Mapping Cellular Hierarchy by Single-Cell Analysis of the Cell Surface Repertoire

Guoji Guo,1 Sidinh Luc,1 Eugenio Marco,3 Ta-Wei Lin,4 Cong Peng,1 Marc A. Kerenyi,1 Somir Beyaz,1 Woojin Kim,1 Jian Xu,1 Partha Pratim Das,1 Tobias Neff,6 Keyong Zou,6 Guo-Cheng Yuan,3 and Stuart H. Orkin1,2,4

1Division of Pediatric Hematology/Oncology, Boston Children’s Hospital and Dana-Farber Cancer Institute, Harvard Stem Cell Institute, Harvard Medical School, Boston, MA 02115, USA
2Howard Hughes Medical Institute, Boston, MA 02115, USA
3Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA 02115, USA
4Molecular Genetics Core Facility, Children’s Hospital Boston, Boston, MA 02115, USA
5Pediatric Hematology/Oncology/BMT, University of Colorado, Aurora, CO 80045, USA
6Boston Open Labs, Cambridge, MA 02138, USA

Correspondence: stuart_orkin@dfci.harvard.edu
http://dx.doi.org/10.1016/j.stem.2013.07.017

SUMMARY

Stem cell differentiation pathways are most often studied at the population level, whereas critical decisions are executed at the level of single cells. We have established a highly multiplexed, quantitative PCR assay to profile in an unbiased manner a panel of all commonly used cell surface markers (280 genes) from individual cells. With this method, we analyzed over 1,500 single cells throughout the mouse hematopoietic system and illustrate its utility for revealing important biological insights. The comprehensive single cell data set permits mapping of the mouse hematopoietic stem cell differentiation hierarchy by computational lineage progression analysis. Further profiling of 180 intracellular regulators enabled construction of a genetic network to assign the earliest differentiation event during hematopoietic lineage specification. Analysis of acute myeloid leukemia elicited by MLL-AF9 uncovered a distinct cellular hierarchy containing two independent self-renewing lineages with different clonal activities. The strategy has broad applicability in other cellular systems.

INTRODUCTION

Cellular differentiation is commonly depicted as a sequential binary commitment process through multiple intermediate stages. By combining markers, different types of stem and progenitor cells have been identified in various systems. Further enrichment and analysis of these populations has aided appreciation of stepwise lineage specification. However, the choice of a small number of markers for enrichment of cell populations often masks potential heterogeneity and may bias an understanding of the cellular hierarchy.

Extensive cellular and molecular studies have contributed to the characterization of vertebrate hematopoietic differentiation pathways (Orkin and Zon, 2006). The prospective identification of mouse hematopoietic stem and progenitor cells (Muller-Sieburg et al., 1986; Visser et al., 1984) and further separation of hematopoietic stem cells (HSCs) from multipotent progenitors (MPPs) (Kiel et al., 2005; Morrison et al., 1997; Morrison and Weissman, 1994; Osawa et al., 1996) suggested a cellular hierarchy, whereby self-renewing HSCs produce transiently amplifying MPP. Subsequent identification of common lymphoid (CLP) and myeloid progenitors (CMP) (Akashi et al., 2000; Kondo et al., 1997) led to the conventional model in which lineage specification first takes place as a lymphoid (CLP) versus myeloid (CMP) bifurcation event. Several findings, however, challenge this simple view. They describe heterogeneity of early progenitor populations and posit that lymphomyeloid lineage commitment may occur upstream of the separation of CLP and CMP (Adolfsson et al., 2005; Arinobu et al., 2007; Pronk et al., 2007). Different marker panels and fluorescence-activated cell sorting (FACS) purification schemes have prevented resolution of these alternative models.

Cells within leukemias are also believed to form a hierarchy, yet descriptions of leukemia stem cells (LSCs) are often seemingly contradictory. Original support for the existence of LSCs rested on the observation that only a rare subset of human acute myeloid leukemia (AML) cells, characterized by a surface phenotype similar to that of hematopoietic stem/progenitor cells, was competent to reinstantiate disease upon transplantation in immunodeficient mice (Bonnet and Dick, 1997). More recent findings derived from a mouse model of AML driven by MLL-AF9 suggest that LSCs display a granulocyte-monocyte progenitor (GMP)-like phenotype and stand at the top of the leukemia hierarchy (Krivtsov et al., 2006). Other reports argue that leukemia cells with immunophenotypes of lineage cells may perform as functional LSCs in mouse AML (Gibbs et al., 2012; Somervaille and Cleary, 2006), adding to the complexity of the leukemia hierarchy.

Single-cell gene expression analysis offers potential to resolve these issues. Recently, several hallmark technical advances...
have been achieved. Single-cell messenger RNA (mRNA) sequencing strategies enable whole transcriptome analysis from individual cells (Islam et al., 2012; Ramsköld et al., 2012; Tang et al., 2009, 2010). Alternatively, single-cell mass cytometry constitutes a powerful system for multiplexed gene expression analysis at the protein level (Bennett et al., 2011). When both sample size and assayed gene number are taken into consideration, high-throughput single-cell quantitative PCR (qPCR) represents a favorable option (Buganim et al., 2012; Dalerba et al., 2011; Guo et al., 2010; Moignard et al., 2013). The qPCR approach is highly sensitive in detecting quantitative differences at mRNA level (Guo et al., 2010).

Here, we sought to improve the utility and value of current single-cell qPCR technology by increasing its throughput so as to assess expression of nearly all commonly used cell surface markers. We illustrate how this enhanced approach provides biological insights into normal and leukemic hematopoiesis. The approach we describe should be applicable to other developmental systems and allow for cross-tissue and cross-experiment comparisons. The method allows dissection of heterogeneous populations and the identification of cellular states at single-cell resolution.

RESULTS

Single-Cell Gene Expression Analysis of the Cell Surface Repertoire

By introducing algorithm-based primer design and optimizing the cycling conditions for highly multiplexed PCR, we have increased the capacity of single-cell mRNA sequence-specific preamplification (Figure 1A). In addition, the use of EvaGreen real-time PCR chemistry (Biotium) and melting curve analysis allows for nonspecific signal control during gene-specific qPCR on the BioMark real-time PCR system (Fluidigm). Finally, inclusion of nested primers filters out primer dimer signals (Figure 1A). In the highly multiplexed PCR preamplification, the chance of forming a dimer between a given primer pair of the same gene is actually very low. The subsequent gene-specific qPCR will select and enrich target amplicons, even from extremely low starting materials (Figures S1A and S1B available online). We have designed and optimized a panel of assays to cover all commonly used cell surface markers (Lai et al., 1998) with a total of 280 genes (a few important transcription factors are also included) in establishing an analysis platform for all mouse cell types. After 280 multiplexed, single-cell preamplification, individual gene expression is quantified on the BioMark real-time PCR system (Fluidigm) using three 96.96 dynamic arrays.

To assess the ability of the assay to discriminate different cell types at single-cell level, we used flow cytometry to sort stem cell populations from a broad range of tissues, including neural, prostate, mammary gland, intestinal, and hematopoietic stem cells, according to published protocols (Table S1A), and applied the single-cell assay for gene expression profiling. As shown in Figure 1B, hierarchical clustering of the single cell data faithfully groups cells of the same origin together. The clustering also reveals lineage-specific markers, such as CD56 for neural stem cells, CeaCam2 for prostate stem cells, Icam1 for mammary gland stem cells, Lgr5 for intestinal stem cells, and Httm1 for hematopoietic stem cells (Figure 1C). False positive signal from a no-cell preamplification control is extremely rare and weak (Figure S1C). These results provide initial evidence on behalf of the robustness of the single-cell approach.
Comprehensive Single-Cell Analysis of the Hematopoietic System

We utilized the single-cell assay for a systematic analysis of the mouse hematopoietic system. To enrich stem cell and progenitor cell populations and represent all possible cellular transitional statuses during differentiation, we used FACS to sort the principal hematopoietic compartments of the bone marrow by use of the cell surface markers Kit and Sca1, as well as a lineage (Lin) cocktail that recognizes mature cells of the major hematopoietic cell lineages, including T lymphocytes, B lymphocytes, monocytes/macrophages, granulocytes, and erythrocytes. Sorted populations include Lin+ Sca1+Kit+ (LSK), Lin-Sca1+Kit+ (LSK-), Lin-Sca1+Kit+ (LSK-) populations (Figure 2A). In addition, we sorted conventionally defined stem and progenitor cell types (including HSCs, MPP, CMP, CLP, common dendritic cell progenitor [CDP] [Oinai et al., 2007], megakaryocyte/erythroid progenitor [MEP], and GMP) as well as a set of differentiated cell types (Figure S2; Table S1). A similar strategy was used to sort CD4+ T cells, CD8+ T cells, CD4+CD8+ double positive T cells, earliest thymic progenitors (ETP), CD4+CD8- double-negative (DN) 2, DN3, and DN4 stage thymocyte progenitors (Figures 2A and S2; Table S1). An average of around 50 single cells are analyzed for each sorted population (Table S1). We analyzed more than 1,500 single cells throughout the mouse hematopoietic system and quantified all 280 genes for each individual cell (Table S2).

Unsupervised hierarchical clustering of the single cell data set reveals high correlation of gene expression clusters with cell-type clusters (Figure 2B). As highlighted by white boxes, CD11c, CD3, Btk, Kit, CD11b, and Gypa clusters correspond to dendritic, T, B, stem and progenitor, myeloid, and megakaryocytic and erythroid (MegE) lineage cells, respectively. The principal lineage-specific gene clusters are summarized in Table S3. Subclusters also exist within these main clusters. In addition, the quantified mRNA level differences correlate with different FACS sorting schemes; Actb and Gapdh expression levels are relatively consistent (Figure S2A). The clustering data suggest that differential global gene expression signatures at the single level are reproducible in both progenitor and differentiated cell types. The clustering pattern may then be used to identify novel markers and populations.

To visualize the overall pattern of gene expression (280 parameters) at the single-cell level, we used the Gene Expression Dynamics Inspector (GEDI) program (Chang et al., 2008) to generate individual expression maps (Figure 2C). The color of each pixel on the map indicates the centroid value of the gene expression level for each minigene cluster generated by the software. Representative single-cell maps from different populations illustrate how the method can be used to identify and classify virtually all cell types. As an example, we show that an incompletely defined LSK+ population is very heterogeneous, as revealed by the clustering (Figure 2C). According to the single-cell gene expression signature, this population contains not only CLP-like progenitors and B cell progenitors but also plasmacytoid dendritic cells (PDC) (Oinai et al., 2007) and neutrophils (Neill et al., 2010). Changes during cellular differentiation may be visualized from the maps. The gradual transition of the GEDI map from sorted MEP to CD71+ erythroid progenitors and then to Ter119+ cells provides an example. Interestingly, we have identified bone marrow MPP-like cells in the spleen (MPP-SP) and thymus (MPP-TH), consistent with the circulation of hematopoietic progenitor cells throughout the body.

Heterogeneity of Hematopoietic Progenitor Cell Types

Having established a robust methodology for single-cell analysis, we proceeded to examine the classically defined hematopoietic progenitor cell types (Figures 3A–3E). Each of these progenitor types reveals marked heterogeneity. For example, we profiled 47 single CMP cells, originally defined by the Lin-IL7R-Kit+Sca1-CD34-CD16/CD32lo profile, and ranked 280 genes by their standard deviation across all CMP samples. The top four most variable genes were CD53, Self, CD55, and Flt3 (Figure 3A). Hierarchical clustering of these variable genes reveals two principal populations with different gene expression patterns. To address whether these gene expression differences reflect stochastic noise (Chang et al., 2008), we applied violin plot analysis to visualize the distribution of gene expression levels. In this plot, the Y- and X-axes correspond to the gene expression level and distribution frequency, respectively. Theoretically, expression noise should exhibit unimodal distribution around a reference level, whereas a multimodal distribution should indicate quantitative differences. As expected, the distributions of Actb and Gapdh levels are unimodal, with a very narrow peak indicative of low variation (Figures 3A and S3B). In contrast, the top four most variable genes within the CMP population show clear bimodal distribution. To confirm mRNA level differences at the protein level, we used available antibodies to the surface marker CD55 to further analyze the CMP compartment. Flow cytometry validated the heterogeneous nature of the CMP population detected by single-cell qPCR (Figure 3A). We then continued to dissect heterogeneity further in the CD55-CMP population and revealed Csfr1 as one of the most differentially expressed markers (Figure 3B). Comparable analyses were performed for GMP, CLP, MEP, ETP, and CDP. We observed discrete heterogeneity within all populations (Figures 3C–3E). The analysis also reveals dynamic changes in LSK heterogeneity during the aging process (Figure S3C) and permits assessment of the purity of HSCs from different enrichment protocols (Figure S3D). The bimodal distribution of mRNA transcripts is present in all the cell types that we have purified, suggesting extensive unknown heterogeneities. Although the mRNA level expression is not always reflective of protein level expression, we argue that it should be indicative of a cell’s transcriptional state and functional potential.

Mapping Hematopoietic Hierarchy by Computational Lineage Progression Analysis

We hypothesized that the similarity of different single-cell signatures and continuity of transitional states during differentiation could form the foundation of an in silico strategy to organize high-dimensional data into ordered, stepwise cell fate commitment pathways. To accomplish this, we first removed redundancy by extracting the average value of 40 distinct gene expression clusters from the entire data set (Table S3) and then used spanning-tree progression analysis of density-normalized events (SPADE) (Bendall et al., 2011; Qiu et al., 2011) analysis to distill 40 dimensional single-cell data down to a single interconnected cluster of transitional cell populations.
Figure 2. Comprehensive Single-Cell Analysis of the Mouse Hematopoietic System

(A) Single-cell sorting strategy to enrich stem and progenitor cells but to cover all possible populations.

(B) A master heatmap showing the hierarchical clustering of gene expression signatures from 1,500 single cells throughout the hematopoietic system. Each row corresponds to a specific gene; each column corresponds to a particular single cell. Strong correlation between gene and cell clusters are highlighted by white boxes and labeled by cell type-specific clusters. Red to yellow suggest high to middle expression, whereas green to blue suggest low to no expression.

(C) GEDI plot allows for visualization of single-cell global signatures. Examples of single-cell GEDI map from different cell types are presented. Color scale is as described in Figure 1B. The lower right corner, which is always red, corresponds to endogenous control genes that are highly expressed in all single-cell samples. From the Lin-Sca1+Kit+ population, there are clusters of single cells (the red lines separate different clusters in the heatmap of Lin-Sca1+Kit+ single-cell data) with nuocyte signature and PDC signature.

See also Figure S2 and Tables S2 and S7.

The unsupervised computationally constructed hierarchy shows high resemblance to the hematopoietic differentiation lineage tree (Figure 4A). Different cell lineages are readily separated into distinct branches, as revealed by the overlaid expression level of different gene clusters. Branches expressing Kit cluster and Gypa cluster genes correspond to stem and progenitor
and to MegE lineage cells, respectively. The dendritic, macrophage, B cell, and T cell branches, as well as lymphomyeloid progenitor cells, are marked by expression of CD11c, CD11b, Btk, CD3, and FLt3 clusters, respectively. The Gapdh endogenous control cluster is expressed broadly.

In the hierarchy generated from single-cell expression data, the MegE lineage branch is closely connected to the long-term repopulating HSC branch. These data suggest that the MegE lineage separates very early from the lymphomyeloid lineage cells. Upon inspection of the composition of different nodes, we found that phenotypic CMP cells are located on two separate differentiation pathways, with half merged to the MegE lineage and half merged to the lymphomyeloid lineage (Figure 4B). This pattern is inconsistent with the conventionally portrayed, classical differentiation scheme that positions the MegE progenitor after the bifurcation of CMP and CLP and is reminiscent of an alternative model (Adolfsson et al., 2005; Pronk et al., 2007).

To validate this alternate scheme functionally, we sought to predict an early MegE lineage-specific marker from our data resource. We compared gene expression differences between the two separated CMP compartments (CMP1 and CMP2) and identified CD55 as the most differentially expressed MegE marker (Figure 4C). In addition, we found that CD55 expression was strongly correlated with the Gata1 transcription factor (Figure 4C and S4A), a master regulator of MegE lineage specification (Arinobu et al., 2007; Fujisawa et al., 1996; Iwasaki et al., 2003). FACS analyses indicate that Lin-Kit+Sca1+ cells can be separated by CD55 into two main compartments (Figure S4B). To overcome the limitation of traditional two-dimensional gating strategy, we used SPADE analysis to analyze multidimensional FACS data from mouse bone marrow stained with CD55, CD150, CD34, CD16/CD32, Sca1, Kit, and lineage antibodies. We focused on Lin−/Kit+ data points and generated a simplified lineage tree with seven dimensional single-cell profiles. Consistent with our qPCR expression findings, the MegE lineage branch is closely connected with the HSC containing cell cluster nodes (Figures 4D and S4C), confirming early MegE specification.

We next separated CMP (Lin−IL7R-Sca1+Kit+CD34+CD16/CD32+ and MPP (Lin−Sca1+Kit+CD34+) compartments into CD55− and CD55+ subpopulations (Figure 4E) and tested their function using in vitro colony-forming assays. Both CD55+ MPP and CD55+ CMP produce predominantly erythroid and megakaryocytic colonies, whereas few MegE colonies arise from CD55− MPP or CD55− CMP, revealing a functional difference in these early progenitor compartments (Figures 4F and
Single-Cell Analysis of Cell Surface Repertoire

A. KIT Cluster
   - BLNK Cluster
   - CD3 Cluster
   - CD11C Cluster
   - CD11B Cluster

B. DC, Macrophage
   - MegE
   - CMP
   - MPP
   - HSCs

C. Top differentially expressed
   - CD55: 7.87
   - ICAM4: 3.98
   - CD274: 3.32
   - MPL: 3.19
   - TEK: 2.83

D. FACS SPADE analysis with CD55, CD158, CD34, CD1832, Scale, Kit and Lineage antibodies
   - MegE
   - MPP
   - HSCs
   - Myeloid

E. Reconstitution of CD55+ and CD55- CMP
   - CD61+
   - Mac1+
   - E220+
   - CD3+

F. Colony type
   - CD55- MPP
   - CD55- CWP
   - CD55+ CWP
   - CD55+ CMP

G. 1w, 2w, 3w, 6w reconstitution analysis

H. Reconstitution of CD55+ and CD55- MPP
   - CD61+
   - Mac1+
   - E220+
   - CD3+

(legend on next page)
S4D). In order to confirm the early MegE separation in vivo, we used Actβ-GFP mice for transplantation studies (Figure 4G). CD55+ CMPs transiently give rise to CD61+ platelets, whereas CD55− CMPs produce mainly myeloid cells (Figure 4H). CD55+ MPPs achieved more than 50% platelet reconstitution, whereas there was no reproducible contribution of CD55− MPPs to CD61+ platelets (Figures 4I and S4E). Importantly, CD55− CMPs and CD55+ MPPs failed to produce platelets in vivo, whereas CD150− progenitors exhibited robust MegE potential (Prok et al., 2007), suggesting that CD55 is an improved marker for separating early MegE progenitors. In conclusion, by computational analysis of single-cell data, we have predicted and validated CD55 as a marker to establish a functional separation between early MegE and lymphomyeloid differentiation at both CMP and MPP stages.

Genetic Network Construction by Single-Cell Analysis

To explore potential molecular mechanisms underlying early hematopoietic lineage specification, we designed primers to assay expression of an additional 180 genes, including lineage-specific transcription factors, epigenetic modifiers, and cell-cycle regulators. We assayed single cells from HSCs (CD48−CD34−CD150+LSK), MPP (CD34+LSK), CMP, MEP, GMP, and CLP populations (Table S4) and calculated gene expression covariance across the data set to uncover hidden regulatory links. We then used Cytoscape software to integrate expression correlations with published chromatin immunoprecipitation sequencing (ChIP-seq) binding data sets for 11 major stem cell transcriptional regulators (including Scl/Tal1, Lyt1, Lmo2, Gata2, Runx1, Meis1, PU.1, Erg, Fli1, and Gfi1b) from HPC-7 cell line (Wilson et al., 2010). The network (Figure S5A) only depicts links in which the coverage was above 0.1 for correlated genes (green edges) or below 0.1 for anticorrelated genes (red edges). The network contains 76 nodes, connected through 71 edges between correlated genes and 74 edges between anticorrelated genes. Figure 5A highlights the transcription factor components of the complete network. As revealed, Gata2, a central hematopoietic stem cell regulator (Tsai et al., 1994; Wilson et al., 2010), lies at the core of the lineage specification pathway (Figure 5A), positively correlates with a MegE lineage module (characterized by Gata1, Gfi1b, Nfia, and Klf1), and negatively correlates with a lymphomyeloid module (characterized by Flt3, Sali, Cebpa, and Notch1). This is also depicted on a gene-to-gene correlation heatmap in Figure 5B. Time-course, single-cell tracing experiments suggest that upregulation or downregulation of Gata2 marks the first molecular event during colony formation (Figure S5B). The correlation in expression level between Gata2 and Gata1 is maintained during both in vivo and in vitro differentiation (Figure S5C). As revealed in Figure 5C, Gata2, Runx1, Meis1, Scl, Lyt1, and Lmo2 co-occupy at Gata1 and Gfi1b regulatory regions (Figure 5C). The stem cell transcription factor Gata2 occupies regulatory elements of multiple MegE lineage-related genes, as well as HSC-enriched genes (Figure S5D).

As a functional test of this predicted genetic network, we examined the consequences of perturbation of the level of Gata2. Because Gata2−/− embryos die early due to hematopoietic failure (Tsai et al., 1994), we analyzed gene expression changes in viable Gata2+/− mice at single-cell resolution. Consistent with a previous report (Rodrigues et al., 2005), we observed a reduction in the size of the LSK population in Gata2 heterozygous mice as compared with wild-type. Single-cell gene expression analysis of Gata2+/+ and Gata2−/− LSK reveals that haploinsufficiency is associated with an altered regulatory network during early lineage differentiation (Figures 5D and S5E), revealing sensitivity of the network to modest quantitative changes in Gata2 expression. Haploinsufficiency for Gata2 leads to downregulation of the MegE marker Gfi1b and Gata1 and upregulation of lymphomyeloid markers, including Flt3, Sali, CD34, CD33, and Cebpa in hematopoietic stem and progenitor cell populations (Figures 5D and S5E). Taken together, single-cell-level gene expression in combination with functional studies validates a genetic network underlying early differentiation of MegE lineage from the lymphomyeloid lineage.

MegE Priming in the Most Primitive HSCs

In single-cell, in vitro tracing experiments (Figure S5B), we noticed that pure megakaryocytic colonies are the first to emerge in cultures of HSCs. These results encouraged us to investigate the heterogeneity and existing MegE network within the most primitive HSC (CD48−CD34−CD150+LSK) population. Remarkably, a MegE module, characterized by expression of...
Figure 5. Genetic Regulation during HSC Differentiation

(A) A genetic network constructed by Cytoscape using transcription factor ChIP-seq binding information and single cell-level gene expression correlation data. It highlights transcription factors components within the complete network in Figure S5A. Green arrow corresponds to positive correlation, whereas red arrow corresponds to anticorrelation. The width of the line corresponds to absolute value of the covariance between two linked gene nodes.

(B) Gene-to-gene correlation heatmaps containing HSC, MegE, myeloid, and lymphoid modules in MPP and CMP.

(C) HSC module transcription factors co-occupy Gata1 and Gfi1b upstream region.

(D) Gene expression level distribution of LSK single cells from wild-type and Gata2 heterozygous mice are presented with violin density plots. The percentages of cells with positive expression levels are marked on the violin plots. Note the decrease in MegE-primed cells (Gfi1b+) or Gata1+ cells and increase in lymphomyeloid primed cells (Cebpα+, FB3+, CD63+, or Scl+) in the Gata2+/− LSK population.

See also Figure S6 and Tables S4 and S7.

Fli1, CD41, CD150, Gata1, vWF, Mpl, and Gfi1b, maintains high correlation in single HSCs (Figure 6A). MegE lineage-specific gene expression is detected in HSCs purified by different enrichment protocols and further confirmed by single-cell NanoString technology (Figures 6B, S6A, and S6B; Table S5). Such transcriptional priming does not appear to be stochastic but rather controlled by an intertwined HSC regulatory network. We ranked single HSCs by the expression level of Gata2 and compared gene expression between Gata2high HSCs (top 50%) and Gata2int HSCs (bottom 50%). CD150 emerged as a candidate marker for separating HSCs according to different levels of Gata2 expression (Figure 6C), as well as different degrees of MegE priming (Figure S6C).

To confirm these differences, we FACS-sorted HSCs into CD150high and CD150int compartments for gene expression analysis (Figure 6D). Indeed, CD150high HSCs express higher levels of Gata2, Gata1, and CD61, as well as other MegE lineage-related genes (Figure 6E). In colony-forming assays, CD150high HSCs generate greater numbers of MegE lineage-containing colonies than CD150int HSCs (Figure 6F). Similar biased differentiation readouts were also seen in HSCs that were separated by relative expression levels of CD55, CD41, or CD9 (Figure S6D). These results suggest that MegE differentiation bias is already established at the HSC level.

The positive correlation of Gata2 with the MegE priming expression suggests that the regulatory network within HSCs is intrinsically unstable. As such, higher levels of Gata2 in...
Single-Cell Analysis of Cell Surface Repertoire

B. MegE Priming in HSCs

C. Top 5 differentially expressed genes between Gata2high HSCs and Gata2int HSCs

<table>
<thead>
<tr>
<th>Gene</th>
<th>Log2FoldChange</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD150</td>
<td>3.33</td>
</tr>
<tr>
<td>CD41</td>
<td>2.34</td>
</tr>
<tr>
<td>GATA1</td>
<td>2.24</td>
</tr>
<tr>
<td>MYCN</td>
<td>2.14</td>
</tr>
<tr>
<td>GATA2</td>
<td>1.98</td>
</tr>
</tbody>
</table>

D. CD48-LSK

E. Colony count and FACS analysis

H. Reduced Gata1 priming in the Gata2 +/- HSCs (HSC single cells ordered by Gata1 expression)

I. Reduced Gfi1b priming in the Gata2 +/- HSCs (HSC single cells ordered by Gfi1b expression)
HSCs may activate MegE lineage expression and promote MegE lineage skewing. When stained with the full panel of HSC markers, we observed a reduced number of CD150high HSCs in the Gata2 haploinsufficient state (Figure 6G). In addition, in the most primitive HSCs of Gata2+/- mice, we observed a reduction in the number of Gata2+ or Gfi1b+ HSCs, as well as the average level of MegE: priming (Figure 6H). Consistent with these findings, overexpression of Gata2 has been reported to promote MegE differentiation (Huang et al., 2003; Kitaizumi et al., 2006). The characterization of MegE priming in HSCs supports the cellular hierarchy and genetic network derived from single-cell expression data and illustrates the power of single-cell analysis in detecting the earliest regulatory events during stem cell differentiation.

Single-Cell Analysis of the AML Cellular Hierarchy

Having obtained a comprehensive data set in the wild-type hematopoietic system, we next applied the single-cell expression approach to characterization of LSCs in MLL-AF9-driven AML, a clinically relevant model of hematopoietic malignancy (Krivtsov et al., 2006; Neff et al., 2012). In this model, LSCs resemble GMPs and are hence described as LSCMs. Others, however, have described alternative cellular hierarchies of AML (Gibbs et al., 2012; Somervaille and Cleary, 2006). We generated MLL-AF9 primary leukemia in mice (Neff et al., 2012) and profiled single cells of the originally defined LSC population (Lin-Il7r-Klt-Sca1+CD34+CD16/323+), as well as the leukemic Lin- (+) population from bone marrow (Figure 7A and Table S1). As shown in Figure 7B, hierarchical clustering of gene expression data from leukemia cells and the wild-type myeloid cells reveals clear separation of the two groups. The Lin+ clusters closely with a group of LSCM cells, suggesting that lineage marker expression does not define a clear hierarchy in the leukemia. Two strong gene clusters are observed in the leukemia: a Csfr1, Ccr2, Ccr5 cluster and a CD24, Vcam1, CD133 cluster (Figure 7B). We adapted SPADE to analyze the data (Figure 7C). To allow for comparison of the wild-type and leukemia lineages, we extracted 40 clusters from the combined data sets of LSCM, Lin+ GM, and Lin- single cells (Table S6). We then used these clusters to infer lineage hierarchy for both cellular systems. From the over laid expression level of different gene clusters, we observed clear separation of the CD24+ lineage branch and the Csfr1+ lineage branch within the tested leukemia cells (Figure 7C). By comparing the two main leukemia cell type signatures with other hematopoietic cell types, we find that MLL-AF9 leukemia cells display a unique signature with high expression of Lmb1, Lmp2, Il1n1, CD47, and CD33 (Figure 7D). Notably, the leukemia cellular state differs from other hematopoietic cellular states both at the single cell and population levels.

In the previously defined LSCM population, in which LSCMs are highly enriched (Krivtsov et al., 2006), we observed clear heterogeneity. Guidance from single-cell data, we separated the LSCM into two populations using CD24 antibody (Figure 7E). To assess potential functional differences of these two compartments, we transplanted each into sublethally irradiated secondary recipients. Both CD24+ LSCM and CD24+ LSCM are capable of initiating AML (Figure 7F). However, mice transplanted with CD24+ LSCMs exhibited a marked delay in disease progression. Analysis of the bone marrow from secondary leukemia mice indicated that CD24+ leukemia cells and CD24+ leukemia cells maintain their respective signatures and fail to reconstitute each other when clonal expansion (Figure 7G). Thus, CD24 marks two distinct, self-renewing clones within MLL-AF9-driven AML. Further profiling of additional intracellular regulators reveals different genetic programs used by CD24+ LSCM and CD24+ LSCMs (Figure S7A). Interestingly, Ezh2, a core polycomb repressive complex 2 (PRC2) component, is overexpressed in CD24+ LSCMs (Figure S7A). Our analysis also reveals high variation of Ezh2 at the single-cell level, which strongly correlates with Ccn2, Ccnb1, and Ccnb2 expression (Figure S7B). Such correlation may account in part for the more aggressive behavior of the CD24+ Ezh2 high leukemia clone, as compared with the CD24+ Ezh2low leukemia clone. In microarray data of synchronized HeLa cells (Whitfield et al., 2002), Ezh2 expression is lowest in G1 and peaks at S phase (Figure S7C). In addition, many cell-cycle regulators are direct targets of PRC2, as assessed from PRC2 chromatin occupancy data (Figures S7D–S7G). Moreover, inhibition of Ezh2 function with the specific inhibitor GSK126 (McCabe et al., 2012) leads to an increase in G1 phase cells and a decrease in S phase cells in MLL-AF9 cultures (Figure S7H). Our findings are in general agreement with the observation that Ezh2 overexpression correlates with poor prognosis in several tumor types (Cavalli, 2012; McCabe et al., 2012).

DISCUSSION

Single-cell analysis technologies provide a powerful approach to the study of rare cell types and cell heterogeneity. For both genome analysis and transcriptome analysis of single cells, amplification of small amounts of material is required and presents technical challenges. For assessment of gene expression,
single-cell, high-throughput qPCR has several advantages. First, it utilizes one-tube one-step single-cell sequence-specific pre-amplification, which involves minimal sample handling time and allows for high-throughput complementary DNA (cDNA) library generation. Second, the targeted PCR approach enables specific amplification of lowly expressed genes. Finally, Biomark (Fluidigm) microfluidic qPCR permits well-controlled, parallel analysis of 96 single-cell samples. The system minimizes technical variation, allowing for comparison of different samples without normalization. A major challenge relates to primer dimer formation during multiplexed, sequence-specific preamplification step, which generates false-positive signals (Guo et al., 2010). To overcome this obstacle, we have introduced multiplexed primer design, lowered the preamplification primer concentration, and included nested primers to avoid primer dimer signals. These optimizations significantly increased the throughput of single-cell qPCR technology and permitted analysis of the cell surface repertoire in many single cells from a broad range of tissue types. We show with functional validation that such data sets can be used for classification of cell type, dissection of heterogeneity, mapping of cellular hierarchy, and computational construction of genetic networks.

We have applied our analysis to examine cellular lineages of the mouse hematopoietic system. Computational lineage progression analysis provided an unbiased view of cellular state transition during differentiation from HSCs. Our findings independently support an alternative hematopoietic hierarchy first proposed by Jacobsen and colleagues (Adolfsson et al., 2005) and provide a molecular model for early MegE lineage separation. The sensitivity of the assay allowed detection of coordinated MegE transcriptional priming within the most highly enriched HSCs. We then extended the analysis to a robust, clinically relevant model of AML. At the level of single cells, we showed that leukemia cells are intrinsically distinct from any of the wild-type hematopoietic lineages. Interestingly, the most significant heterogeneity within the leukemia corresponded to two independent, disease-initiating clones. Moreover, we found that Ezh2 is overexpressed in the more highly proliferative leukemia cells and uncovered a link with cell-cycle progression.

No two cells are identical, a concept evident in our data set. Even the most closely related cells, which correspond to two erythroid progenitor cells from our complete data-clustering heatmap (Figure 2B), exhibit differences in gene expression patterns. Variation detected in single-cell gene expression data may reflect biological and/or technical noise or may correspond to function. Distinguishing these two types of variation is very important. Here, we assayed many single-cell samples and searched for correlated gene clusters rather than variation in expression of individual genes. Such correlated gene expression behavior is more likely to represent genetic network function rather than biological noise. By this approach, we correlated MegE priming in the HSCs and correlated Ezh2 dynamics with cell cycle, which were then both functionally validated. Single-cell gene expression data is extremely valuable for extracting such correlated gene expression clusters, because single cells represent the fundamental unit of genetic network regulation.

The mammalian epigenetic landscape contains numerous transitional cellular states within lineage differentiation pathways. Comprehensive mapping of this landscape requires single-cell gene expression analysis in order to represent all possible states. Such an assay needs to be both quantitative and thorough, so that the data are experimentally robust and reflects all major cell types. We suggest that the strategy described here satisfies these requisites. In order to validate biological differences in cell populations, we have relied extensively on study of cell surface markers, as available antibodies can then be used for prospective cell isolation. The approach is readily applicable to other biological contexts. Its use should facilitate identification of new surface markers for functional assessment of stem and progenitor cells and the construction of cellular hierarchies in other organ systems. The strategy is suitable for deconvoluting cellular heterogeneity within different types of cancers. Further accumulation of data sets from diverse contexts should eventually allow for the mapping of all nonredundant cellular states on the mammalian differentiation hierarchy.

### Experimental Procedures

**Multiplexed Primer Design for Single-Cell Analysis**

Gene symbol list for commonly used surface markers is summarized from two resources: a comprehensive mouse cell surface antigens review paper (Lai et al., 1999) and the eBioscience website mouse cellular antigen charts (http://www.ebioscience.com/resources/mouse-cc-chart.html). Gene symbols are then converted to mRNA refseq ID by DAVID tools (http://david.abcc.ncifcrf.gov/). mRNA sequences for each gene are retrieved from University of California Santa Cruz table browser; only common regions are used for genes with different isoforms. Multiplexed, gene-specific primers are designed using a Primer3-based (http://primer3.wi.mit.edu/) algorithm to ensure that each primer within the designed pool has a maximum complimentary sequence of 7 bp to all the other primers. All primers (Table S7) are synthesized and provided by Boston Open Labs (http://boltresearch.com/).

**Figure 7. Single-Cell Analysis of the AML Cellular Hierarchy**

(A) Single-cell sorting strategy for different leukemia compartments in the MLL-AF9 AML mouse model.

(B) A heatmap showing hierarchical clustering of gene expression signatures from 390 single cells from wild-type myeloid cells (Lin-1I7+Klt+Sca1+CD34+CD16/CD32+ or Lin- bone marrow) and MLL-AF9 primary leukemia cells (Lin-1I7+Klt+Sca1+CD34+CD16/CD32+ or Lin+ bone marrow). Each row corresponds to a specific gene; each column corresponds to a particular single cell. White boxes highlight strongly correlated gene and cell clusters. Color scale is as described in Figure 1B.

(C) SPADE analysis of the wild-type myeloid hierarchy and leukemia hierarchy using the high-dimensional, single-cell data. Overlaid expression pattern of different gene clusters helped to define distinct cell lineages in the MLL-AF9 leukemia system.

(D) Gene expression clustering heatmap of gene expression from the main leukemia cell clusters and all the hematopoietic cell clusters reveals distinct leukemia-specific expression.

(E) Dissection of heterogeneity in the LGMP (Lin-1I7+Klt+Sca1+CD34+CD16/CD32+) according to described method in Figure 3.

(F) Survival of secondary recipient mice receiving 800 CD24+ LGMP or CD24+ LGMP cells.

(G) Reconstitution of two leukemia lineages in the secondary recipient bone marrow.

See also Figure S7 and Tables S6 and S7.
FACS and Single-Cell Collection
Seven- to twelve-week-old C57B/6 mice or Actb-GFP C57B/6 transgenic mice were used throughout this study (except for fetal liver and aged mice). Bone marrow cells were isolated by crushing iliac crest bones, femur, and tibiae in PBS containing 5% fetal calf serum and 2 mM EDTA. After red blood cell lysis, the remaining cells were stained with monoclonal antibodies, analyzed, and sorted on the BD FACSAria II (BD Biosciences). Individual cells were sorted directly into 96 well PCR plates loaded with PCR buffer under single-cell mode. Monoclonal antibodies and conjugations used in this study are found in Table S7. All data were analyzed with FlowJo (Tree Star).

One-Tube Single-Cell Sequence-Specific Preampification
Individual primer sets (total of 300) were pooled to a final concentration of 0.1 mM for each primer. Individual cells were sorted directly into 96 well PCR plates loaded with 5 nl RT-PCR master mix (2.5 nl CellsDirect reaction mix, Invitrogen; 0.5 nl primer pool; 0.1 nl RT/taq enzyme, Invitrogen; 1.9 nl nucleic acid water) in each well. Sorted plates were immediately frozen on dry ice. After brief centrifugation at 4°C, the plates were immediately placed on PCR machine. Cell lysates and sequence-specific reverse transcription were performed at 50°C for 60 min. Then, reverse transcriptase inactivation and Taq polymerase activation was achieved by heating to 95°C for 3 min. Subsequently, in the same tube, cDNA went through 20 cycles of sequence-specific amplification by denaturing at 95°C for 15 s, annealing, and elongation at 60°C for 15 min. After preampification, PCR plates were stored at -80°C to avoid evaporation.

High-Throughput Microfluidic Real-Time PCR
Preamplified products were diluted 5-fold prior to analysis. Amplified single-cell samples were analyzed with Universal PCR Master Mix (Applied Bio-Systems), EvaGreen Binding Dye (Bioc-line), and individual qPCR primers using 96.96 Dynamic Arrays on a Biomek System (Fluidigm). Three Dynamic Arrays loaded with different primer sets were used for each sample plate. Threshold crossing (Ct) values were calculated using the Biomek Real-Time PCR Analysis software (Fluidigm).

Single-Cell NanoString
Reporter probes are designed and synthesized by NanoString R&D team. Target sequences are amplified from single cells using one-tube single-cell sequence-specific preamplification as described before. Twenty-five percent of the amplified cDNA are subject to gene expression quantification using the GENZ2 Digital Analyzer. Raw counts are compiled, normalized, and analyzed using nSolver. The data are then subtracted with the background signal and transformed to Log2 scale before analysis.

Computational Processing of Single-Cell Data
A background Ct of 28 was used for all real-time signals. Samples with low Actb expression level (Ct higher than 18) are outliers of normal distribution and are excluded from the analysis. These samples had low or no expression for all the other genes, suggesting that they correspond to empty wells or bad single-cell samples. Hierarchical clustering was done with MultiExperiment Viewer program. For all hierarchical clustering heatmaps, the rainbow scheme color map was used and adjacent genes are set from 1 to 14, corresponding to Log2 gene expression above background of 28. GEDI plots were generated using the gene expression dynamics inspector. Each pixel on the 10.10 GEDI map corresponds to a particular micogene cluster generated by the software. Violin plot, box plot, and correlation heatmap were generated with R software. SPADE analysis was performed with Matlab. Lineage specific gene lists for 180 intracellular regulator assay set and for Figure S5D are generated from the Imgen website analysis tool. CHIP-seq peak visualization was done with Integrative Genomics Viewer program. The genetic networks in Figures S5A and S3A were constructed using Geyescpe 3 software.

SUPPLEMENTAL INFORMATION
Supplemental Information for this article includes Supplemental Experimental Procedures, seven figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.stem.2013.07.017.

ACKNOWLEDGMENTS
We thank H. Skalsky from Whitehead Institute for extensive help on the multiplexed primer design; Y. Fujisawa, E. Baona, O. Yilmaz, M. Nguyen, X. Han, V. Berga, D. Linn, and J. Buchman for help with different parts of the sample preparations; and H. Huang, Z. Li, D. Scadden, X. Xie, and H. Heck for insightful discussions on the project. This work was supported by funding from the National Institutes of Health and the Harvard Stem Cell Institute (S.H.O.). S.H.O. is an investigator of the Howard Hughes Medical Institute.

Received: May 24, 2013
Revised: July 4, 2013
Accepted: July 22, 2013
Published: September 12, 2013

REFERENCES


