the SAGE tags into 3′ ESTs with much longer lengths. This increase can be achieved through the GLGI approach (4). Standard reverse

Table 4. Top genes expressed in CD15$^+$ myeloid progenitor cells

SAGE tag	Copy	No. matched UniGene cluster*	GLGI confirmation	Gene
CCCATCGTCC	949	Hs.151604		Ribosomal protein S8
TGTGTTGAGA	945	6	Hs.181165 (X03558)	Eukaryotic translation elongation factor 1 α1
TACCTGCAGA	924	Hs.100000,Hs.256957,Hs.253884	Hs.100000 (NM_002964)	S100 calcium-binding protein A8 (MRP8)
GTTGTGGTTA	853	Hs.75415	Hs.75415 (AB021288)	β_2-Microglobulin
CTAAGACTTC	802	Hs.80562		Gelsolin
GTGGCCACGG	779	Hs.112405		S100 calcium-binding protein A9 (MRP9)
AGCCCTACAA	732	Hs.180532,Hs.43445	Hs.180532 (AW070665)	Glucose-phosphate isomerase
ACTAACACCC	726	Hs.125753		Hypothetical protein FLJ20302
TTCATACACC	599	Hs.75760,Hs.199250,Hs.278302,Hs.108661	Hs.75760 (BE676905)	Sterol carrier protein 2
ATGTAAAAAA	569	8	Hs.204040 (AF004230)	Leukocyte immunoglobulin-like receptor, subfamily B
CACCTAATTG	521	Hs.282283,Hs.289107	Hs.282283 (AW270021)	Mitochondrial DNA
CTCATAAGGA	516	Hs.251871,Hs.151604,Hs.80562	Not detected	
GCAAGCCAAC	461	Hs.80562,Hs.119000	Hs.80562 (AV698237)	Gelsolin
GTGAAGGCAG	439	Hs.77039,Hs.4221	Hs.77039 (L13802)	Ribosomal protein S3A
TTGGGGTTTC	415	11	Hs.62954 (L20941)	Ferritin, heavy polypeptide 1
ATTTGAGAAG	400	5	Hs.169921 (AW778986)	General transcription factor II, i, pseudogene 1
ATGGCTGGTA	394	Hs.182426,Hs.254246	Hs.182426 (X17206)	Ribosomal protein S2
CCACTGCACT	385	175	Hs.271396 (T90298)	EST
ACCCTTGGCC	370	Hs.279009,Hs.77608,Hs.80562	Hs.279009 (AI065095)	EST
ACGCAGGGAG	358	Hs.180532		Glucose-phosphate isomerase
AGCTCTCCCT	342	Hs.82202,Hs.284836	Hs.82202 (X53777)	Ribosomal protein L17
AAAACATTCT	341	Hs.75525,Hs.80562,Hs.75621	Not detected	
ATCAAGGGTG	326	Hs.157850,Hs.277180	Hs.157850 (U21138)	Ribosomal protein L9
AAAAAAAAAA	325	592	Not detected	
TGATTTCACT	322	Hs.240443,Hs.38503,Hs.24322,Hs.7149	Hs.240443 (AI133606)	FLJ23538 fis, clone LNG08010
TTGGTGAAGG	301	Hs.75968,Hs.288031	Hs.75968 (M17733)	Thymosin, β4

Table 4. Continued

SAGE tag	Copy	No. matched UniGene cluster*	GLGI confirmation	Gene
GTGCACTGAG	300	5	Hs.277477 (X58536)	Major histocompatibility complex, class I, C
CTGACCTGTG	292	Hs.77961,Hs.277477,Hs.181244	Hs.77961 (D83956)	Major histocompatibility complex, class I, B
CCTGTAATCC	284	430	Hs.292308 (AW975509)	ESTs
GTGAAACCCC	281	325	Hs.241392 (AI687343)	Small inducible cytokine A5
CACAAACGGT	271	Hs.195453	Hs.195453 (U57847)	Ribosomal protein S27
TTGGTCCTCT	263	Hs.108124,Hs.12328,Hs.9739,Hs.112845	Hs.108124 (AF026844)	Ribosomal protein L41
CTTGACATAC	254	Hs.171695	Hs.171695 (NM_004417)	Dual specificity phosphatase 1
ATAATTCTTT	254	6	Hs.539 (L31610)	Ribosomal protein S29
GGGCATCTCT	243	Hs.76807,Hs.75061	Hs.76807 (K01171)	Major histocompatibility complex, class II, DR α
TAGGTTGTCT	242	5	Hs.119252 (X16064)	Tumor protein, translationally controlled 1
GGGCTGGGGT	238	Hs.183698,Hs.118757,Hs.90436	Hs.183698 (U10248)	Ribosomal protein L29
GGATTTGGCC	230	5	Hs.119500 (NM_002268)	Karyopherin α4 (importin α3)
GCCGTGTCCG	224	Hs.241507,Hs.230982	Hs.241507 (M20020)	Ribosomal protein S6
GTTCACATTA	210	Hs.84298	Hs.84298 (AW768633)	CD74 antigen
TGGAAAGTGA	207	Hs.25647,Hs.23317,Hs.187890,Hs.214906	Hs.25647 (V01512)	FOS
TCACCCACAC	200	Hs.234518,Hs.131965,Hs.288372	Hs.234518 (AI378597)	Ribosomal protein L23
ACTTTCCAAA	198	Hs.180532	Hs.180532 (BE387582)	Glucose-phosphate isomerase
AGCACCTCCA	195	Hs.75309	Hs.75309 (AW874543)	Eukaryotic translation elongation factor 2
GTAAGTGTAC	188	Hs.153423		KIAA0493 protein
GCATAATAGG	187	Hs.184108	Hs.184108 (U14967)	Ribosomal protein L21
TGGGCAAAGC	185	Hs.2186,Hs.289975	Hs.2186 (Z11531)	Eukaryotic translation elongation factor 1 γ
CGCTGGTTCC	179	Hs.179943,Hs.266803,Hs.132525,Hs.283429	Hs.179943 (L05092)	Ribosomal protein L11
AAGACAGTGG	178	Hs.184109,Hs.282786,Hs.3352	Hs.184109 (X66699)	Ribosomal protein L37a
AGGGCTTCCA	176	Hs.29797,Hs.276544	Hs.29797 (AB007170)	Ribosomal protein L10
CTGGGTTAAT	168	9	Hs.126701 (M81757)	Ribosomal protein S19
CATTTGTAAT	166	Hs.240443,Hs.8417	Hs.240443 (AW439275)	FLJ23538 fis, clone LNG08010
CCCTGGGTTC	166	Hs.111334,Hs.52891	Hs.111334 (M11147)	Ferritin, light polypeptide
GGACCACTGA	160	Hs.119598,Hs.150580,Hs.74637	Hs.119598 (X73460)	Ribosomal protein L3
TTGTAATCGT	151	Hs.125078	Hs.125078 (D89870)	Ornithine decarboxylase antizyme 1
GAAAAAAAAA	147	169	Hs.169921 (AW075720)	General transcription factor II, i, pseudogene 1
TAATAAAGGT	145	9	Hs.151604 (X67247)	Ribosomal protein S8
TTCAATAAAA	144	5	Hs.177592 (M17886)	Ribosomal protein, large, P1
TGTGCTAAAT	142	Hs.250895		Ribosomal protein L34
ACATCATCGA	141	Hs.182979,Hs.289690	Hs.182979 (L06505)	Ribosomal protein L12
CAATAAATGT	139	Hs.179779,Hs.66151,Hs.163109,Hs.225767	Hs.179779 (L11567)	Ribosomal protein L37
CCTCAGGATA	138	Hs.169921,Hs.279932	Hs.169921 (AW376254)	General transcription factor II, i, pseudogene 1
CGCCGCCGGC	136	Hs.182825		Ribosomal protein L35
AAGGAGATGG	136	Hs.184014,Hs.164170	Hs.184014 (X69181)	Ribosomal protein L31
CTGTTGGTGA	135	Hs.3463		Ribosomal protein S23
TGGTGTTGAG	135	Hs.275865	Hs.275865 (X69150)	Ribosomal protein S18
ATTCTCCAGT	133	Hs.234518,Hs.287464	Hs.234518 (NM_000978)	Ribosomal protein L23
CGCCGGAACA	130	Hs.286		Ribosomal protein L4
TACCATCAAT	130	6	Hs.169476 (AF261085)	Glyceraldehyde-3-phosphate dehydrogenase
CCCACAACCT	127	Hs.252136		Ficolin 1
TCAGACGCAG	126	Hs.250655		Prothymosin, α
GTGCTGAATG	125	Hs.77385,Hs.198689	Hs.77385 (M31212)	Myosin, light polypeptide 6
CCAGAACAGA	121	Hs.111222		Ribosomal protein L30
GTGAAACCCT	120	155	Hs.282725 (AV661681)	EST
GAGGGAGTTT	120	Hs.76064	Hs.76064 (U14968)	Ribosomal protein L27a
TCAGATCTTT	116	6	Hs.75344 (NM_001007)	Ribosomal protein S4
GCGACGAGGC	116	Hs.2017		Ribosomal protein L38
ACTTTTTCAA	115	46	Hs.169921 (AW129653)	General transcription factor II, i, pseudogene 1
CCAGAGAACT	113	Hs.6975,Hs.243886	Hs.6975 (AW582859)	PRO1073 protein
CAAGCATCCC	110	Hs.153423		KIAA0493 protein
TGCACGTTTT	110	7	Hs.169793 (X03342)	Ribosomal protein L32
GCAGCCATCC	109	Hs.4437		Ribosomal protein L28
TCGAAGCCCC	107	Hs.118223		Microfibrillar-associated protein 4
GCCTGTATGA	105	Hs.180450		Ribosomal protein S24
TTGGAACAAT	105	Hs.80305		Human clone 23719 mRNA sequence
ACCCGCCGGG	105	0	BF352207	EST
GAACACATCC	104	Hs.75879	Not detected	
AGAAAGATGT	103	Hs.78225,Hs.260622,Hs.243561,Hs.224788	Hs.78225 (NM_000700)	Annexin A1
AAAAGAAACT	101	Hs.172182,Hs.256309,Hs.230481,Hs.152860	Hs.172182 (Y00345)	Poly(A)-binding protein, cytoplasmic 1
AGGCTACGGA	100	Hs.119122,Hs.211582	Hs.119122 (AB028893)	Ribosomal protein L13a

*All 59 tags with more than one match and 12 tags with single match were tested by GLGI. The number of clusters is listed for the ones over four clusters.

transcription–PCR with primers or Northern blot with a probe from one of the matched sequences cannot prove the correctness of the assignment unless all of the matched sequences are tested. In fact, as we show in Table 5, many genes in the database are matched by the same SAGE tag. However, not all of the matched genes exist in the initial cDNA material used for SAGE analysis. Therefore, the correctness of each SAGE tag/gene association in each experiment has to be enforced in a very precise manner in order for the data to be valid.

Genes Highly Expressed in Myeloid Progenitor Cells. Genes highly expressed in myeloid progenitor cells include basic structural and functional genes for protein synthesis and catalytic enzymes for metabolism. There are five genes represented only by EST sequences among these highly expressed genes; their identity and functions need to be studied. Myeloblast cells and monoblast cells differentiate from the same committed myelomonocytic cell. We compared the 90 most highly expressed genes from myeloid progenitor cells reported here with the top 80 genes from mature monocytes (ref. 15; www.prevent.m.u-tokyo.ac.jp/Monocytes.html). The techniques used for confirming the correctness of each of these monocytic genes, given the problem of multiple matches, were not stated clearly. In both sets of genes, ribosomal protein genes account for about one third of the total. Among other genes, some are common in both cell types such as β_2-microglobulin, translational elongation factors, ferritin heavy and light chains, MHC complex genes, *MRP8*, *MRP14*, thymosin β4, and translationally controlled tumor protein; many others are different between these two cell types. The genes detected only in myeloid progenitor cells included annexin A1, CD74 antigen, cellular repressor of E1A-stimulated genes, dual-specificity phosphatase 1, *FOS*, gelsolin, KIAA0493 protein, poly(A)-binding protein, small inducible cytokine A5, and sterol carrier protein 2. Genes detected only in monocytes were Bak protein, *CAPL* protein, cell division cycle 2-like 1, cytochrome b-245, α polypeptide, *DAP12*, *HSP27*, *IP30*, *LLR* ep3, multitransmembrane protein, *SMCX*, etc. The differences in these highly expressed genes between these two cell types suggest that these genes may play roles in differentiation from myelomonocytic stem cells to myeloid or monocytic cells, as well as reflect the unique functions performed by these different cell lineages.

Genes Important for Myeloid Differentiation. Genes that are functionally important for myeloid differentiation include transcriptional factors, growth factors, or specific functional genes (12). Although we detected a number of genes considered important for myeloid-cell differentiation (Table 5), some others were not detected in this analysis. The reason for the lack of detection of these genes may be that they might be expressed at low levels, below the detection threshold. Alternatively, these genes might not be expressed in the cells we analyzed. Our studies analyzed gene expression in primary cells, in contrast to other studies with transformed cell lines (16, 17). It has been observed that the expression pattern can be significantly different between pri-

Table 5. Expression level of genes important for myeloid differentiation

Genes		Presence in the SAGE tag set			
Name	UniGene ID	Tag	Tag copy	Cluster(s) matched by tag*	GLGI confirmation†
AML1	Hs.129914	GCCCCCTCCG	1	Hs.129914,Hs.127765	N
CEBPA	Hs.76171	GGCAACTGCG	2	Hs.76171	N
CEBPB	Hs.99029	GCTGAACGCG	24	Hs.99029 (AI702318)	Y
CEBPD	Hs.76722	CTCACTTTTT	24	Hs.76722 (NM_005195)	Y
CBP	Hs.23598	ACTACAAGGA	2	Hs.23598 (AI953646)	Y
		CAAGCGCTCT	2	Hs.23598 (U89355)	Y
CD11b	Hs.172631	AGAAGCCGTG	10	Hs.172631 (AI631857)	Y
CD18	Hs.83968	GAGACTTGAG	15	Hs.83968 (AI683192)	Y
ETO	Hs.31551	AGCATTGGAT	4	Hs.31551	N
FOS	Hs.25647	TGGAAAGTGA	207	Hs.25647 (AW337322),Hs.187890,Hs.18127,Hs.214906	Y
		TCAAAAGACC	30	Hs.25647 (R84834),Hs.274707	Y
		GATATAGCTA	7	Hs.25647 (BE177025)	Y
G-CSF receptor	Hs.2175	CTCCATCCAG	33	Hs.2175 (AI273981)	Y
	Hs.174142	TGGCTGGCCA	4	Hs.174142	N
Cathepsin G	Hs.100764	AGGAGGGGAA	3	Hs.100764	N
JUN	Hs.78465	CTAACGCAGC	32	Hs.78465 (BF221858)	Y
		CCTTTGTAAG	12	Hs.78465 (AV736058)	Y
		ATGTCTTCGT	2	Hs.77039,Hs.78465,Hs.144926	Hs.77039 (NM_001006)
JUNB	Hs.198951	ACCCACGTCA	22	Hs.198951,Hs.50915	N
MLL	Hs.199160	TGCACGTTTT	110	Hs.169793,Hs.199160,Hs.279943,Hs.36927, Hs.5338,Hs.83450,Hs.173902,Hs.183506	Hs.169793 (AW073833)
		TCAAGTTTAT	3	Hs.116468,Hs.130707,Hs.199160,Hs.15871	Hs.116468 (AA631929)
		TGTTATTTTG	3	Hs.199160 (W16724)	Y
		TATAACAGAT	1	Hs.199160	N
MLL2		GAAAGGACAT	1	Hs.114765 (AW960268),Hs.199160	Y
MRP8	Hs.100000	TACCTGCAGA	924	Hs.100000 (AI826354),Hs.256957,Hs.253884	Y
		CCGTCTACAG	13	Hs.100000 (AA321386)	Y
		AAGAAAGCCA	8	Hs.100000 (T99219)	Y
Myeloperoxidase	Hs.1817	GCTCCCCTTT	36	Hs.1817,Hs.173103	Hs.292231 (AI054296)
		TATGTGCGAA	3	Hs.1817 (AA883501)	Y
Myeloid zinc finger protein	Hs.169832	GTCAGAACAC	2	Hs.169832 (AW449715)	Y

*Cluster numbers underlined indicate the GLGI-confirmed genes.

†Y, confirmed by GLGI; N, not confirmed by GLGI, possibly because they were PCR artifacts. The clusters in this column show the detected sequences that do not match to the expected genes.

mary cells and cell lines (18). The biologically relevant pattern of gene expression should be preserved more accurately in primary cells than in cell lines. When specific genes are analyzed with the SAGE technique, one issue needs to be addressed: SAGE tags are located at the most 3′ end of a transcript. If a gene has different splicing forms with different 3′ sequences, each of these templates may contribute a unique tag. These tags are different from one another but the templates contributing to these tags are within the same UniGene cluster because of overall sequence homology. This explanation can explain why one gene can have several unique tags. A possibility exists that tags may be generated from sequences upstream of the last CATG because of incomplete *Nla*III digestion; however, it is unlikely that this problem contributes significantly to the event. From this study, it should be noted that the SAGE technique could detect the different splice variants by identifying each template through a unique tag if their 3′ parts differed. For example, three different tags with different copy numbers were detected for the *MRP8* gene. Editing the full-length expressed transcripts can generate different isoforms for different functions, which is one of the major control mechanisms for gene expression (19). Further exploration of this mechanism may provide significant insights for monitoring the function of these splice variants.

Applying SAGE and GLGI Techniques for Gene Identification. On the basis of our data and those of many other SAGE analyses, about half of the SAGE tags from various cell types are novel (6). Therefore, a large number of genes expressed at low levels in the genome have not been identified despite intensive efforts in the past decades. Further studies on gene identification and gene function need to focus on this category. Current estimates for the gene contents in different genomes are based on computational predictions or some model systems or rely on known gene and EST sequences that are far from complete (20–22). Because SAGE can identify genes expressed at low levels that are difficult to detect with other current techniques, its application should provide comprehensive SAGE-tag indices of the expressed genes in various cell types, as shown in SAGE analysis of gene expression in brain (www.nabi.nih.gov/SAGE). However, the SAGE-tag index itself cannot be converted automatically into the gene index because of the low specificity caused by the short length of SAGE tags and the lack of matches to known expressed sequences for the tags. We have shown that by conversion of the SAGE tags into the longer 3′ ESTs through GLGI, the SAGE-tag index can be converted into a gene index.

An important issue for gene identification in the postgenome era is the identification of genes expressed at low levels. The conventional EST approach, or random sequencing of clones from regular cDNA libraries, primarily identifies the genes expressed at high levels, as shown that only thousands of genes can be identified in a library by using this strategy regardless of the large-scale efforts, and most of these genes are already known (23–25). Applying the subtraction/normalization strategy can identify more genes expressed at lower levels, many of which are novel genes (5, 26). The application of SAGE and GLGI techniques provides tools for identification of genes expressed at low levels, which represent the majority of genes expressed from the genome.

We thank Dr. R. Chen for technical advice on SAGE performance and Drs. N. Zeleznik-Le and P. Domer for their thoughtful comments on this manuscript. Work was supported by National Institutes of Health Grants CA42557 (J.D.R.) and CA78862-01 (J.D.R. and S.M.W.), American Cancer Society IRG-41-40 (S.M.W.), and the G. Harold and Lelia Y. Mathers Foundation (S.M.W.).

1. Robert, I., Handin, S. E. & Thomas, P. S., eds. (1995) *Principles and Practice of Hematology* (Lippincott, Philadephia).
2. Rowley, J. D. *Semin. Hematol.* (1999) **36,** 59–72.
3. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995) *Science* **270,** 484–487.
4. Chen, J.-J., Rowley, J. D. & Wang, S. M. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 349–353.
5. Wang, S. M., Fears, S. C., Zhang, L., Chen, J-J. & Rowley, J. D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 4162–4167.
6. 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.
7. Warner-Bartnicki, A. L., Murao, S., Collart, F. R. & Huberman, E. (1993) *Exp. Cell Res.* **204,** 241–246.
8. Gondo, H., Kudo, J., White, J. W., Barr, C., Selvanayagam, P. & Saunders, G. F. (1987) *J. Immunol.* **139,** 3840–3848.
9. Keyse, S. M. & Emslie, E. A. (1992) *Nature (London)* **359,** 644–647.
10. Woodgett, J. R. (1990) *Semin. Cancer. Biol.* **1,** 389–397.
11. Abate, C. & Curran, T. (1990) *Semin. Cancer. Biol.* **1,** 19–26.
12. Gordon, S. & Clarke, S. (1998) *J. Leukoc. Biol.* **63,** 153–168.
13. Bettelheim, P. (1989) in *Leukocyte Typing IV*, ed. Knapp, W. (Oxford Univ. Press, Oxford), pp. 798.
14. Bishop, J. O., Morton, J. G., Rosbach, M. & Richardson, M. (1974) *Nature (London)* **250,** 199–204.
15. Hashimoto, S., Suzuki, T., Dong, H. Y., Yamazaki, N., Matsushima, K. (1999) *Blood* **94,** 837–844.
16. Hohn, P. A., Popescu, N. C., Hanson, R. D., Salvesen, G., Ley, T. J., (1989) *J. Biol. Chem.* **264,** 13412–13419.
17. Zimmer, M., Medcalf, R. L., Fink, T. M., Mattmann, C., Lichter, P. & Jenne, D. E. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 8215–8219.
18. Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. & Kinzler, K. W. (1997) *Science* **276,** 1268–1272.
19. Newman, A. (1998) *Curr. Biol.* **8,** R903–R905.
20. Fields, C., Adams, M. D., White, O. & Venter, J. C. (1994) *Nat. Genet.* **7,** 345–346.
21. Smaglik, P. (2000) *Nature (London)* **405,** 264.
22. Ewing, B. & Green, P. (2000) *Nat. Genet.* **2,** 232–234.
23. Adams, M. D., Dubnick, M., Kerlavage, A. R., Moreno, R., Kelley, J. M., Utterback, T. R., Nagle, J. W., Fields, C. & Venter, J. C. (1992) *Nature (London)* **355,** 632–634.
24. Lanfranchi, G., Muraro, T., Caldara, F., Pacchioni, B., Pallavicini, A., Pandolfo, D., Toppo, S., Trevisan, S., Scarso, S. & Valle, G. (1996) *Genome Res.* **6,** 35–42.
25. 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.
26. Bonaldo, M. F., Lennon, G. & Soares, M. B. (1996) *Genome Res.* **6,** 791–806.