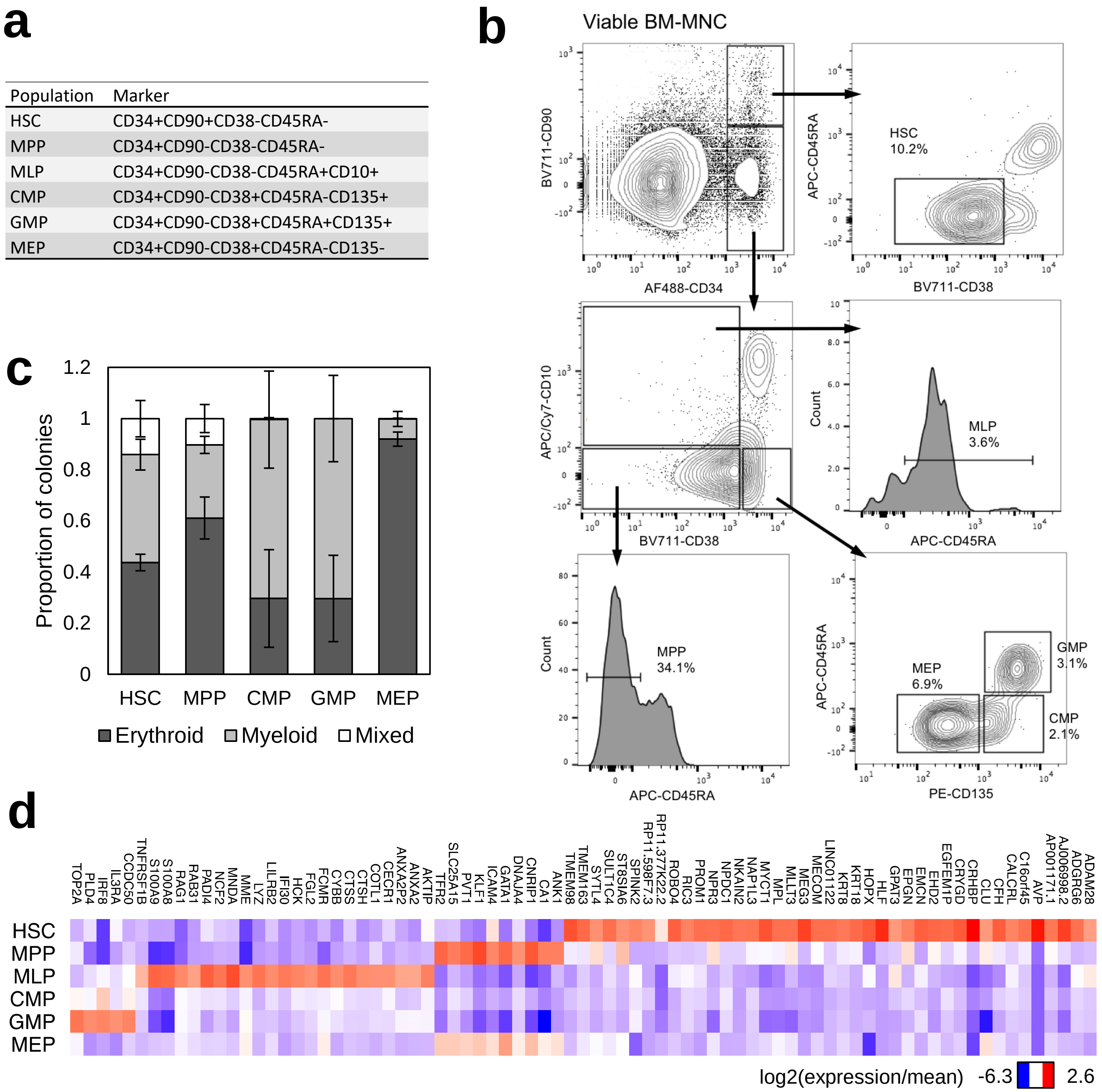
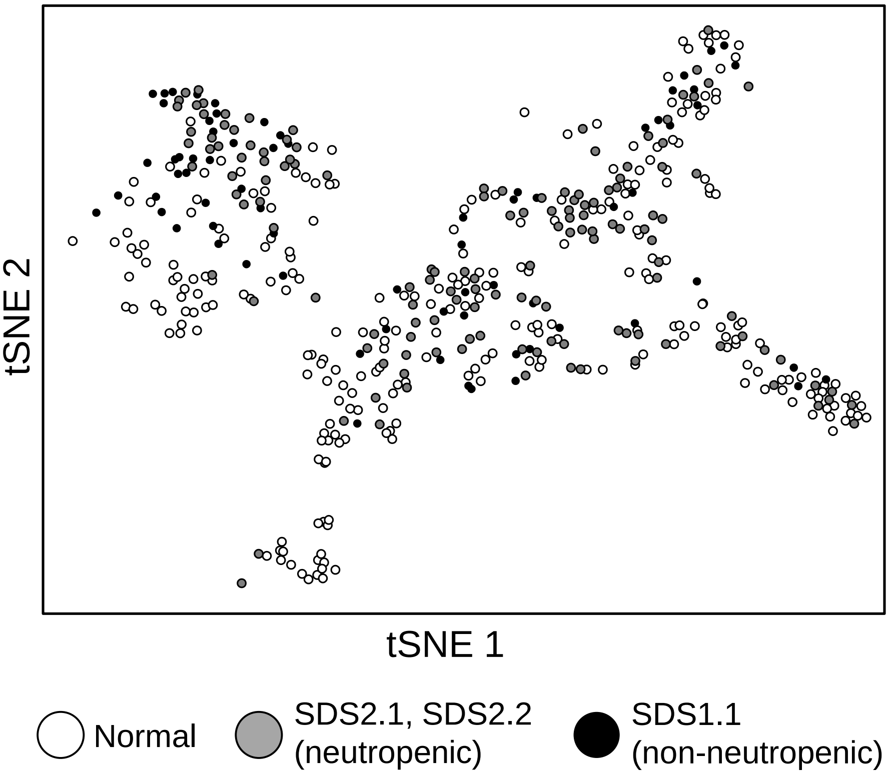
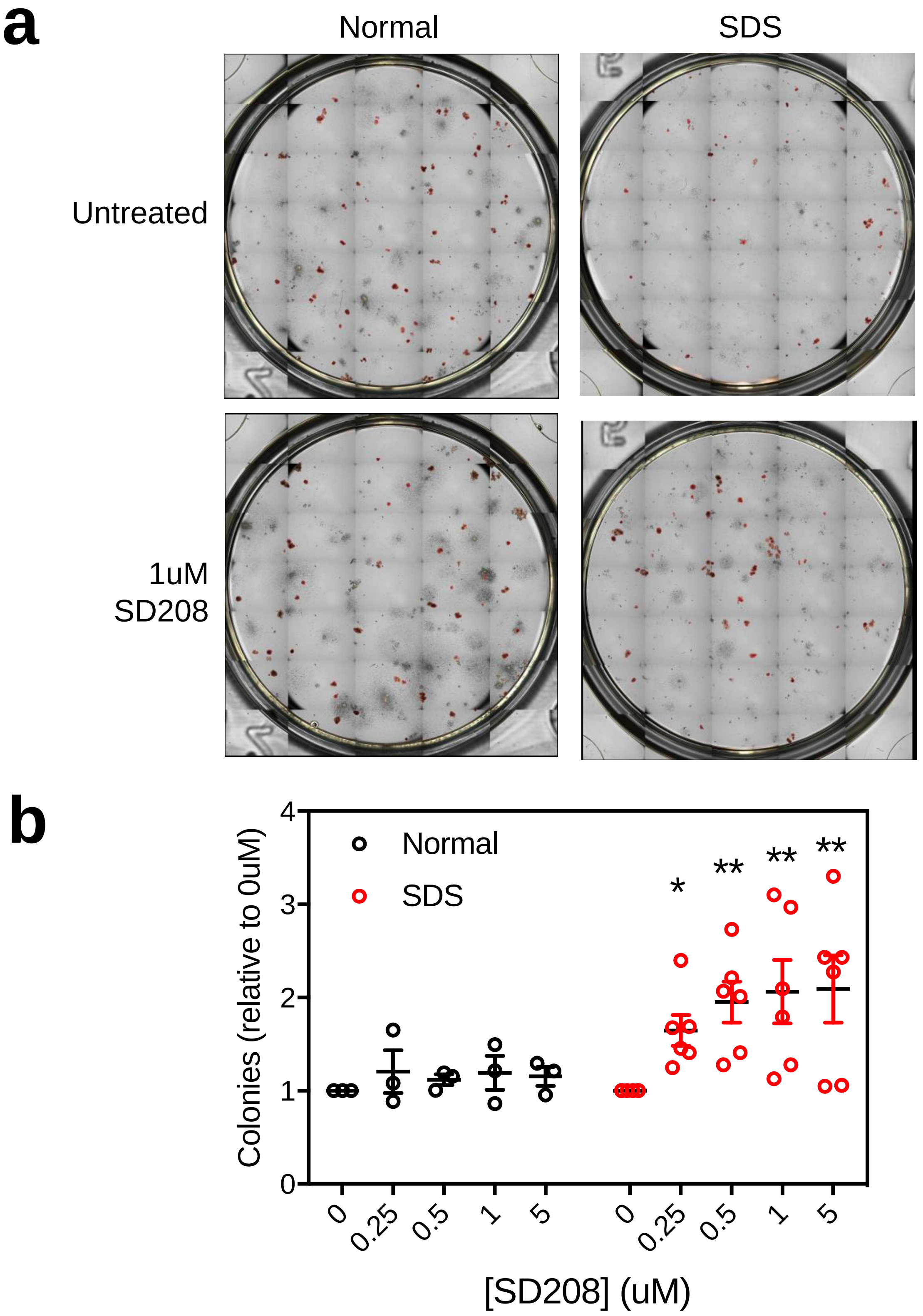


Population	Marker
HSC	CD34+CD90+CD38-CD45RA-
MPP	CD34+CD90-CD38-CD45RA-
MLP	CD34+CD90-CD38-CD45RA+CD10+
CMP	CD34+CD90-CD38+CD45RA-CD135+
GMP	CD34+CD90-CD38+CD45RA+CD135+
MEP	CD34+CD90-CD38+CD45RA-CD135-

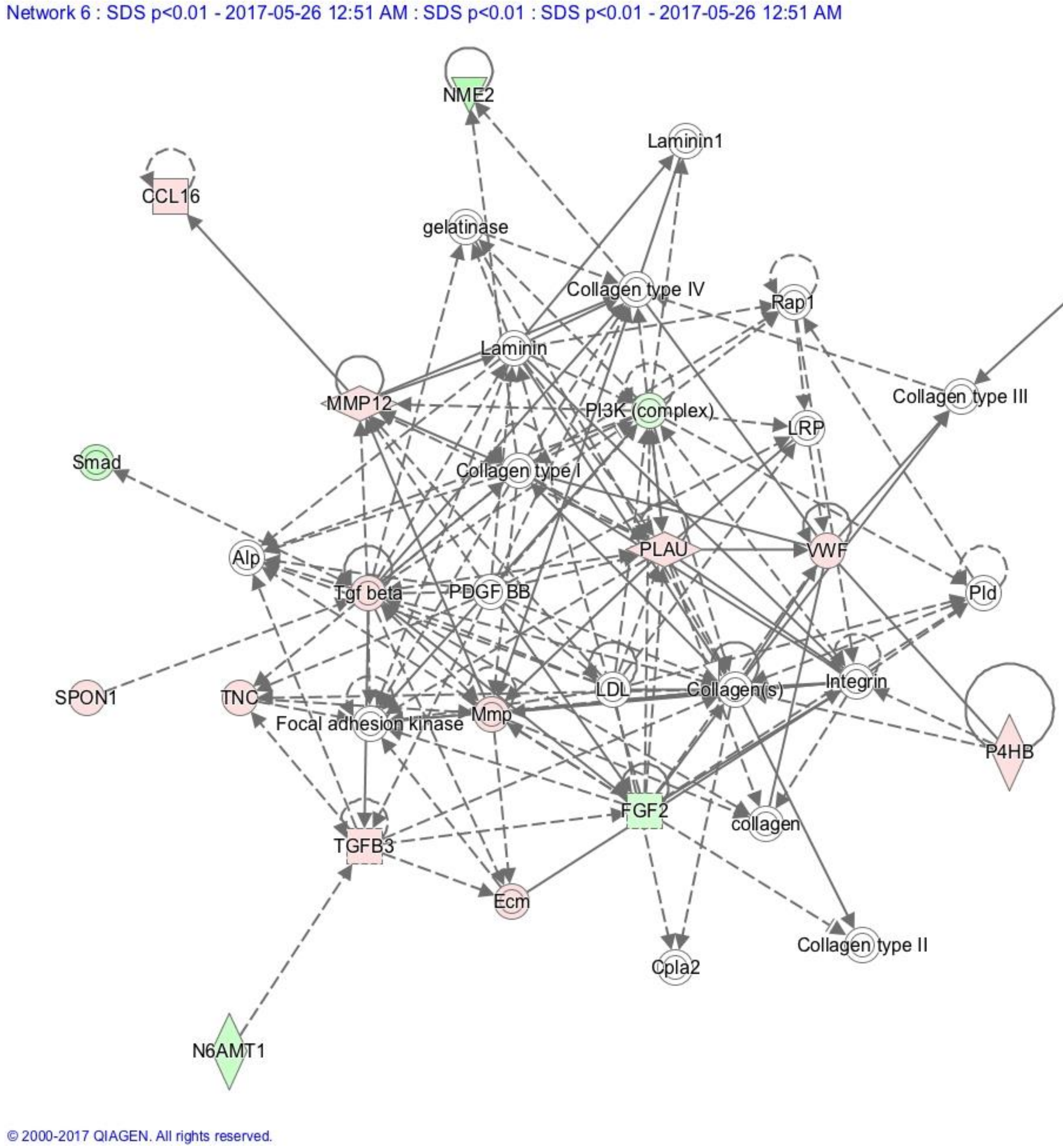
Supplementary Figure 1. Derivation of lineage commitment gene expression signature. (a) Immunophenotypes and (b) representative gating scheme used to purify CD34$^+$ subsets in human BM. Percentages = % of CD34$^+$ cells. (c) Hematopoietic colony forming assays demonstrating enrichment of mixed colonies from HSC and MPP gates, myeloid colonies from CMP and GMP gates, and erythroid colonies from the MEP gate. (d) Heatmap showing a 79 gene signature derived from sequencing 100 cells purified from each gate. Expression values reflect the average expression of each gene across two biological and two technical replicates per subset. High expression of erythroid genes such as GATA1 and KLF1 in the MPP subset is likely due to the recently reported enrichment of MEP in the CD34$^+$CD38^{mid}CD45RA$^-$CD135$^-$ population[1], which was gated as MPP under our sorting strategy adapted from Laurenti *et al.*[2]. Immunophenotypic MPPs did not cluster with GATA1-expressing MEP, as shown in Fig. 1D.



Supplementary Figure 2. SDS GMP deficiency is present in the absence of symptomatic neutropenia. tSNE plot of hematopoietic lineage commitment was derived from an empirically-derived gene expression signature, colored based on SDS diagnosis and active neutropenia (absolute neutrophil count < 1500/ul).



Supplementary Figure 3. a) Representative, full-well images from methylcellulose colony forming assays performed on primary bone marrow mononuclear cells from SDS patients and a normal donor in the presence or absence of 1uM SD208. b) Number of colonies formed by normal donor and SDS patient BM-derived mononuclear cells with increasing concentrations of SD208, normalized to the 0uM treatment. Significance was determined by two-way ANOVA, with Holm-Sidak's multiple comparisons test. *p<0.05, **p<0.01.



Supplementary Figure 4. Dysregulated protein network including TGFB3 and associated factors in SDS patient plasma. Significant networks were assembled from differentially expressed proteins using Ingenuity Pathway Analysis.

SUPPLEMENTARY DATA REFERENCES

1. Sanada, C., Xavier-Ferrucio, J., Lu, Y.C., Min, E., Zhang, P.X., *et al*. Adult human megakaryocyte-erythroid progenitors are in the CD34+CD38mid fraction*.* Blood, 2016. **128**(7): p. 923-33.

2. Laurenti, E., Doulatov, S., Zandi, S., Plumb, I., Chen, J.*, et al.* The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nature immunology* **14**, 756-763 (2013).

METHODS

Sample processing. For scRNA-seq of all SDS samples, and normal donors N1 and N2: 7-20 ml of fresh BM were diluted to 35ml in MACS buffer (PBS/2mM EDTA/0.5% BSA), layered onto 15ml Ficoll-paque (GE Healthcare, Uppsala, Sweden), and spun for 30 min at 1400 rpm and 20°C with no brakes. Mononuclear cells were collected from the interface, washed once, pelleted for 5 min at 1200 rpm and 20°C, and resuspended at 40 ul per 10^7 cells in MACS buffer + 1 ul/ml RNaseOUT (Thermo Fisher Scientific, Waltham, MA, USA). CD34+ cells were positively selected on an AutoMACS instrument using the Indirect CD34 MicroBead Kit (Miltenyi, Bergisch Gladbach, Germany), and singulated on the C1 Instrument (Fluidigm, San Francisco, CA, USA). cDNA libraries were prepared using the SMARTer Ultra Low RNA Kit (Clontech, Mountain View, CA, USA). For samples N3 and N4, protocol conditions were modified to ascertain immunophenotypes from single cells, and in accordance with the newest available methods. For these samples: red blood cells were lysed with ammonium chloride (Stem Cell Technologies, Vancouver, CA). Mononuclear cells were pelleted for 5 min at 1200 rpm and 20°C, washed twice, and resuspended in PBS + 1 ul/ml RNaseOUT. Cells were stained as described below. Single CD34+ cells were sorted into 5ul TCL buffer (Qiagen, Hilden, Germany) in 96 well plates using a FACS Aria II instrument (BD, Franklin Lakes, NJ, USA) on index mode. Two technical replicates of 100 cells from each gated CD34+ subset – HSC, MPP, MLP, CMP, GMP, MEP – were sorted into 5 ul TCL buffer in separate 96 well plates. cDNA libraries were prepared using the SMART-Seq v4 Ultra Low RNA Kit (Clontech). Libraries from all samples were sequenced on a HiSeq 2500 Instrument (Illumina, San Diego, CA) to a read depth of ~3 M paired-end, 25 bp reads per single cell, or ~12 M paired-end, 25 bp reads per 100 cells.

Antibodies and staining. Cells were stained at a density of $1x10^6$ per 100 ul in PBS + 1 ul/ml RNaseOUT because staining buffers contain proteins that can inhibit SMARTer-seq (Clontech) cDNA synthesis reactions. The staining panel was adapted from an analysis of human cord blood progenitors(1). in accordance with the parameters of our flow cytometer. Antibodies used were: brilliant violet 421-anti-CD90 (BD 562556, 1:20), alexa fluor 488-anti-CD34 (Biolegend, San Diego, CA 343518, 1:20), brilliant violet 711-anti-CD38 (BD 563965, 1:20), allophycocyanin-anti-CD45RA (BD 550855, 1:5), phycoerythrin-anti-CD135 (BD 558996, 1:5), and allophycocyanin-cyanine 7-anti-CD10 (Biolegend 312212, 1:20). Live/dead staining was performed immediately prior to sorting using Zombie Aqua Fixable Viability Dye (Biolegend). Cells were sorted on a FACSAria II instrument (BD), and data analysis was performed in FlowJo v10.0.8.

Data processing and availability. Paired-end reads were mapped to the hg38 human transcriptome (Gencode v24) using STAR v2.4.2a(2). Aligned reads are available through dbGaP (phs001845.v1.p1). Gene expression levels were quantified as transcript-per-million (TPM) in RSEM(3). Cells with at least 1000 expressed genes (defined by TPM>1) and genes expressed in at least 50 single cells were kept. This resulted in 11094 genes in 583 single cells. The same set of 11094 genes was analyzed to derive lineage signature genes from 100 cell libraries made from FACS-purified CD34+ subsets.

Gene selection based on bulk expression data. We used the Gini index(4) to identify cell type-specific genes from HSC, MPP, CLP, CMP, MEP, and GMP 100 cell libraries. We first calculated maximum TPM value of each gene, and genes with maximum value lower than the 20-quantile of all maximum values were filtered out because those genes could have high Gini index due to their low expression. We then identified the top 500 high Gini index genes for each of the biological (*n*=2) and technical (*n*=2) replicates for each cell type. The cell type specific gene signatures were chosen as the intersection of high Gini genes across all replicates for each cell type.

tSNE analysis. We divided TPM values by 10 to better reflect the complexity of single cell libraries which is estimated to be ~100,000 transcripts(5). The data were log2 transformed (log2(TPM/10

+1)). The expression of the 79 genes identified by bulk data across the 583 single cells was used for Principal Component Analysis (PCA) in the Seurat Package in R(6). Using a jackstraw approach implemented in the Seurat package with num.replicate = 200 and each time randomly permuting three genes, the top four principal components (PCs) were identified as significant (*p*-value < $1x10^{-4}$). To aid visualization, these top four PCs, were subject to t-distributed Stochastic Neighbor Embedding (t-SNE)(7) analysis in Seurat with 2000 iterations.

Clustering analysis. The tSNE coordinates were used for partitioning around medoids (PAM), a more robust version of *k*-means clustering implemented in the "cluster" package in R with default parameters (https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/pam.html). To determine the optimal *k*, we assessed the average Silhouette value(8) for each clustering result (from *k*=2 to *k*=10) and selected *k* =5, which gave the largest mean Silhouette value.

Differential gene expression and pathway analysis. Differential gene expression analysis was performed on SDS versus normal cells in each cluster (and in all clusters combined) using the MAST package in R(9) *p*-values were adjusted for multiple testing using the "p.adjust" function in R with "fdr" method(10) We focused on genes with an FDR adjusted *p*-value < 0.05 and |log2(fold change)| >1 in at least one cluster. Enriched pathways and functions were determined in Ingenuity Pathway Analysis (Qiagen) using the 11094 detected genes as the reference gene set. Split violin plots were generated using the ''vioplot'' package and "vioplot2" function in R.

Immunofluorescent staining and imaging. Primary BM-derived mononuclear cells were cultured for 30-32h in StemSpan SFEM II (Stem Cell Technologies) supplemented with 100 ng/mL of SCF, TPO, Flt3L and 20 ng/mL of IL-3 (PreproTech, Rocky Hill, NJ). CD34+ cells were sorted using CD34 Microbeads (Millitenyi) according to manufacturer's protocol, and allowed to recover in culture medium for 14-16h, plus an additional 2h in the presence of 0.6μg/ml AVID200 for relevant samples. 25,000-50,000 cells were spun onto coverslips (ES0117580, Azer Scientific, Morgantown, PA) using a cytospin instrument (Thermo Shandon) at 380rpm for 5min; fixed with 4% PFA in 1X PBS for 10min at room temperature (RT); washed 2X with 1X PBS; permeabilized with 0.3% TritonX in 1X PBS solution for 10min at RT; washed 2X with 1X PBS; blocked in 10% FBS, 0.1% NP40 in 1X PBS for 1h at RT; incubated with 1:250 anti-p-smad2 (Invitrogen, 44-244G) in blocking solution for 14-16h at 4°C; washed 3X with 0.1% NP40 in 1X PBS at RT for 10min; incubated with 1:1,000 diluted anti-rabbit IgG-Alexa488 antibody (Invitrogen, A21206) in blocking solution for 1h at RT; and washed 3X with 0.1% NP40 in 1X PBS at RT for 10min. Stained coverslips were mounted on glass slides with VectaShield with DAPI (H-1200, Vector Laboratories, Burlingame, CA) diluted 1:1 in VectaShield without DAPI (H-1000). Slides were imaged on a LeicaSP5 confocal microscope with constant laser power (30% for DAPI, 70% for Alexa488) and identical resolution, offset, and gain settings for all slides. Z stack images were captured with 40-80μm step range, and the plane with the best nuclear representation was analyzed using Fiji software. Background was calculated using four randomly selected empty regions for each image. Mean signal intensity for p-SMAD2 (Alexa Fluor-488) was calculated within each nucleus, and background signal was subtracted.

Colony formation assays. Primary BM-derived mononuclear cultured for 24h in StemSpan SFEM II (Stem Cell Technologies) supplemented with 100 ng/mL of SCF, TPO, Flt3L and 20 ng/mL of IL-3 (PreproTech, Rocky Hill, NJ). Cells were resuspended at 10,000 cells/mL for control and 20,000 cells/mL for SDS in the presence or absence of 0, 0.25, 0.5, 1, or 5 µM SD208 (Tocris, Bristol, UK), and incubated for 1hr at 37˚C/5% CO_2. 200 µL of cell suspension was mixed with 3 mL of Methocult H4434 (Stem Cell Technologies), and 1 mL was plated in triplicate in a SmartDish 6-well plate (Stem Cell Technologies). After 14 days of growth at 37˚C/5% CO_2, colonies were manually scored by two independent, blinded investigators using standard criteria(11).

SOMAscan proteomic analysis. SOMAscan (SomaLogic, Boulder, CO) was performed on 50 ul of EDTA-plasma from six patients and six normal controls at the BIDMC Genomics, Proteomics, Bioinformatics and Systems Biology Center. Samples were prepared and run using the SOMAscan Assay Kit for Human Plasma, 1.3k (cat. # 900-00011), according to the manufacturer's protocol. Five pooled controls and one no-protein buffer control provided in the kit were run in parallel with the samples. Median normalization and calibration of the data was performed according to the standard quality control protocols at SomaLogic. All samples passed the established quality control criteria. Proteins with *p*-values<0.01 were analyzed. Benjamini-Hochberg adjusted *p*-values are reported in Extended Data Table 4.

Statistics. In figure 2a, statistical significance was determined by the chi-squared test; the frequency of cells in each cluster was compared between SDS and normal. In Figure 2b, 3c, 4d, and Extended Data Figure 3b, statistical significance was determined by two-way ANOVA with Holm-Sidak's multiple correction test in GraphPad Prism 7. In Figure 2b, the frequency of cells was compared between SDS and normal cells within each cluster. In Figure 3c, log2 expression was compared between SDS and normal cells within each cluster. In Figure 4d and Supplementary Figure 3b, relative colony number was compared between each drug dose and the 0uM treatment. In Figure 4b and 4c, statistical significance was determined by one-way ANOVA with Holm-Sidak's multiple correction test in GraphPad Prism 7; SDS samples were compared to normal samples that were stained and imaged concurrently.

Study approval. Subjects provided written, informed consent for protocols approved by the institutional review board of Boston Children's Hospital (Boston, MA) and Dana-Farber Cancer Institute (Boston, MA), in accordance with the Declaration of Helsinki's Ethical Principles of

Medical Research Involving Human Subjects. All subjects provided informed consent prior to their participation in the study.

METHODS REFERENCES

1. Laurenti E, et al. The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nat Immunol.* 2013;14(7):756-63.
2. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15-21.
3. Li B, and Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics.* 2011;12:323.
4. Jiang L, Chen H, Pinello L, and Yuan GC. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol.* 2016;17(1):144.
5. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, 2nd, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science.* 2016;352(6282):189-96.
6. Satija R, Farrell JA, Gennert D, Schier AF, and Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol.* 2015;33(5):495-502.
7. Maaten LJPVD, and Hinton GE. *Visualizing High-Dimensional Data using t-SNE.* 2008.
8. Rousseeuw P. *Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65.* 1987.

9. Finak G, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 2015;16:278.

10. Benjamini Y, and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological).* 1995;57(1):289-300.

11. Eaves C., Lambie K. Atlas of Human Hematopoietic Colonies: An Introduction to the Recognition of Colonies Produced by Human Hematopoietic Progenitor Cells Cultured in Methylcellulose Media. *StemCell Technologies.* 1995.