Gene expression profile of murine long-term reconstituting vs. short-term reconstituting hematopoietic stem cells

Jiang F. Zhong*†‡, Yi Zhao*§, Susan Sutton¶, Andrew Su¶, Yuxia Zhan*, Lunjian Zhu*, Chunli Yan*, Tim Gallaher*, Patrick B. Johnston*‖, W. French Anderson*†, and Michael P. Cooke¶

*Gene Therapy Laboratories, †Department of Biochemistry and Molecular Biology, and §Division of Hematology of the Department of Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033; and ¶Genomics Institute of the Novartis Research Foundation, San Diego, CA 92121

Communicated by J. Craig Venter, Center for the Advancement of Genomics, Rockville, MD, December 24, 2004 (received for review September 8, 2004)

The hematopoietic stem cell (HSC) compartment is composed of long-term reconstituting (LTR) and short-term reconstituting (STR) stem cells. LTR HSC can reconstitute the hematopoietic system for life, whereas STR HSC can sustain hematopoiesis for only a few weeks in the mouse. Several excellent gene expression profiles have been obtained of the total hematopoietic stem cell population. We have used five-color FACS sorting to isolate separate populations of LTR and STR stem cell subsets. The LTR HSC has the phenotype defined as Lin^- Sca^+ Kit^+ 38^+34^-; two subsets of STR HSC were obtained with phenotypes of Lin^- Sca^+ Kit^+ 38^+34^+ and Lin^- Sca^+ Kit^+ 38^- 34^+. The microarray profiling study reported here was able to identify genes specific for LTR functions. In the interrogated genes (≈12,000 probe sets corresponding to 8,000 genes), 210 genes are differentially expressed, and 72 genes are associated with LTR activity, including membrane proteins, signal transduction molecules, and transcription factors. Hierarchical clustering of the 210 differentially expressed genes suggested that they are not bone marrow-specific but rather appear to be stem cell-specific. Transcription factor-binding site analysis suggested that GATA3 might play an important role in the biology of LTR HSC.

microarray | regulation

Murine adult hematopoietic stem cells (HSC) reside in the bone marrow (BM) and are regulated by a complex network of gene interactions that maintain the proper balance between self-renewal, differentiation, and apoptosis. Classified by their multilineage repopulating ability in irradiated recipients, HSC can be divided into two groups: long-term reconstituting (LTR) cells that can sustain hematopoietic systems for the life of the animal, and short-term reconstituting (STR) cells that can repopulate lymphoid and myeloid cells for several weeks (1–3). The LTR HSC is considered the true stem cell, and its frequency is estimated to be ≈1 in 100,000 murine BM cells (4, 5). Previous studies indicated that LTR activity is in the Lin^- Sca^+ Kit^+ fraction of murine BM (6–8). The CD34 and CD38 are two important markers for HSCs. In general, human HSCs are believed to be $CD38^-$ $CD34^+$ (9). However, Zanjani *et al.* (10–12), using the human/sheep competitive engraftment model, demonstrated that human BM $CD34^-$ cells are capable of long-term multilineage engraftment *in vivo*.

In mouse, it has been accepted that the HSCs are in the BM $CD34^-$ fraction (13–15). The HSCs in yolk sac, BM, and fetal liver have been reported to be $CD38^+$ cells (16–19). Using five-color FACS and antibodies to cell-surface markers, we were able to fractionate this population of cells into three subsets: Lin^- Sca^+ Kit^+ $CD38^+$ $CD34^-$ [38^+34^-, Lin^- Sca^+ Kit^+ $CD38^+$ $CD34^+$ (38^+34^+), and Lin^- Sca^+ Kit^+ $CD38^-$ $CD34^+$ (38^-34^+)] (19). Competitive repopulation studies demonstrated that the 38^+34^- cells could provide long-term reconstitution of the hematopoietic system in primary, secondary, and tertiary BM transplantation experiments. On the other hand, the 38^+34^+ and 38^-34^+ subsets had minimal LTR activity but possessed excellent STR ability. We were able to demonstrate (19) that the pattern of differentiation *in vivo* in the mouse is:

$$38^+34^- \rightarrow 38^+34^+ \rightarrow 38^-34^+.$$

Global gene expression profiling provides a tool to understand the regulation of hematopoiesis. Using microarray or other molecular biological approaches, several groups have analyzed gene expression in the total HSC population (20–24). In some of these studies, however, HSC were compared with lineage-committed progenitors or fetal liver cells (20–22, 24). Because LTR HSC are quite different from lineage-committed or whole BM cells, a large number of genes were identified in these studies, only a subset of which are likely to directly control LTR stem cell activity. Indeed, comparison of genes identified in two of these studies (21, 22) revealed little overlap between the studies, suggesting that substantial differences exist in either HSC populations used by each group or the analysis methods used to identify stem cell genes. A more direct comparison of LTR (Rh^{lo} Sca-1^+ c-kit^+ $lin^{-/lo}$) and STR (Rh^{hi} Sca-1^+ c-kit^+ $lin^{-/lo}$) stem cells was conducted by Park *et al.* (23), who used cDNA libraries constructed from LTR cells and enriched for LTR genes by subtractive hybridization by using lineage-committed progenitors. Differential filter hybridization using probes prepared from LTR or STR HSC led to the identification of LTR-enriched transcripts in the libraries. Although useful, this technique is constrained by the difficulty of preparing comprehensive and representative libraries from limited numbers of cells and by bias arising from the need to prepare large amounts of probe for differential filter hybridization. Therefore, a direct comparison of the LTR and STR cells inside the stem cell compartment is desired.

In the present study, cRNA probes were prepared directly from LTR (38^+34^-) and STR (38^+34^+ and 38^-34^+) HSC for microarray profiling. Because the STR (38^+34^+ and 38^-34^+) HSC are immediately differentiated from LTR (38^+34^-) HSC, they are closely related, and such an analysis is more likely to identify genes that are gained or lost as a result of the transition from the LTR to the STR cell. In this study, ≈1.7% of the 12,000 interrogated probe sets were differentially expressed among the three subsets. From the differentially expressed genes, 72 have an expression pattern correlating with stem cell activity and are classified as LTR stem cell-related genes. Comparison of the expression of these genes across a mouse body atlas comprised of 45 tissues revealed that three genes had their peak expression within the HSC compartment, suggesting that

Freely available online through the PNAS open access option.

Abbreviations: HSC, hematopoietic stem cells; BM, bone marrow; LTR, long-term reconstituting; STR, short-term reconstituting; GNF, Genomics Institute of the Novartis Research Foundation.

‡To whom correspondence should be addressed. E-mail: jzhong@usc.edu.

‖Present address: Mayo Clinic, Rochester, MN 55905.

www.pnas.org/cgi/doi/10.1073/pnas.0409459102

their primary role may be to direct HSC differentiation. Additionally, comparison of these LTR-HSC genes with previously published HSC RNA expression data revealed substantial overlap, despite the different methods used to purify HSCs in each study. These data identify the transcription factor GATA3, certain membrane-associated proteins (Bdkrb, Fxyd3, Fzd4, Kit, Map17, Mpl, Ormdl3, Ptpro, and Sell), and signal transduction pathway regulatory proteins that may play important roles in the regulation of LTR HSC.

Materials and Methods

Isolation of HSC Subsets. C57BL/6J (Ly5.1) male mice were obtained from The Jackson Laboratory. Isolation of the LTR (Lin^- Sca^+ Kit^+ $CD38^+$ $CD34^-$ [abbreviated 38^+34^-]) and the STR (Lin^- Sca^+ Kit^+ $CD38^+$ $CD34^+$ [38^+34^+] and Lin^- Sca^+ Kit^+ $CD38^-$ $CD34^+$ [38^-34^+]) murine HSC subsets was as described (19). All animal procedures were approved by the University of Southern California Animal Care and Use Committee.

RNA Isolation and Microarray Processing of HSC Subsets. Cells from the three HSC subsets (38^+34^-, 38^+34^+, and 38^-34^+) were sorted with a MoFlo flow cytometer (Cytomation, Fort Collins, CO) into lysis buffer directly, as described (19). To minimize the biological variability, each stem cell subset RNA sample was pooled from multiple independent FACS sorting experiments by using identical gating schemes. For these experiments, four independent FACS sorts were conducted to collect the three subsets from BM cells. For each sorting experiment, 10 mice were used for the fresh Lin^- cell preparations and surface marker labeling. Cells were sorted directly into lysis buffer, and lysates were pooled for RNA extraction. Pooling samples from four independent experiments serves to minimize the variability arising from experiment-to-experiment variability and ensure that the expression measurements are representative of the cell populations under study.

Total RNA from 5×10^5 cells of each subset was extracted by Triazol (Sigma). Two aliquots of total RNA from each subset were used for microarray processing on Affymetrix Genechip MG-U74A at the Genomics Institute of the Novartis Research Foundation (GNF), as described (25). To determine the reliability of microarray processing, each pooled RNA sample was split into two aliquots and used for amplification, labeling, and hybridization to independent arrays. To evaluate the quality of these replicates and the reproducibility of the data derived, we analyzed these data by comparing the number of genes called present in each replicate and the correlation of gene expression measurements for genes detected in each replicate. On average, >90% of the genes that were called present in one sample are also called present in the companion replicate. For genes called present, the correlation of the expression data for the two replicates was extremely high, with r^2 values of 0.999 (see Table 1 and Fig. 6, which are published as supporting information on the PNAS web site). Thus, analysis of the individual hybridizations demonstrates a high degree of reproducibility of the data for each sample. These technical replicates indicate that the overall procedure from RNA preparation to data acquisition and analysis was highly reproducible.

RT-PCR Analysis. To confirm the observed expression differences, fresh RNA samples were prepared from independently sorted stem cell populations and used for semiquantitative RT-PCR verification. Hypoxanthine phosphoribosyl transferase, a housekeeping gene, was used to normalize all RT-PCR fractions for comparison (RT-PCR Primer sequences are listed in the *Supporting Text*, which is published as supporting information on the PNAS web site).

Data Analysis. Affymetrix Murine Genome U74A chips were used to monitor each of the three HSC subsets in duplicate. Expression data were extracted from image files by Affymetrix MICROARRAY SUITE 5.0 and were scaled to 200 expression units as the median. Raw expression values were normalized within each chip by dividing the median expression value of the chip. For each gene, the expression values were further normalized across chips by dividing the median of the six expression values among the six chips. Statistical analysis was performed with MATLAB (Mathworks, Natick, MA) software. Differentially expressed genes were selected by three filters: (*i*) ANOVA statistic ($P < 0.01$); (*ii*) >100 average expression units in at least one subset; and (*iii*) >2-fold differences between any two subsets. To compare data with previous studies, image files were kindly supplied by Ihor Lemischka (Princeton University, Princeton, NJ) (21), Douglas Melton (Harvard University, Boston) (22), and GNF for expression data extraction, as described above. Principal component analysis and K-mean clustering were performed with GENESPRING (Silicon Genetics, Redwood City, CA). Hierarchical clustering was performed by GENESPRING and the ROSETTA LUMINATOR SYSTEM (Agilent Technologies, Palo Alto, CA), with gene expression data from different data sets. Expression data of the 45 mouse body tissues were provided by GNF (26). Promoter searching for genes was performed with a sequence database containing a 10-kb upstream sequence from the translational start methionine. Randomization tests were performed by changing the orders of the nucleotides in the binding sites. *P* values for binding sites were generated by statistical bootstrap procedures.

Results

Expression of Interrogated Genes in the Three HSC Subsets. To compare gene expression levels within the stem cell compartment, RNA samples were prepared from the three HSC subsets sorted as described (19). Differentially expressed genes were selected by three criteria: (*i*) ANOVA statistic $P < 0.01$; (*ii*) >100 average expression units in at least one subset; and (*iii*) >2-fold in the differences between any two subsets. Because the three HSC subsets are closely related, with regard to both genotype and phenotype (all are Lin^- Sca^+ Kit^+ and comprise a three-unit set in the BM), the majority of genes (97.4%) are not differentially expressed among the three subsets. Among the interrogated genes (≈12,000 probe sets corresponding to 8,000 genes), 2.6% (210) genes passed these differential expression criteria (Fig. 1).

Selection of Stem Cell-Related Genes. To correlate stem cell activity of the three HSC subsets with gene expression, a hypothetical stem cell activity pattern corresponding to the *in vivo* repopulating activity of the three subsets was generated and used for comparison of the normalized expression levels of each differentially expressed gene identified above. To identify gene expression patterns in a manner that does not assume prior knowledge of stem cell activity, we performed Principal Component Analysis (PCA) on the stem cell expression data. PCA of the differentially expressed genes revealed that the most significant pattern in the data correlated positively with the hypothetical stem cell activity curve defined above: high in $CD38^+CD34^-$ and low in the other two subsets (see Fig. 1 *Inset*, red line). In addition, the second most significant pattern in the data were anticorrelated with the stem cell activity: low in $CD38^+CD34^-$ and high in the other two subsets (see Fig. 1 *Inset*, blue line). This analysis suggests that many of the differentially expressed genes fall into two classes of genes whose expression positively correlated with LTR HSC activity or genes whose expression negatively correlated with LTR HSC activity.

Correlation analysis of the gene expression patterns of differentially expressed genes with stem cell activity identified 72 genes with highly significant (Pearson $R > 0.95$) positive (Table 2, which is published as supporting information on the PNAS web site) or negative (Table 3, which is published as supporting information on the PNAS web site) correlations. Among these genes, there are six transcription factor genes, nine genes encoding plasma membrane proteins, six genes encoding known signal transduction molecules, and 17 genes with unknown functions. Fifty-two of the genes had

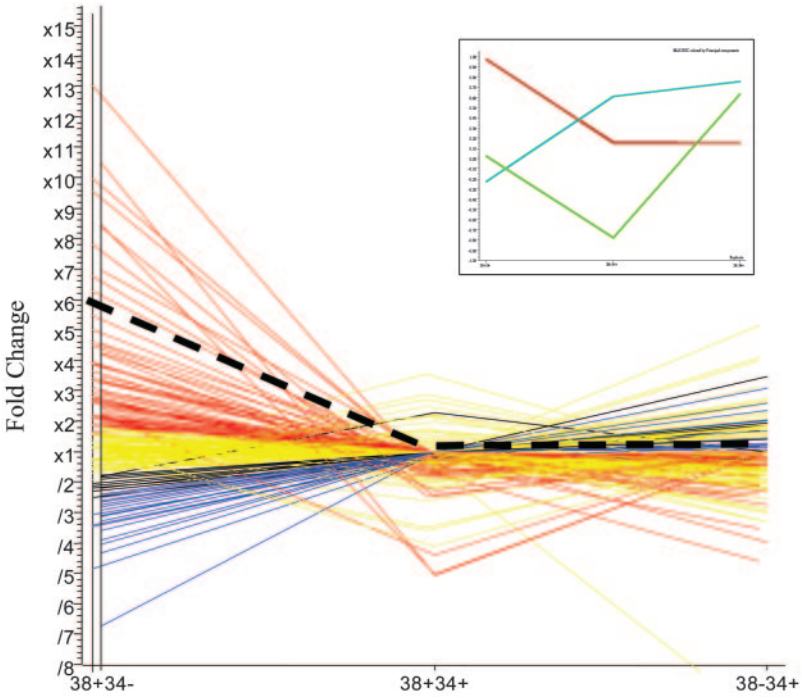

Fig. 1. Expression patterns of differentially expressed genes. There are 210 differentially expressed genes among the 8,000 interrogated genes. Fold changes are calculated from normalized expression values. The hypothetical stem cell activity pattern based on repopulation units is shown by a heavy dashed black line. The relative intensity of each gene is indicated by color based on intensity value in $38^{+}34^{-}$ subsets, blue being the lowest and red, the highest. (*Inset*) The three major patterns. The red line was the most significant pattern, and the blue line, the second most significant pattern.

expression patterns that correlated positively with LTR HSC activity. These included two well characterized stem cell membrane protein genes, the stem cell factor receptor, c-kit, and the thrombopoietin receptor, c-Mpl. Another membrane protein identified is Fzd4, which binds the Wnt proteins [recently shown to control HSC self renewal (27, 28)]. In addition, two genes that negatively regulate cytokine signaling *Inpp5d* (inositol polyphosphate-5-phosphatase) and *Socs2* (suppressor of cytokine signaling 2) are also highly expressed in LTR HSC. Because the expression of Socs2 increases after cytokine stimulation, it is possible that LTR HSC may have recently experienced cytokine stimulation. Alternatively, the high basal levels of genes such as *Socs2* and *Inpp5d* in LTR HSC may render these cells insensitive to cytokines that promote differentiation and thus bias cytokine signaling in LTR-HSC toward self renewal.

Twenty genes were identified that were anticorrelated with stem cell activity (Table 3), including several genes related to adhesion and mobility of cells. These include protein tyrosine phosphatase O (Ptpro), a transmembrane PTPase expressed in epithelial cells, which is also expressed in lymphocytes as an alternatively spliced isoform lacking the extracellular domain; the lymph node homing receptor L-selectin (Sell), and the chemokine Ccl9/MIP1-γ, which binds CCR1 and mediates osteoclast and dendritic cell chemotaxis (29). Because HSC reside in the BM in the proximity of osteoclasts (30, 31), this chemokine may attract CCR1-expressing osteoclasts and facilitate osteoclast–HSC interactions. Because these homing and adhesion molecules are induced during the transition from LTR HSC to STR HSC, they are likely to promote changes in the trafficking of STR HSC or facilitate the recruitment of additional cell populations important for differentiation.

In addition to these characterized genes, we also sought to characterize several previously unstudied genes whose expression pattern correlates with HSC activity. Protein sequences from the 17 genes with unknown functions in Tables 2 and 3 were used to search for functional domains and similar proteins in the National Center for Biotechnology Information protein databases (Table 4, which is published as supporting information on the PNAS web site). Eight of the 17 genes with unknown functions positively correlate with HSC activity and have no similarity to any previously reported proteins or motifs. The eight remaining genes are either similar to hypothetical proteins of unknown function or contain recognizable protein domains.

The expression of nine known transcription regulation factors was found to correlate positively with LTR HSC activity. These include *Cited2*, *GATA3*, *Hdac3*, *Irf6*, *Jun B*, *Nmyc1*, *Rnps1*, *Xbp1*, and *Zfp292*. Little is known regarding the role of these specific transcription factors in the control of HSC biology; however, targeted disruption of either *GATA3* or *Cited2* results in early embryonic lethality due to defective cardiac development (*Cited2*) (32) or innervation (*GATA3*) (33). Because HSC are also mesodermally derived, it is possible that these essential transcription factors may play a role in regulating HSC development and differentiation.

To determine whether any of the differentially expressed transcription factors are themselves regulating transcription in LTR HSC, we performed a search of putative upstream regulatory regions (10 kb upstream of start codons) of the interrogated genes for binding sites of the nine transcription factors. Statistical analysis of these results revealed that only the binding sites of GATA3 (NNGATARNG) were significantly enriched ($P < 0.05$) within the 210 differentially expressed genes. Potential GATA3 target genes along with the number of putative GATA3-binding sites are listed in Table 5, which is published as supporting information on the PNAS web site. Interestingly, this list contains a large fraction (20 of 52) of the genes whose expression positively correlated with HSC activity, suggesting the possibility that GATA3 may play an important role in the control of LTR HSC biology. A small number of genes (3 of 20) whose expression is negatively correlated with HSC activity also contained GATA3-binding sites, suggesting the possibility that low levels of GATA3 expressed in STR HSC may influence gene expression at later stages.

Confirmation of Microarray Expression Data. To confirm the observed expression differences, fresh RNA samples were prepared from independently sorted stem cell populations and used for semiquantitative RT-PCR verification for three of the LTR HSC genes identified. These included the transcription factors *GATA3* and *Jun B*, as well as the thrombopoietin receptor *c-Mpl*, a gene

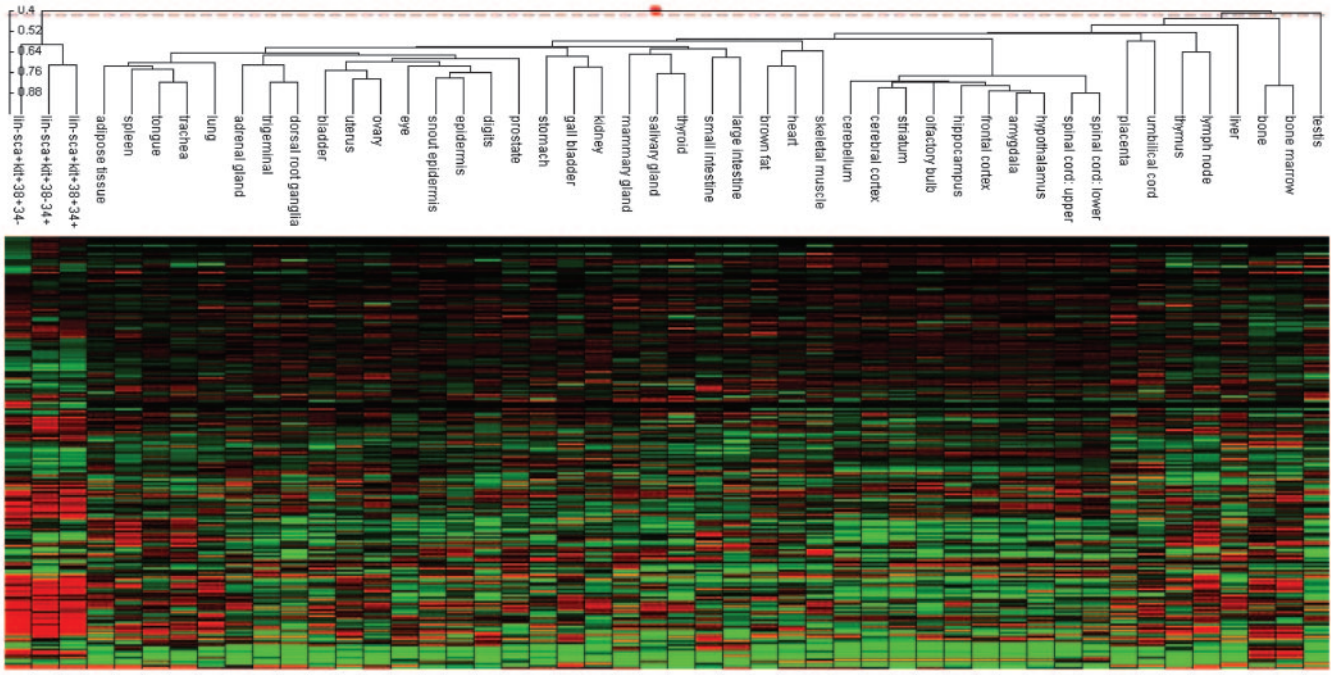

Fig. 2. Hierarchical clustering of differentially expressed genes in different tissues. A hierarchical clustering tree illustrates the expression similarity of the 210 genes among the three HSC subsets and 45 other tissues. The expression levels of the 210 genes are indicated by the color bar: green, low; black, medium; and red, high.

previously shown to be differentially expressed in LTR HSC. As shown in Fig. 7, which is published as supporting information on the PNAS web site, this analysis demonstrated that all three mRNAs are expressed at a significantly higher level in 38^+34^- (LTR HSC) cells compared with the other two HSC subsets. The intensity and expression pattern of the genes selected for verification correlated well with the microarray data. Although these examples represent only a fraction of the total genes analyzed, the high correlation of the RT-PCR analysis with our microarray expression analysis indicates that the data derived are highly reproducible.

Hierarchical Clustering. The previous analysis examined RNA expression within the HSC compartment. Next, we sought to determine whether genes differentially expressed within the HSC compartment are also expressed in other tissues. To perform this analysis, we compared the gene expression levels of the 210 differentially expressed HSC genes with a database composed of 45 normal tissues (available at expression.gnf.org) (26). Hierarchical clustering of these data were used to group tissues and genes with similar expression patterns (Fig. 2). The three HSC subsets formed a distinct branch in this analysis, with LTR-enriched 38^+34^- cells forming a discrete branch compared with the STR cells (38^+34^+ and 38^-34^+). This clustering pattern agrees with the stem cell activity pattern within the three subsets. Importantly, the HSC samples do not cluster near the bone or BM samples, suggesting that the differentially expressed HSC genes are not BM-related. This analysis also demonstrates that the majority of these genes are not ubiquitously expressed, although most are expressed at comparable levels in at least one other tissue. Three of the 72 genes have their peak expression within the HSC compartment. These were: the scaffolding protein Gab1 (GRB2-associated-binding protein 1); the gene A430017F18, which displays the highest level of expression in the LTR-enriched $CD38^+CD34^-$ cells; and the *Pdgfrb* gene (platelet-derived growth factor receptor, β polypeptide), which peaks within the 38^+34^+ STR HSC subset. Although the majority of these 72 genes are also expressed at comparable levels in other tissues, it is important to note that in many cases the level of expression in HSC subsets was at or near the peak expression determined for these genes across the entire 45-tissue panel. The high relative expression within HSC of this subset of genes suggests that they are likely to play an important role in the biology of HSC.

Comparison of Stem Cell Microarray Studies. Two other groups have used identical microarrays to search for stem cell and HSC-restricted genes (21, 22). Curiously, however, analysis of the resulting stem cell-restricted gene lists revealed that they have little in common with each other (21, 22) or with our own list of 72 LTR and STR HSC genes. This observation could reflect either differences in the methods used to define and purify HSC or differences in the analysis methods used to identify HSC-enriched genes. To address this question, we obtained the raw data files from the other two studies and used identical analysis methods to look for HSC-restricted genes among the three HSC profiling studies. Gene expression data were extracted with MAS 5 (Affymetrix) from our data set and image files provided by the Lemischka and Melton groups and normalized to the same BM sample. Genes that were at least 2-fold enriched in HSC compared with BM were identified in each of the data sets, and the resulting gene lists were compared. Approximately 10% (1,230) of the interrogated genes were identified as HSC-enriched in any one of the subsets, with 34% (423) of the 1,230 genes satisfying this criterion in at least two of the studies and 13% (160) of the 1,230 genes identified in all three studies (Fig. 3).

This degree of overlap is significantly higher than that obtained by direct comparison of the published gene lists that were generated from each study (which used different analysis methods). The results indicate that both the analysis methods and different HSC isolation protocols and/or technical differences all contribute to the low overlap of the three studies. Regardless of the cause for differences among the studies, that genes were identified in three independent studies, which each used different HSC isolation procedures and technical methods, highlights that these genes are of considerable interest for further analysis.

Hierarchical clustering was performed to compare gene expression among different tissues, including HSC, in the three data sets (Fig. 4). The comparison included five different stem cell popula-

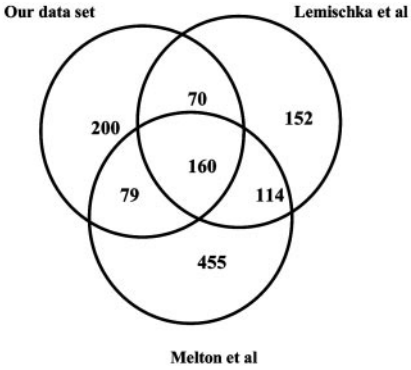

Fig. 3. Comparison of the three data sets with the same analysis method. Data sets from three different studies [Lemishka (21), Melton (22), and this study] were normalized to the same BM samples and analyzed with the same method. There are 160 genes selected by all three data sets, and 50% of the selected genes in each data set have also been selected in other data sets.

tions in the three data sets and 45 other body tissues (data provided by GNF). Expression patterns are clustered among tissues and then clustered among genes. All five stem cell populations clustered together, which further indicated that their expression profiles are different from the expression profiles of other body tissues. This analysis indicates that 69 of the 160 genes are up-regulated only in HSC populations (Table 6, which is published as supporting information on the PNAS web site). Inspection of this list reveals the presence of many of the known markers for HSC, including the cytokine receptors Mpl, c-kit, and Flt3.

Discussion

The mRNA levels of 12,000 probe sets (≈8,000 genes) in the three murine HSC subsets (LTR, 38^+34^-; STR, 38^+34^+ and 38^-34^+) were measured by microarray analysis. The close relationship of the three HSC subsets makes the differentially expressed genes among the three subsets most likely related to long-term stem cell activity. Principal component analysis of these differentially expressed genes indicated that the two major expression patterns resembled the stem cell activity of the three subsets; expression was significantly higher or lower in 38^+34^- (LTR) HSC compared with the other two subsets (38^+34^+ and 38^-34^+cells), whereas there was little difference between the 38^+34^+ and 38^-34^+ STR HSC subsets themselves. These results suggest that long-term stem cell activity is regulated by a group of genes differentially expressed in the 38^+34^- LTR HSC, either up- or down-regulated. A total of 210 genes were differentially expressed, with 72 genes having expression patterns correlated with stem cell activity among the three subsets: high in LTR HSC and low in STR HSC. Many known stem cell-related genes are included in these two groups of genes. These known stem cell-related genes further validate the microarray data and analysis methods.

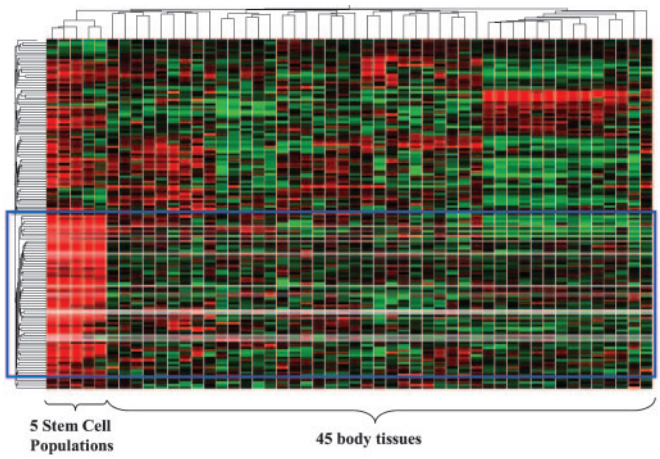

Fig. 4. Clustering of stem cells with other body tissues indicates that 69 of the 160 selected stem cell genes up-regulated only in the five stem cell populations. The blue box indicates genes up-regulated exclusively in stem cell populations. From left to right, the five stem cell populations are: Lin^- Sca^+ Kit^+ $CD34^-$ SP cells [Melton *et al.* (22)], 'Lin$^-$ Sca$^+$ Kit$^+$ Rho^{low} [Lemischka *et al.* (21)], Lin^- Sca^+ Kit $CD38^+$ $CD34^-$, Lin^- Sca^+ Kit^+ $CD38^+$ $CD34^+$, and Lin^- Sca^+ Kit^+ $CD38^-$ $CD34^+$ (this study). The 45 body tissue data were provided by GNF and, from left to right, are: adipose, trachea, ovary, uterus and bladder, umbilical cord, lung, placenta, adrenal gland, BM, bone, lymph node, spleen, thymus, brown fat, heart, skeletal muscle, digits, epidermis, snout epidermis, tongue, gall bladder, liver, kidney, large intestine, small intestine, stomach, mammary gland, salivary gland, thyroid, prostate, amygdala, frontal cortex, hippocampus, cerebral cortex, olfactory bulb, striatum, hypothalamus, spinal cord lower, spinal cord upper, cerebellum, dorsal root ganglion, trigeminal, testis, and eye.

Among the 210 differentially expressed genes, we were particularly interested in the following categories: membrane-associated proteins, signal transduction pathway proteins, and transcription regulation factors. Membrane proteins can be used as markers in FACS analyses and can be directly bound by antibody, whereas signal transduction pathway proteins and transcription regulation factors can reveal the regulatory network of LTR activity. We found nine membrane-associated proteins (see Tables 2 and 3 for detailed information).

Because transcription factors regulate multiple genes, and stem cell activity is the result of orchestrated multiple gene interactions, manipulation of stem cell activity is most likely achieved by manipulating transcription regulation factors. There are 10 annotated transcription factors among the selected stem cell genes; 9 are stem cell activity-correlated genes (*Cited2*, *GATA3*, *Hdac3*, *Irf6*, *Jun B*, *Nmyc1*, *Rnps1*, *Xbp1*, and *Zfp292*), and one is an anticorrelated gene (*Satb1*). These genes are discussed in detail in the *Supporting Text.* These 10 transcription regulation factors are involved in various regulatory mechanisms and pathways; these data suggest that LTR HSC activity, at least in part, is regulated by these factors.

There is strong evidence for the involvement of GATA3 in stem cell activity. GATA 3 is a zinc-finger transcription factor that regulates IL-4, -5, and -13 (34–36). GATA3-deficient murine embryonic stem cells exhibit an enhanced capacity to differentiate into adipocytes, and defective GATA3 expression is associated with obesity (37). Microarray data indicate that GATA3 is highly expressed in the LTR HSC subset but not in the other two subsets. These data suggest that GATA3 might keep LTR HSC from differentiating and thus allow them to maintain their LTR status. This result supports the stem cell maturation pathway proposed by Zhao *et al.* (19). In murine repopulation assays, retroviral vector-forced high-level expression of GATA3 in Sca^+ Kit^+ donor BM cells resulted in a lower level of differentiated blood cells in recipients (38). This result augments the suggestion that GATA3 may play a role in inhibiting differentiation of HSC. A sequence analysis of 10 kb upstream of the 8,000 interrogated genes indicates that many stem cell-related genes have GATA3-binding sites, including *Mpl* and *Kit* (Table 5). Among the selected potential stem cell-related genes, 30% of the stem cell-correlated genes (*9030401P18Rik*, *Cited2*, *Fzd4*, *Gtpi-pending*, *Hdac3*, *Igtp*, *Iigp-pending*, *Inpp5d*, *Irf6*, *kit*, *lcn7*, *Mpl*, *Nedd4*, *serpinb6*, *Xbp1*, and *Zfp292*), and 21% of the anticorrelated genes (*Mpo*, *Rev11*, *Satb 1*, and *Sell*) have GATA3-binding sites. GATA3 might play a role in regulation of these genes. Therefore, GATA3 is a good candidate gene for *in vitro* manipulation of stem cell activity.

There are also 17 genes with unknown functions among the 72 listed potential stem cell-related genes. Because little information of stem cell activity is known today, these genes might well play an important role in stem cell activity.

Previous microarray profiling studies were performed by comparing stem cell-enriched populations with differentiated cell populations (20–24). Because of the hierarchical structure of the HSC compartment and the heterogeneity of the total cell population, which includes many different types of cells, it is difficult to associate gene expression differences with stem cell activity, particularly LTR activity, in these studies. The most interesting stem cell genes are those differentially expressed between LTR and STR HSC, i.e., before progenitor lineages are committed. In this study, we have investigated the expression profiles of LTR and STR HSC within the stem cell compartment. To identify genes that might regulate long-term stem cell proliferation activity, the differentially expressed genes were further selected by correlation of expression patterns with stem cell activity among the three subsets. The close genotype/phenotype of the analyzed subsets of cells and the strict selection criteria used in our study have resulted in a strong association among differentially expressed genes and stem cell activity. That many known stem cell-related genes were identified by our selection method confirms the strong association among differentially expressed genes and stem cell activity in this study.

It is troubling that two previously published data sets that also sought to identify "HSC-enriched genes" had little in common in the resulting HSC genes. Thus it was important to compare our LTR-HSC genes with those identified in previous reports. Of the 52 genes identified in this study that correlated positively with LTR-HSC activity, approximately half were also present in at least one of the two previously published data sets, which compared HSC to BM (21, 22). Thirteen of the 25 genes were selected in all three data sets, and each data set has a number of genes that are not identified in the other two data sets (Fig. 5).

Twenty-seven of the 52 stem cell activity positively correlated genes, despite being up-regulated in LTR HSC compared with STR HSC, did not have higher expression when compared with BM cells. These 27 genes might not be stem cell-specific genes. However, it seems plausible that genes that play a role in regulating LTR activity inside the stem cell compartment may also play a role in other cell types and tissues. Therefore, these genes might not be HSC-specific but may still be LTR HSC-specific.

That nearly half of the genes identified in the present study as LTR-HSC-specific were not identified in previous studies highlights the impact of the reference population and methods used to identify HSC-enriched genes and suggests that comparing HSC with different samples, i.e., whole BM or different progenitor cells, will dramatically impact the gene identification process. When all three studies were analyzed in an identical manner by using the same sample for normalization and criteria for selection of HSC enriched genes, the overlap of stem cell genes in the lists was ≈50% between any two studies. Because each study used different stem cell populations (different HSC definitions) and were performed by different laboratories, this high degree of overlap suggests these microarray experiments indeed reflect the true gene expression profiles of different cell populations. Hierarchical clustering of these data demonstrated that all five stem cell populations were clustered together and separated from the other 45 body tissues, further supporting this notion.

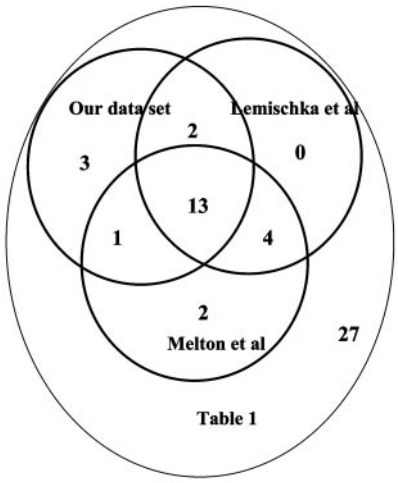

Fig. 5. Comparison of different stem cell selection methods. Twenty-five of the 52 stem cell genes selected by positive correlation of *in vivo* stem cell activity to gene expression also were selected by up-regulation in HSC populations compared with BM cells.

Collectively, these data demonstrate that LTR-HSC-enriched genes can be reproducibly identified via RNA microarray analysis. These genes and others identified in future more comprehensive genome-wide surveys of this type should provide fertile ground for subsequent experiments directed at determining the role of these candidate stem cell regulatory genes in directing the process of stem cell self renewal, death, and differentiation.

This work was supported by grants from the G. Harold and Leila Y. Mathers Charitable Foundation and Farmal Biomedicine LLC. We thank Drs. Doug Melton (Harvard University) and Ihor Lemischka (Princeton University) for providing microarray data and Drs. Lemischka and Natalia Ivanova (Princeton University) for reviewing the manuscript.

1. Zhong, R. K., Astle, C. M. & Harrison, D. E. (1996) *J. Immunol.* **157,** 138–145.
2. Harrison, D. E. & Zhong, R. K. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10134–10138.
3. Jones, R. J., Celano, P., Sharkis, S. J. & Sensenbrenner, L. L. (1989) *Blood* **73,** 397–401.
4. Harrison, D. E., Jordan, C. T., Zhong, R. K. & Astle, C. M. (1993) *Exp. Hematol.* **21,** 206–219.
5. Harrison, D. E., Stone, M. & Astle, C. M. (1990) *J. Exp. Med.* **172,** 431–437.
6. Uchida, N., Jerabek, L. & Weissman, I. L. (1996) *Exp. Hematol.* **24,** 649–659.
7. Morrison, S. J. & Weissman, I. L. (1994) *Immunity* **1,** 661–673.
8. Spangrude, G. J., Heimfeld, S. & Weissman, I. L. (1988) *Science* **241,** 58–62.
9. Novelli, E. M., Ramirez, M. & Civin, C. I. (1998) *Leukemia Lymphoma* **31,** 285–293.
10. Zanjani, E. D., Almeida-Porada, G., Livingston, A. G., Zeng, H. & Ogawa, M. (2003) *Exp. Hematol.* **31,** 406–412.
11. Zanjani, E. D., Almeida-Porada, G., Livingston, A. G., Porada, C. D. & Ogawa, M. (1999) *Ann. N. Y. Acad. Sci.* **872,** 220–231.
12. Zanjani, E. D., Almeida-Porada, G., Livingston, A. G., Flake, A. W. & Ogawa, M. (1998) *Exp. Hematol.* **26,** 353–360.
13. Sato, T., Laver, J. H. & Ogawa, M. (1999) *Blood* **94,** 2548–2554.
14. Osawa, M., Hanada, K., Hamada, H. & Nakauchi, H. (1996) *Science* **273,** 242–245.
15. Goodell, M. A., Rosenzweig, M., Kim, H., Marks, D. F., DeMaria, M., Paradis, G., Grupp, S. A., Sieff, C. A., Mulligan, R. C. & Johnson, R. P. (1997) *Nat. Med.* **3,** 1337–1345.
16. Tajima, F., Deguchi, T., Laver, J. H., Zeng, H. & Ogawa, M. (2001) *Blood* **97,** 2618–2624.
17. Randall, T. D., Lund, F. E., Howard, M. C. & Weissman, I. L. (1996) *Blood* **87,** 4057–4067.
18. Dagher, R. N., Hiatt, K., Traycoff, C., Srour, E. F. & Yoder, M. C. (1998) *Biol. Blood Marrow Transplant.* **4,** 69–74.
19. Zhao, Y., Lin, Y., Zhan, Y., Yang, G., Louie, J., Harrison, D. E. & Anderson, W. F. (2000) *Blood* **96,** 3016–3022.
20. Terskikh, A. V., Easterday, M. C., Li, L., Hood, L., Kornblum, H. I., Geschwind, D. H. & Weissman, I. L. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 7934–7939.
21. Ivanova, N. B., Dimos, J. T., Schaniel, C., Hackney, J. A., Moore, K. A. & Lemischka, I. R. (2002) *Science* **298,** 601–604.
22. Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R. C. & Melton, D. A. (2002) *Science* **298,** 597–600.
23. Park, I. K., He, Y., Lin, F., Laerum, O. D., Tian, Q., Bumgarner, R., Klug, C. A., Li, K., Kuhr, C., Doyle, M. J., *et al.* (2002) *Blood* **99,** 488–498.
24. Phillips, R. L., Ernst, R. E., Brunk, B., Ivanova, N., Mahan, M. A., Deanehan, J. K., Moore, K. A., Overton, G. C. & Lemischka, I. R. (2000) *Science* **288,** 1635–1640.
25. Scherer, A., Krause, A., Walker, J. R., Sutton, S. E., SerAn, D., Raulf, F. & Cooke, M. P. (2003) *BioTechniques* **34,** 546–550.
26. Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99,** 4465–4470.
27. Willert, K., Brown, J. D., Danenberg, E., Duncan, A. W., Weissman, I. L., Reya, T., Yates, J. R., Nusse, R., Ailles, L., Domen, J., *et al.* (2003) *Nature* **423,** 448–452.
28. Reya, T., Duncan, A. W., Ailles, L., Domen, J., Scherer, D. C., Willert, K., Hintz, L., Nusse, R. & Weissman, I. L. (2003) *Nature* **423,** 409–414.
29. Lean, J. M., Murphy, C., Fuller, K. & Chambers, T. J. (2002) *J. Cell. Biochem.* **87,** 386–393.
30. Schneider, G. B. & Relfson, M. (1988) *Bone* **9,** 303–308.
31. Udagawa, N., Takahashi, N., Akatsu, T., Tanaka, H., Sasaki, T., Nishihara, T., Koga, T., Martin, T. J. & Suda, T. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 7260–7264.
32. Yin, Z., Haynie, J., Yang, X., Han, B., Kiatchoosakun, S., Restivo, J., Yuan, S., Prabhakar, N. R., Herrup, K., Conlon, R. A., *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99,** 10488–10493.
33. Lim, K. C., Lakshmanan, G., Crawford, S. E., Gu, Y., Grosveld, F. & Engel, J. D. (2000) *Nat. Genet.* **25,** 209–212.
34. Zhang, D. H., Yang, L. & Ray, A. (1998) *J. Immunol.* **161,** 3817–3821.
35. Lee, G. R., Fields, P. E. & Flavell, R. A. (2001) *Immunity* **14,** 447–459.
36. Lavenu-Bombled, C., Trainor, C. D., Makeh, I., Romeo, P. H. & Max-Audit, I. (2002) *J. Biol. Chem.* **277,** 18313–18321.
37. Tong, Q., Dalgin, G., Xu, H., Ting, C. N., Leiden, J. M. & Hotamisligil, G. S. (2000) *Science* **290,** 134–138.
38. Chen, D. & Zhang, G. (2001) *Exp. Hematol.* **29,** 971–980.