SUPPLEMENTARY APPENDIX

Long noncoding RNAs of single hematopoietic stem and progenitor cells in healthy and dysplastic human bone marrow

Zhijie Wu,[1]* Shouguo Gao,[1]* Xin Zhao,[1] Jinguo Chen,[2] Keyvan Keyvanfar,[1] Xingmin Feng,[1] Sachiko Kajigaya[1] and Neal S. Young[1]

[1]Hematology Branch, National Heart, Lung, and Blood Institute, National Institutes of Health and [2]Trans-NIH Center for Human Immunology, Autoimmunity, and Inflammation, National Institutes of Health, Bethesda, MD, USA

**ZW and SG contributed equally to this work.*

 doi:10.3324/haematol.2018.208926

Received: October 12, 2018.

Accepted: November 22, 2018.

Pre-published: December 13, 2018.

Correspondence: *ZHIJIE WU* zhijie.wu@nih.gov

Long noncoding RNAs of single hematopoietic stem and progenitor cells in healthy and dysplastic human bone marrow

Short title: single cell lncRNAs in hematopoiesis

Zhijie Wu,[1,*] Shouguo Gao,[1,*] Xin Zhao,[1] Jinguo Chen,[2] Keyvan Keyvanfar,[1] Xingmin Feng,[1] Sachiko Kajigaya[1] and Neal S. Young[1]

*Z.W. and S.G. contributed equally to this work.

Supplementary Methods

scRNA-seq bioinformatics assembly

High-quality filtered cells[1] were used to define lncRNAs with a high-confident model.[2-5] For each subject (healthy donors or MDS patients), reads were pooled and mapped to the human hg19 reference genome using aligner TopHat2, which was supplied with GENCODE GTF file as a gene model reference. We used a previously published two-step mapping strategy[6] to obtain and combine splice junctions from all subjects. In the first round, novel splice junctions were identified in all subjects and combined to obtain a set of novel junctions. In the second round, mapping was performed, with the combined junctions using option "-j" (--raw-juncs) and "--no-novel-juncs". The sorted BAM files obtained from the second round were input into Cufflinks to assemble transcripts. New transcripts were assembled for each subject individually using Cufflinks with discovery mode ("-g" option) and with parameters --min-isoform-fraction =0.0, --minfrags-per-transfrag =1 and --upper-quartile-norm. Finally, Cuffmerge[7] was used to combine the transcriptomes of all subjects. These transcripts included lncRNAs, protein-coding mRNAs, other noncoding RNAs, and even DNA contamination, whereby further filtering was needed.

High-Confidence filtering pipeline to identify novel lncRNAs

Bedtools, bedops, and perl/R scripts were utilized to filter out assembled transcripts, based on published lncRNA identification protocols:[6,8]

1. Transcripts shorter than 200 nt in length and with only one exon (a high chance to be DNA contamination or pre-splicing mRNAs) were removed.

2. We used BEDTools to intersect our *de novo* transcript models with transcript models from the RefSeq, UCSC and Ensembl databases, and discarded all transcripts overlapping at least 1 bp with any known mRNA exons.

3. We removed the transcripts with low coverage below 3 reads per base.

4. Filtering with tools of assessing protein coding potential by BlastX/hmmer/getORF. BlastX: Blast tools v6 were used to translate and align assembled transcripts against a subset of the non-redundant (NR) protein database to peptide homology with command: blastx -query <input.fasta> -strand plus -db db/nr -evalue 0.000001 -outfmt 6 -gilist <gilist> - max_target_seqs. Only hits above a significance threshold of Expect value < 10^{-6} were considered significant. HMMER/Pfam: PfamScan (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/PfamScan.tar.gz, version 7) was used to identify transcripts containing protein-coding domains annotated in Pfam database. Both PfamA and PfamB-quality families were provided as input to HMMER3 (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/, files Pfam-A.hmm, Pfam- A.hmm.dat, Pfam-B.hmm and Pfam-B.hmm.dat). getORF: GetORF9 in EMBOSS was used in Galaxy with the following parameters [Code: Standard; Minimum nucleotide ORF size: 300] to identify any open reading frames longer than 300 nt starting with methionine.

5. CPAT: we used an online tool, CPAT,[9] to exclude transcripts with characteristic coding features independent of their conservation. The CPAT score of a given transcript indicates its protein-coding

potential through logistic regression of four sequence features of known protein-coding transcripts and noncoding RNA. As a control, we applied CPAT on annotated proteins and lncRNAs in GENCODE (Figure 1C).

6. Filtering known lncRNAs and pseudogenes. We removed transcripts which overlapped with pseudogenes (combining Yale Pseudo60 genes, Vega pseudogenes and ENSEMBL) and known lncRNAs in GENCODE.[10,11]

Integration with CAGE to get more reliable lncRNAs and cell type specificity confirmation

When assembling lncRNAs with gene expression data alone, only high expressed transcripts were included in ours and other's pipelines[2-6,8] to guarantee the quality. Recently, Hon *et al*. utilized the advantage of CAGE datasets, which can accurately capture 5' ends of lncRNAs, to refine existed transcript models and build an atlas of human lncRNAs with accurate 5' ends.[12]

Supported by CAGE, we lowered the cutoff of expression levels for transcripts filtering with a consideration that transcripts with medium expression levels that were supported by CAGE datasets were expressed in human $CD34^+$ cells. We used the same strategies through considering CAGE cluster expression levels, exon number and exon length of transcripts, and the distance between CAGE cluster to transcript, and identified another 281 lncRNAs and deposited into GEOs.

Further, we utilized gene expression of large numbers of cell types in CAGE study to validate the cell type specificity of lncRNAs. We sought to compare if those lineage-specific lncRNAs in the current study also showed preferential expression in corresponding samples in CAGE data. Though no perfectly matched cell types could be found in a CAGE dataset, we chose the closest cell types in CAGE (ie. erythrocytes progenitors, defined as MEP-like) to check the expression of both mRNAs and lncRNAs. Expression correlation was calculated as: (expression of mRNAs or lncRNAs in

MEP)/(expression of mRNAs or lncRNAs in non-MEP) in our study, with (expression of mRNAs or lncRNAs in MEP-like)/(expression of mRNAs or lncRNAs in all other cell types) in CAGE data.

Evaluation of cell and cell type specificity

We plotted variance against mean gene expression across cells or cell types for variation. To quantify whether lncRNAs exhibited higher cell type-specific expression patterns than did protein-coding genes, we calculated a tissue specificity score for each transcript using an entropy-based metric that relies on Jensen-Shannon (JS) divergence.[6] The specificity score is defined as 1-(JSdist(p,q)) where p is the density of expression [probability vector of log10(TPM+ 1)] of a given gene across all conditions, and q is the unit vector for that cell type (ie. perfect expression in that particular cell type). JSdist is a function that used to calculate pairwise Jensen-Shannon distances between columns in R package "cummeRbund". This specificity metric (ranging from 0 to 1) quantifies the similarity between transcripts' expression patterns across cell types. JS specific score = 1 means a transcript is expressed exclusively in that condition.[13]

Quantitative RT-PCR and bioinformatics analysis

FACS-sorted bulk samples from three new healthy donors were used for quantitative RT-PCR validation. Combination of surface markers for each cell population[14] is illustrated in *Online Supplementary Figure S2*. Total RNA was isolated using the Agencourt RNAdvance Cell v2 kit (Beckman Coulter), and assessed by using a Nanodrop spectrophotometer (NanoDrop Technologies). Complementary DNA was synthesized by using the Fluidigm Reverse Transcription Master Mix (Fluidigm). cDNA from this new set of bulk samples and whole transcriptome amplification (WTA) products from 391 single $CD34^+$ cells were subjected to gene expression analysis. A list of lineage-

specific lncRNAs (n = 39) and mRNAs (n = 14) and two housekeeping genes (*Online Supplementary Table S6*) were preamplified and analyzed using quantitative RT-PCR following manufacturer's protocol (Fluidigm). Gene expression preamplification was performed with Fluidigm Preamp Master Mix (Fluidigm) and TaqMan Assays (Thermofisher). Subsequently, gene expression analysis was performed in a 96.96 quantitative PCR Dynamic Array on the Fluidigm Biomark instrument using Fast TaqMan Assays (Thermofisher) as described previously.[15]

Single-cell and bulk sample gene expression data were firstly analyzed with the Fluidigm Data Collection software. Expression values over the cutoff of the machine (CT value > 27) were set to 28.[16] For assays that were performed in duplicate, the mean of the duplicate was used for subsequent analysis. After filtering, ΔCT values were calculated by cell-wise normalization to the mean expression level of two housekeeping genes (ACTB and GAPDH) through subtracting of their mean CT value. Cell type specificity was calculated as ΔΔCT, which defined as the difference of ΔCT values of one cell type with all the others.

Functional enrichment analysis of lncRNA-encoding gene sets by "guilt-by-association" approach

We adopted a comprehensive "guilt-by-association" strategy comprising three methods to identify associated protein-coding gene sets of lncRNAs (*Online Supplementary Results*).

1) Co-expression: co-expression has been widely used to annotate functions of lncRNAs.[13, 17-20] We downloaded lncRNA-mRNA co-expression relationships from lncPath (LncPath package in R), which was based on 28 scRNA-seq data sets. To check the applicable, we confirmed that their co-expressed pairs also showed higher-co-expression in our dataset. We chose to use theirs, because network build

from the large dataset thus is more powerful to capture functional similarity. All co-expressed mRNAs in lncPath of given lncRNAs were defined as associated gene sets.

2) Considering the cis role refers to an lncRNA acting on neighboring target genes, we searched coding genes 10-kb upstream and downstream of lncRNAs, and identified possible cis target gene sets.[21]

3) We used Linc2Go as a functional annotation resource for lncRNAs.[22] Lnc2GO integrated microRNA-mRNA and microRNA-lncRNA interaction data to generate comprehensive functional annotations for lncRNAs based on common shared miRNAs, as lncRNAs and mRNAs targeted by same miRNAs should be functionally similar.

The resulting gene sets from these approaches were subjected to functional GO enrichment analysis. Statistically significant GO terms in Biological Process domain were identified with the database for annotation by topGO.[23] The false discovery rate (FDR) calculated with the Benjamini-Hochberg method was used to measure the significance level of the GO terms. Pathway gene sets were downloaded from Pathway Commons (https://www.pathwaycommons.org/).[24] Fisher's exact test was used to calculate the statistical significance of this over-representation.[25] Genomatix (https://www.genomatix.de) was used to identify the enriched functions and pathways of gene sets.[26]

Weighted Gene Co-expression Network Analysis (WGCNA)

Coexpression analysis of lncRNAs and protein-coding transcripts was performed with WGCNA package in R.[27] Log transformed TPM values of transcripts were used as input for WGCNA and signed weighted correlation networks were built by calculating correlation coefficients between the top 2000 variable gene pairs. Specifically, adjacency matrices were constructed by setting the soft threshold

value of each correlation matrix at 6 (identified by scale-free examination), and the adjacency matrix was transformed to a topological overlap matrix (TOM). The matrix 1-TOM was used as the input to group genes using the average linkage hierarchical clustering method. Finally, we utilized dynamic tree cut algorithm and the merge cut height and minimum module size were set to 0.2 and 10, respectively. Furthermore, an expression profile of each module was represented by its first principal component (module eigengene), which can explain the most variation of the module expression levels. We constructed cell type-specific mathematical vectors. In those vectors, each element represented one cell, with a value of 1, if the cell belonged to a distinct cell type and 0 otherwise. The Pearson correlation between eigengenes of network modules and the condition vectors was then calculated for association estimation. Our analysis revealed seven modules (Figure 3C and *Online Supplementary Table S4*), which were then characterized for GO term enrichment on the protein-coding genes using topGO.[23] GO SamSim was used to cluster significant enriched GO terms and visualize representative ones.

Reconstruction of a pseudotemporal order of hematopoietic cells

Diffusion map was used to estimate a differentiation lineage trajectory with the Destiny package in R.[28] Pseudotime ordering was implemented with Diffusion Pseudo Time.[29]

Protein-protein-lncRNA network analysis

A network was created through integrating STRING 7.0,[30] co-expression annotated in LncPath, and chromosomal neighborliness. The network analysis was done with Cytoscape.[31]

ChIP-seq and histone modification analysis

GATA1 binding (Encode ID: ENCSR000EFT), H3K27Ac binding (Encode ID: ENCSR000AKP), H3K4me2 binding (ENCODE ID: ENCSR000AKT), H3K27me3 binding (ENCODE ID: ENCSR000AKQ), and H3K79me2 binding (ENCODE ID: ENCSR000APD) data were downloaded and used for analysis of protein-coding and lncRNA-encoding genes in our study.

Supplementary Results

Integration with CAGE to get more lncRNAs and cell type specificity confirmation

In Chung-Chau Hon's study,[12] they performed FAMTOM5 cap analysis of gene expression (CAGE) and generated a comprehensive atlas of 27,919 human lncRNA genes with high-confidence 5' ends and expression profiles across 1,829 samples from the major human primary cell types and tissues. We referred to CAGE data as a reliable source to include those transcripts with medium expression that were filtered by expression levels in our high-confidence filtering pipeline illustrated above.

We referred to CAGE data in three ways. 1) In *de novo* transcriptome reconstruction in the current study, we filtered the data to remove low expressed transcripts. Comparing with CAGE, we lowered our cutoff for gene abundance in lncRNA assembly pipeline and defined the assembled transcripts with medium expression level but could be supported by CAGE. We also included these 281 transcripts (*Online Supplementary File* 2) as defined lncRNAs in human $CD34^+$ cells and uploaded to GEO (GSE99095). 2) By comparing defined lncRNAs in the current study with Chung-Chau Hon's study, we found that around 1/3 assembled lncRNAs overlapped with CAGE annotation (*Online Supplementary Figure S10A*), which was reasonable for several reasons. Chung-Chau Hon's study included 1829 samples from the major human primary cell types and tissues, including a lot more cell types other than human $CD34^+$ cells, and subpopulation of human bone marrow derived $CD34^+$ cells were barely included in CAGE data. Also, as we argued, some cell type-specific lncRNAs that could

be diluted in population samples might be annotated in our single cell data. Those lncRNAs defined in our study overlapped with CAGE data were preferential candidates for functional studies. 3) Next, we utilized CAGE data set to check if those lineage-specific lncRNAs in the current study also showed preferential expression in corresponding samples in CAGE data. In correlation analysis, R values were 0.442 for mRNA expression and 0.243 lncRNA expression (*Online Supplementary Figure S10B*), which suggested rough consistency of preferred expression of lineage-specific lncRNAs in both datasets.

Characterization of lncRNAs defined in human CD34$^+$ hematopoietic cells

lncRNAs present in human CD34$^+$ cells, both previously annotated and putative novel ones, were distributed across the human genome (*Online Supplementary Figure S3A*). On average, lncRNAs were much less frequent than were mRNAs (*Online Supplementary Figure S3B*). Compared with coding transcripts (7.4 exons and 65 K nt, on average), lncRNAs had fewer exons and were shorter in length, while novel lncRNAs (2.4 exons and 1 K nt, on average) tended to even lower exon numbers (*Online Supplementary Figure S3C*) and shorter length (*Online Supplementary Figure S3D*) compared to annotated lncRNAs (3.1 exons and 26 K nt, on average). A PhastCons score was used to evaluate conservation of transcripts.[32]

We found that novel and annotated lncRNAs were similarly less well conserved compared to protein coding transcripts (*Online Supplementary Figure S3E*). We extracted transcript start sites (TSS) of all protein coding genes, and ngs.plot was used to calculate and visualize the locations of annotated and novel lncRNAs relative to TSS. Novel lncRNAs exhibited a similar distance to TSS as did annotated lncRNAs (*Online Supplementary Figure S3F*). We then analyzed expression patterns of defined lncRNAs that were localized within 10 kb from a coding gene. lncRNAs showed higher correlation

with expression of neighboring protein-coding genes than did any random pair of chromosomal neighbors, and also higher than the correlation of paired neighboring coding genes ($P < 0.0001$, Kolomogorv-Smirnov test) (*Online Supplementary Figure S3G*).

"Guilt by association" strategy

lncRNAs are weakly conserved in sequences;[33,34] and data for interaction between lncRNAs and protein-coding genes are mostly lacking. Functional studies by experimental methods are only conducted for limited number of lncRNAs. Additionally, a large number of novel lncRNAs have been annotated with the use of whole transcriptome sequencing, such that computational analysis has been used to impute putative functions of lncRNAs.

One genome wide method to impute functions of lncRNAs is analysis of their association with mRNAs of known functions, or "guilt by association".[18] lncRNAs can influence the expression of neighboring genes in a cis manner. Functions of lncRNAs have been determined by analyzing proximal protein-coding genes.[35-41] Conversely, lncRNAs have also been suggested to primarily affect gene expression in trans,[42-44] and functions of lncRNAs have been inferred from co-expressed protein-coding genes.[34,45-47]

We hypothesized that lncRNAs and protein-coding genes involved in the same biological functional pathways were likely neighbors or coordinately regulated.[32,34,48] To this end, we adopted a comprehensive "guilt-by-association" approach, by taking into account either physically proximate, co-expressed or co-regulated coding genes, in order to impute putative functions of lncRNAs present in human $CD34^+$ cells and differentially expressed lncRNAs in MDS and aneuploid cells. "Guilt-by-association" also was applied to WGCNA analysis, which was based on unsupervised clustering of co-expressed mRNAs and lncRNAs (*Online Supplementary Methods*).

Consistent expression profiles with dataset GSE75478

Identification of cell types of individual cells within the $CD34^+$ cell population has been described previously.[1] Specifically, an HSPC type was assigned to each t-SNE cluster based on significance in overlapping between HSPC- and cluster-specific genes (http://www.jdstemcellresearch.ca/index_files/ResearchData.htm)[49] with one-tailed fisher test.

We then compared both protein-encoding and lncRNA-encoding gene expression between assigned types of cells in our study and Velten's study[50] with RNA sequencing results of sorted single human HSPCs, and there was overall agreement.

For MEPs as an example, we first filtered to generate a list of mRNAs and lncRNAs captured in both datasets, for which there were 11,427 mRNAs and 775 lncRNAs. We then calculated fold change of expression levels of all captured mRNAs and lncRNAs as: LogTPM (gene expression in MEP) - LogTPM (gene expression in non-MEP). Log (fold change of expression level) of both captured mRNAs and lncRNAs showed significant linear correlation in two datasets (*Online Supplementary Figure S7A*). Next, we analyzed preferentially expressed genes in MEPs by one-tailed Pearson comparison ($P < 0.05$) in both datasets, seeking common genes between two datasets (*Online Supplementary Figure S7B*). There were 3,934 common mRNAs and 85 common lncRNAs that were preferentially expressed in MEP in both datasets, and the Log (fold change) of these gene expression levels showed significant linear correlation. Similar results were observed for other lineages such as HSC (*Online Supplementary Figure S8*).

In brief, agreement was observed between our data and online datasets based on different sample sources, cell processing, sorting strategy and cell type defining, composition of cell populations, and different platforms of scRNA-seq, indicating the recurrence of defined lncRNA patterns in our study.

Quantitative RT-PCR analysis of lineage-specific lncRNAs and mRNAs

For validation, we applied multiplex quantitative RT-PCR for a list of lineage-specific lncRNAs and mRNAs (*Online Supplementary Table S6*) to aliquoted WTA of 391 single $CD34^+$ cells from four healthy donors and a new set of flow cytometry sorted bulk samples from three new healthy donors, in order to validate the expression of defined potential novel lncRNAs and lineage-specific expression patterns of lncRNAs. Indeed, all 39 lineage-specific lncRNAs including 20 potential novel ones were detectable with RT-PCR in single cell and bulk samples, while negative controls (no-template control and no-assay control) were all undetectable. By comparing their expression in one cell type with all the others, 35/39 lineage specific lncRNAs showed preferred expression in the corresponding lineages compared with other cells (*Online Supplementary Figure S9A*), which was recurrent in independent sorted bulk samples (*Online Supplementary Figure S9B*). As controls, well-recognized lineage-specific mRNAs (four for HSC, four for MEP, and six for GMP/ProB/ETP) were also analyzed. Similarly, they showed preferred expression in corresponding sorted bulk lineage samples (*Online Supplementary Figure S9*).

lncRNAs exhibit differential expression in myelodysplastic syndromes and aneuploid cells

By comparing gene expression of cells from MDS patients with cells from healthy donors, there were 3,373 mRNAs upregulated and 2,082 downregulated ($P < 0.05$). Pathway analysis revealed that upregulated genes were enriched in the oncogene p53 pathway; cell survival such as TNF alpha/NF-kB,

AKT signaling; and the proteasome complex. Downregulated genes showed enrichment in the immune related pathways (such as IL-2, IL-12, and STAT4 signaling of Th1 development, the BCR signaling pathway) and the DNA repair pathways (such as the Fanconi anemia pathway and BRCA1 dependent UB ligase activity) (*Online Supplementary Table S9*). Similarly, we identified lncRNAs to be differentially expressed between MDS patients and healthy donors ($P < 0.05$): 372 and 590 were upregulated and downregulated in MDS patients, respectively (*Online Supplementary Table S10*).

Supplementary References

1. Zhao X, Gao S, Wu Z, et al. Single-cell RNA-seq reveals a distinct transcriptome signature of aneuploid hematopoietic cells. Blood. 2017;130(25):2762-2773.
2. Alvarez-Dominguez JR, Hu W, Yuan B, et al. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. Blood. 2014;123(4):570-581.
3. Paralkar VR, Mishra T, Luan J, et al. Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. Blood. 2014;123(12):1927-1937.
4. Hu G, Tang Q, Sharma S, et al. Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. Nat Immunol. 2013;14(11):1190-1198.
5. Ranzani V, Rossetti G, Panzeri I, et al. The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. Nat Immunol. 2015;16(3):318-325.
6. Cabili MN, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011;25(18):1915-1927.
7. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562-578.
8. Liu SJ, Nowakowski TJ, Pollen AA, et al. Single-cell analysis of long non-coding RNAs in the developing human neocortex. Genome Biol. 2016;17:67.
9. Wang L, Park HJ, Dasari S, et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013;41(6):e74.
10. Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012;22(9):1775-1789.

11. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760-1774.
12. Hon CC, Ramilowski JA, Harshbarger J, et al. An atlas of human long non-coding RNAs with accurate 5' ends. Nature. 2017;543:199-204.
13. Zhang K, Huang K, Luo Y, et al. Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data. BMC Genomics. 2014;15:845.
14. van Galen P, Kreso A, Mbong N, et al. The unfolded protein response governs integrity of the haematopoietic stem-cell pool during stress. Nature. 2014;510:268-72.
15. Treutlein B, Brownfield DG, Wu AR, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 2014;509:371-375.
16. Moignard V, Macaulay IC, Swiers G, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nat Cell Biol. 2013:363-72.
17. Han J, Liu S, Sun Z, et al. LncRNAs2Pathways: Identifying the pathways influenced by a set of lncRNAs of interest based on a global network propagation method. Sci Rep. 2017;7:46566.
18. Guttman, M. et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458(7235):223–227.
19. Huarte M, Guttman M, Feldser D, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell. 2010;142(3):409-419.
20. Yan X, Hu Z, Feng Y, et al. Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. Cancer Cell. 2015;28(4):529-540.
21. Paralkar VR, Mishra T, Luan J, et al. Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. Blood. 2014;123(12):1927-1937.
22. Liu K, Yan Z, Li Y, et al. Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis. Bioinformatics. 2013;29(17):2221-2222.
23. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006;22(13):1600-1607.
24. Cerami EG, Gross BE, Demir E, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39(Database issue):D685-690.
25. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23(2):257-258.
26. Frisch M, Klocke B, Haltmeier M, et al. LitInspector: literature and signal transduction pathway mining in PubMed abstracts. Nucleic Acids Res. 2009;37(Web Server issue):W135-40.

27. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
28. Angerer P1, Haghverdi L1, Büttner M1, et al. destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics. 2016;32(8):1241-1243.
29. Haghverdi L, Büttner M, Wolf FA, et al. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods. 2016;13(10):845-848.
30. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(Database issue):D447-452.
31. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-2504.
32. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034-50.
33. Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet. 206;22(1):1-5.
34. Wang J, Zhang J, Zheng H, et al. Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. Nature. 2004;431(7010):1 p following 757; discussion following 757.
35. Ranzani V, Rossetti G, Panzeri I, et al. The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4. Nat Immunol. 2015;16(3):318-325.
36. Brazão TF, Johnson JS, Müller J, et al. Long noncoding RNAs in B-cell development and activation. Blood. 2016;128(7):e10-9.
37. Washietl S, Kellis M, Garber M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Res. 2014;24(4):616-628.
38. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145-166.
39. Collier SP1, Collins PL, Williams CL, et al. Cuttingedge: influence of Tmevpg1, a long intergenic noncoding RNA, on the expression of Ifng by Th1 cells. J Immunol. 2012;189(5):2084-2088.
40. Gomez JA1, Wapinski OL, Yang YW, et al. The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-γ locus. Cell. 2013;152(4):743-754.
41. Ørom UA, Derrien T, Beringer M, et al. Long noncoding RNAs with enhancer-like function in human cells. Cell. 2010;143(1):46-58.
42. Guttman M, Donaghey J, Carey BW, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature. 2011;477(7364):295-300.

43. Garitano-Trojaola A, Agirre X, Prósper F, et al. Long non-coding RNAs in haematological malignancies. Int J Mol Sci. 2013;14(8):15386-15422.
44. He C, Hu H, Wilson KD, et al. Systematic Characterization of Long Noncoding RNAs Reveals the Contrasting Coordination of Cis- and Trans-Molecular Regulation in Human Fetal and Adult Hearts. Circ Cardiovasc Genet. 2016;9(2):110-118.
45. Hu G, Tang Q, Sharma S, et al. Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. Nat Immunol. 2013;14(11):1190-1198.
46. Schwarzer A, Emmrich S, Schmidt F, et al. The non-coding RNA landscape of human hematopoiesis and leukemia. Nat Commun. 2017;8(1):218.
47. Kim DH, Marinov GK, Pepke S, et al. Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. Cell Stem Cell. 2015;16(1):88-101.
48. Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012;22(9):1775-1789.
49. Laurenti E, Doulatov S, Zandi S, et al. The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. Nat Immunol. 2013;14(7):756-763.
50. Velten L, Haas SF, Raffel S, et al. Human haematopoietic stem cell lineage commitment is a continuous process. Nat Cell Biol. 2017;19(4):271-281.

Table Legends

***Online Supplementary File 1*. Identified lncRNAs by *de novo* transcriptome assembly in single human CD34$^+$ cells**

***Online Supplementary File 2*. Identified lncRNAs supported by CAGE data in single human CD34$^+$ cells**

***Online Supplementary Table S3*. Differentially expressed lncRNAs in CD38$^-$ and CD38$^+$ cells**

***Online Supplementary Table S4*. WGCNA Modules**

***Online Supplementary Table S5*. Lineage specific lincRNAs**

***Online Supplementary Table S6*. List of lncRNAs for RT-PCR validation**

***Online Supplementary Table S7*. Sequentially expressed mRNAs along two differential trajectories**

***Online Supplementary Table S8*. Sequentially expressed lncRNAs along two differential trajectories**

***Online Supplementary Table S9*. Differentially expressed mRNAs in MDS cells and pathway analysis**

***Online Supplementary Table S10*. Differentially expressed lncRNAs in MDS cells**

***Online Supplementary Table S11*. Differentially expressed lncRNAs in monosomy 7 cells**

Figure Legends

Supplementary Figure S1. FACS sorting strategy and samples for RNA sequencing (A) Bone marrow mononuclear cell (BM-MNC) isolation was performed with fresh BM samples. BM-MNCs were stained with lineage markers in Pacific blue (anti-CD3, anti-CD14, and anti-CD19), anti-CD34-PE, and anti-CD38-APC. (B) For four healthy donors and patient 4, Lineage-$CD34^+CD38^-$ and Lineage-$CD34^+CD38^+$ cell populations were sorted and subjected to sequencing separately, while for patients 1, 2, 3, and 5, only the $CD34^+$ cell population was sorted and subjected to sequencing due to limited cell numbers. Total numbers of cells of individual samples subjected to further analysis after sequencing data filtering were shown.

Supplementary Figure S2. FACS sorting strategy and bulk samples for RT-PCR (A) BM-MNC isolation was performed with fresh BM samples. BM-MNCs were stained with lineage markers in Pacific blue (anti-human lineage cocktail: anti-CD3, anti-CD14, anti-CD16, anti-CD19, anti-CD20, and anti-CD56), anti-CD34-PE, anti-CD38-APC, and anti-CD45RA-BV510. (B) For three new healthy donors, HSC/MLP (Lin-$CD34^+CD38^-$), B/NK/GMP (Lin-$CD34^+CD38^+CD45RA^+$), and CMP/MEP (Lin-$CD34^+CD38^+CD45RA^-$) populations were sorted and subjected to RNA extraction and RT-PCR analysis in triplicate.

Supplementary Figure S3. Characterization of lncRNAs expressed in single human $CD34^+$ cells. (A) Distribution of lncRNAs expressed in human $CD34^+$ cells across human genome. Outer track shows human chromosomes 1 to 22; middle track shows positions of 2,892 lncRNAs defined in single human $CD34^+$ cells, with gene names on inner track. (B) Percentages of single human $CD34^+$ cells with expression of lncRNAs or mRNAs. The number of exons (C), length of transcripts (D), species conservation (E), and distribution of distance of annotated and novel lncRNAs to TSS of mRNA (F) in single human $CD34^+$ cells. (G) Co-expression of lncRNAs or mRNAs with neighboring mRNAs. CDF, cumulative distribution function.

Supplementary Figure S4. Single cell RNA sequencing is more powerful than bulk approach in assessing gene expression with high variation. Expression of Gene A and Gene B in single cells and in bulk samplse composed with the same cells are shown on the left and right, respectively. Mean/median expression levels of

Gene A and Gene B in single cells, standard deviation, and standard error of mean are shown on the right. Gene B had the same mean expression levels with Gene A in bulk samples but higher cell-to-cell variation. Given the threshold of detecting gene expression is 30, using bulk samples, both Gene A and Gene B were undetectable. However, with single cell approach, Gene B could be detected in two single cells (labeled in dark red) while Gene A was still undetectable.

Supplementary Figure S5. Identification of highly variable genes (HVGs) across cells and batch correction. (A) Seurat software was employed to identify HVGs across lncRNAs expression in single cell RNA sequencing data, in which a z-score cutoff of 0.5 was applied to increase the power of unsupervised dimensional reduction techniques: only lncRNA-encoding genes showing expression > 0.5 and variance > 0.5 STD were retained for analysis (in black square). (B) Single cells from four healthy donors were plotted in a tSNE plot and colored respectively, without batch correction (left) and with batch correction (right).

Supplementary Figure S6. Cell-to-cell variation of lncRNA and mRNA expression in individual cell populations. Variance of lncRNA and mRNA expression in HSC, MLP, MEP, GMP, ProB, and ETP were plotted, respectively. x axis, Log(TPM+1) of gene expression; y axis, variance. Grey and red dots represent mRNA and lncRNA expression in single cells, respectively.

Supplementary Figure S7. Agreement of mRNA and lncRNA expression in MEP between the current study and the single cell RNA sequencing dataset GSE75478. GSE75478 is a study to characterize early human hematopoiesis on a single cell level which combines flow-cytometry, single cell transcriptome, and single cell lineage fate data.[49] Healthy human bone marrow cells were labeled with FACS surface makers to identify different subpopulations of human HSPCs before subjecting to single cell RNA sequencing and other experiments. (A) Log fold changes of mRNA (left) and lncRNA (right) expression in MEPs vs. non-MEPs in the current study were plotted on x axis, while those in the GSE75478 dataset were plotted on y axis, showing a linear correlation for both mRNA and lncRNA expression between two studies. (B) mRNAs (left) and lncRNAs (right) that were preferentially expressed in MEPs were defined using one-side Pearson comparison between

MEPs and non-MEPs with a *P* value < 0.05. There were 3,934 common mRNAs and 85 common lncRNAs that were preferentially expressed in MEPs in both studies, which were significantly higher than expected number of common genes ($P < 0.0001$).

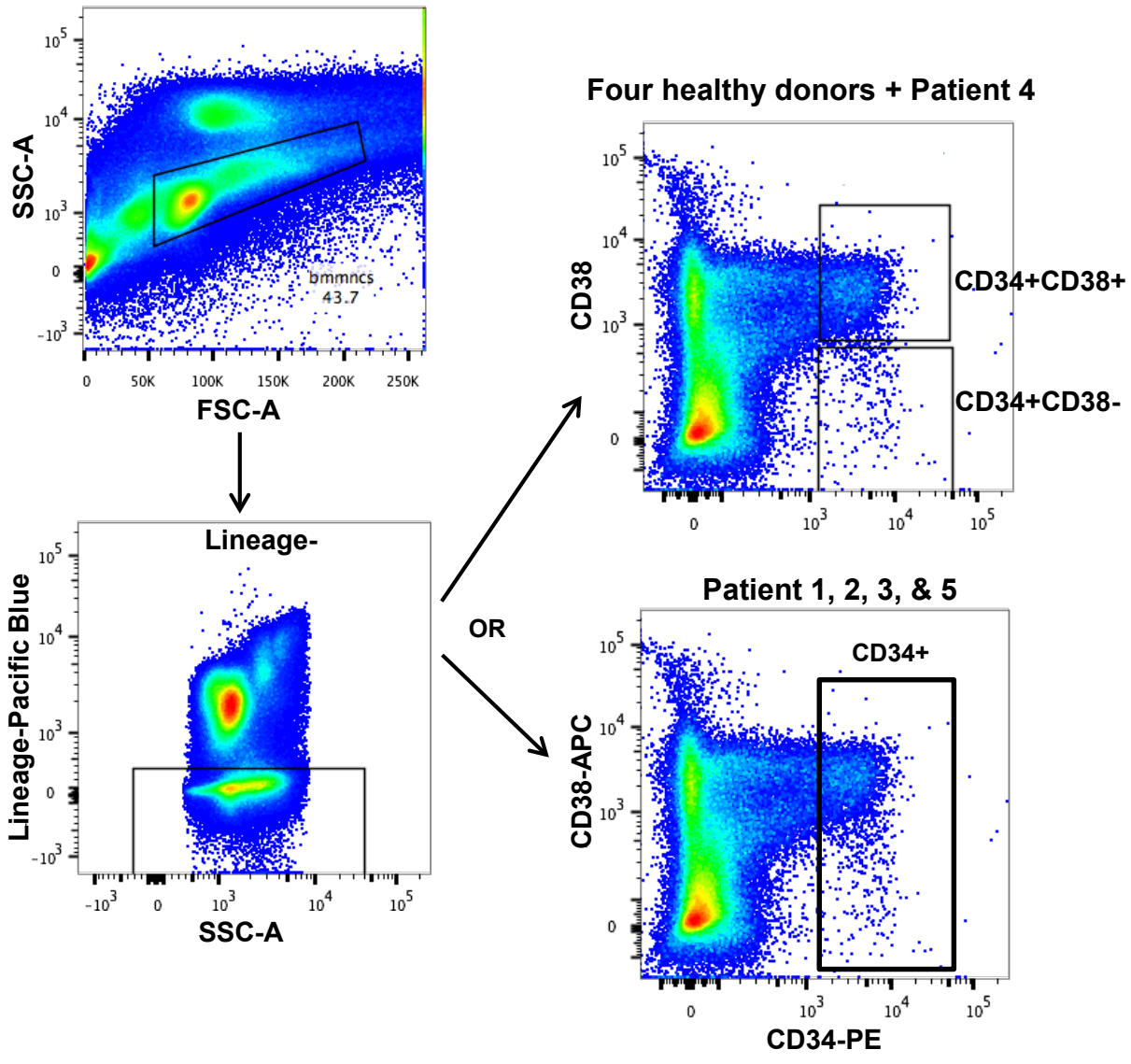
Supplementary Figure S8. Agreement of mRNA and lncRNA expression in HSCs between the current study and the single cell RNA sequencing dataset GSE75478. (A) Log fold changes of mRNA (left) and lncRNA (right) expression in HSCs vs. non-HSCs in the current study were plotted on x axis, while those in the GSE75478 dataset were plotted on y axis, resulting in a linear correlation was observed for both mRNA and lncRNA expression between two studies. (B) mRNAs (left) and lncRNAs (right) that were preferentially expressed in HSCs were defined using one-side Pearson comparison between HSCs and non-HSCs with a *P* value < 0.05. There were 269 common mRNAs and 19 common lncRNAs that were preferentially expressed in HSCs in both studies, which were significantly higher than expected number of common genes ($P < 0.0001$).

Supplementary Figure S9. Quantitative RT-PCR analysis of the expression of signature lncRNAs and mRNAs. (A) Single cell samples. Lineage specific mRNAs and lncRNAs were plotted on x axes. y axes show ΔΔCT values (ddCT) of gene expression in HSC/MLPs vs. non-HSC/MLPs (top); MEPs vs. non-MEPs (middle); and ProBs vs. non-ProBs, ETPs vs. non-ETPs, and GMPs vs. non-GMPs (bottom). (B) Flow cytometry sorted bulk samples. Lineage specific mRNAs and lncRNAs were plotted on x axes. y axes represent ΔΔCT values (ddCT) of gene expression in HSC/MLPs vs. non-HSC/MLPs (top) and MEPs vs. non-MEPs (bottom). (C) Expression of representative lineage-specific mRNAs for HSC/MLP, MEP, and ProB/ETP/GMP along differentiation trees, measured by quantitative RT-PCR analysis. Expression are presented as relative quantity in one population vs. expression in all the others.

Supplementary Figure S10. Integration with CAGE dataset. (A) A schematic diagram of defined lncRNAs by *de novo* transcriptome assembling in the current study and in the CAGE data. There was overlap of lncRNAs defined by our pipeline and the CAGE study. Low expressed transcripts were filtered in our pipeline; however, we lowered threshold of expression levels, and those transcripts with medium expression levels and supported

by CAGE were also considered to be expressed in human $CD34^+$ cells. The CAGE study defined much more lncRNAs than our study probably because of a large number of cell types included. On the other hand, our single cell approach may have advantage in defining transcripts in a minority of cells. (B) Log fold changes of mRNA (left) and lncRNA (right) expression in MEP-likes vs. non-MEP-likes in the CAGE study were plotted on x axis, while those in the current study were plotted on y axis. A correlation with R value = 0.442 (mRNAs) and R value = 0.243 (lncRNAs) was observed between two studies.

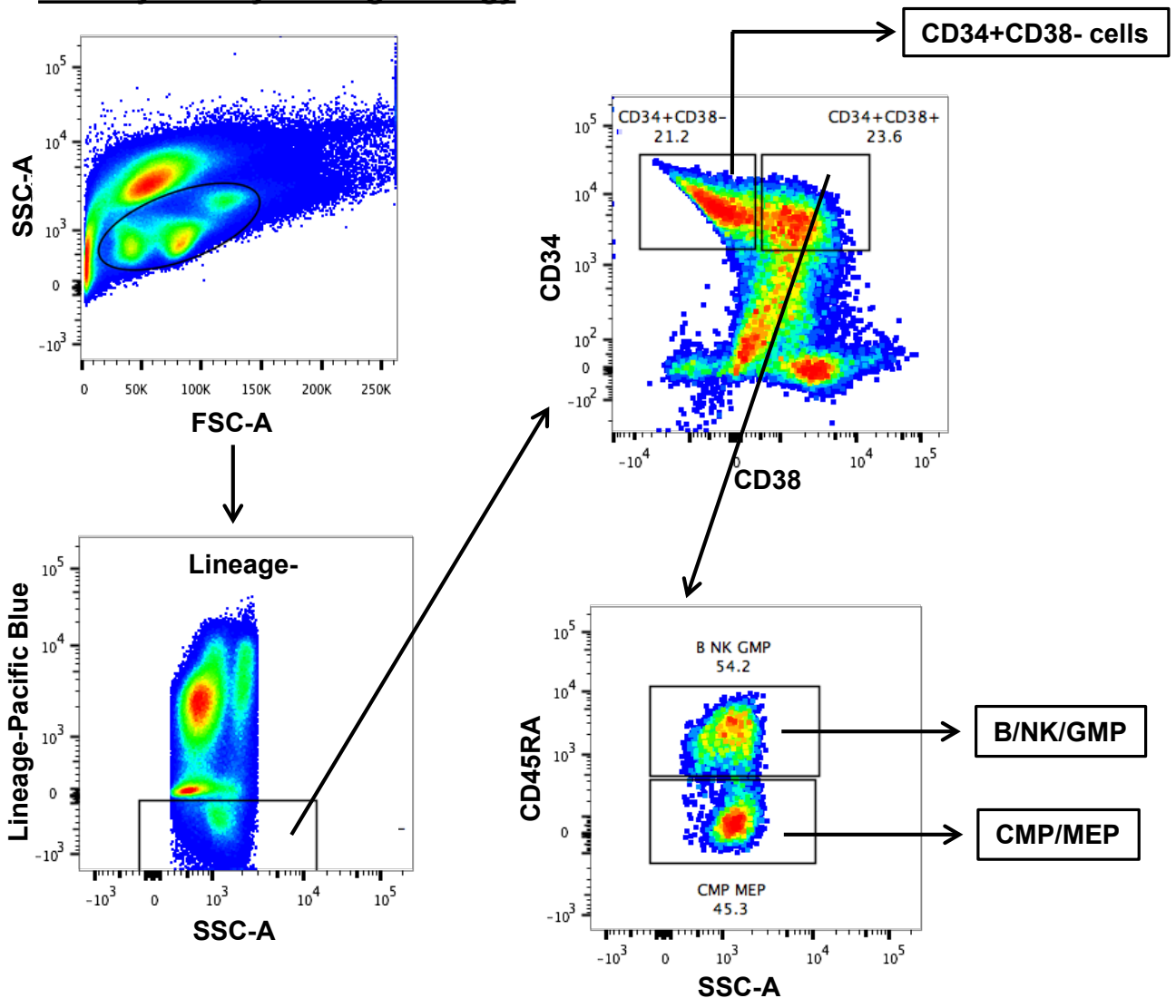
A **Flow cytometry sorting strategy**



B **Purification and scRNA-seq of human HSPCs**

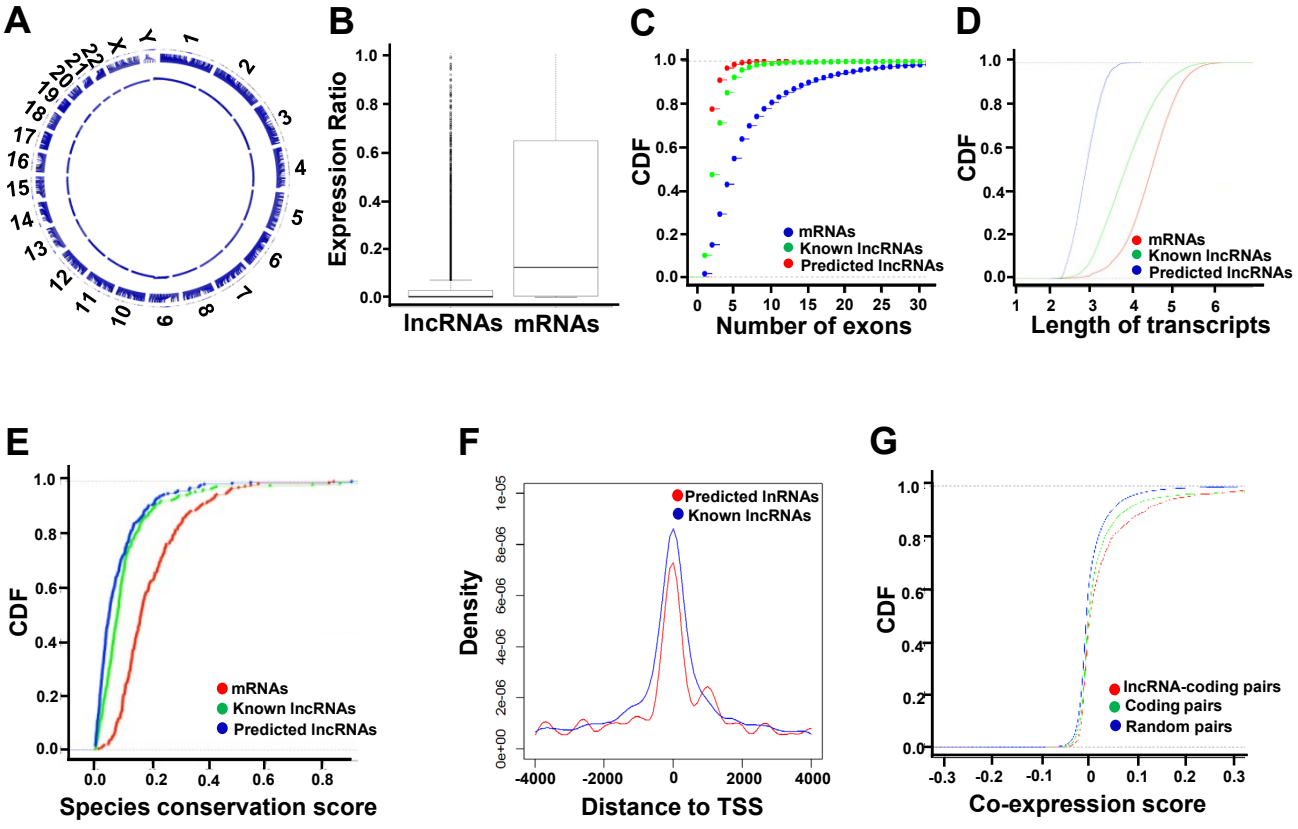
Subject	Age, y /sex	Sorting phenotype	#Total CD34+ Cells after analytical fitering
Healthy Donor 1	58/M	Lin-CD34+CD38-, Lin-CD34+CD38+	76
Healthy Donor 2	34/M	Lin-CD34+CD38-, Lin-CD34+CD38+	90
Healthy Donor 3	57/F	Lin-CD34+CD38-, Lin-CD34+CD38+	98
Healthy Donor 4	31/M	Lin-CD34+CD38-, Lin-CD34+CD38+	127
Patient 1	54/F	Lin-CD34+	78
Patient 2	6.5/M	Lin-CD34+	148
Patient 3	56/F	Lin-CD34+	103
Patient 4	59/M	Lin-CD34+CD38-, Lin-CD34+CD38+	206
Patient 5	67/M	Lin-CD34+	53

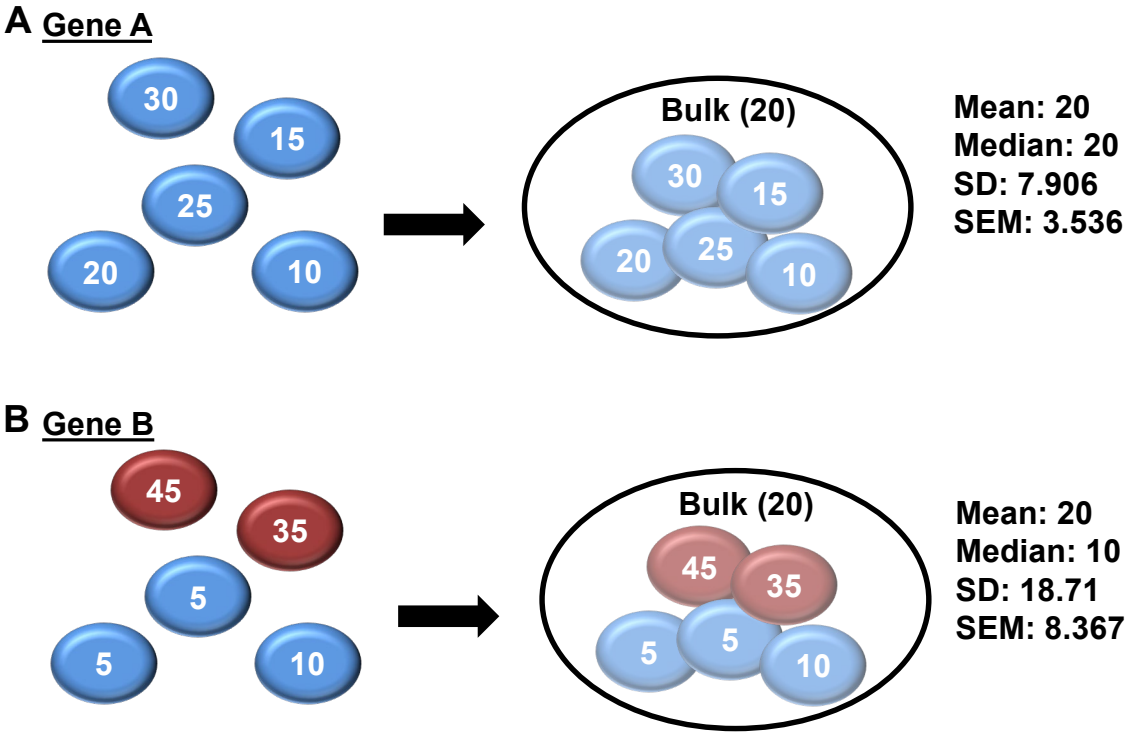
A Flow cytometry sorting strategy



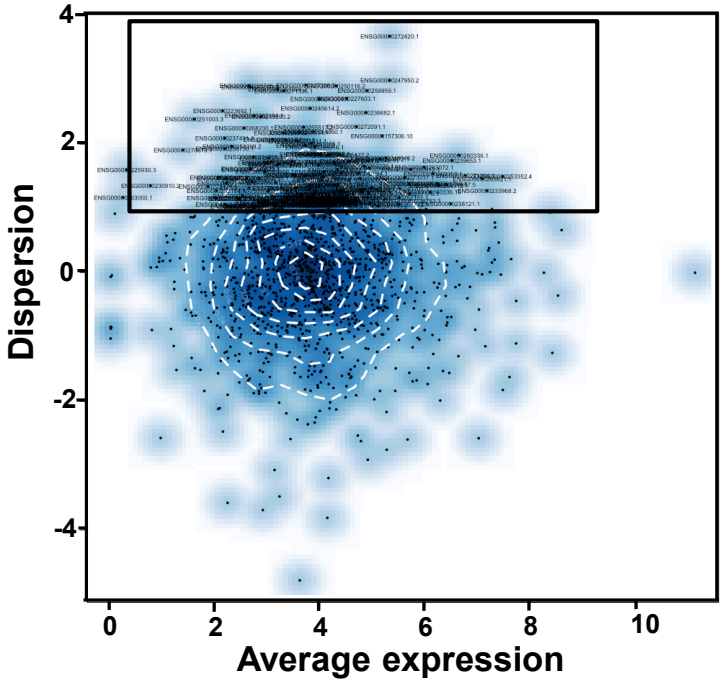
B Purification and RT-PCR of human HSPCs bulk samples

Subject	Age, y/sex	Sorting phenotype	Cell Population
Healthy Donor 5	47/M	Lin-CD34+CD38-	HSC/MLP
Healthy Donor 6	51/M	Lin-CD34+CD38+CD45RA+	B/NK/GMP
Healthy Donor 7	44/F	Lin-CD34+CD38-CD45RA-	CMP/MEP

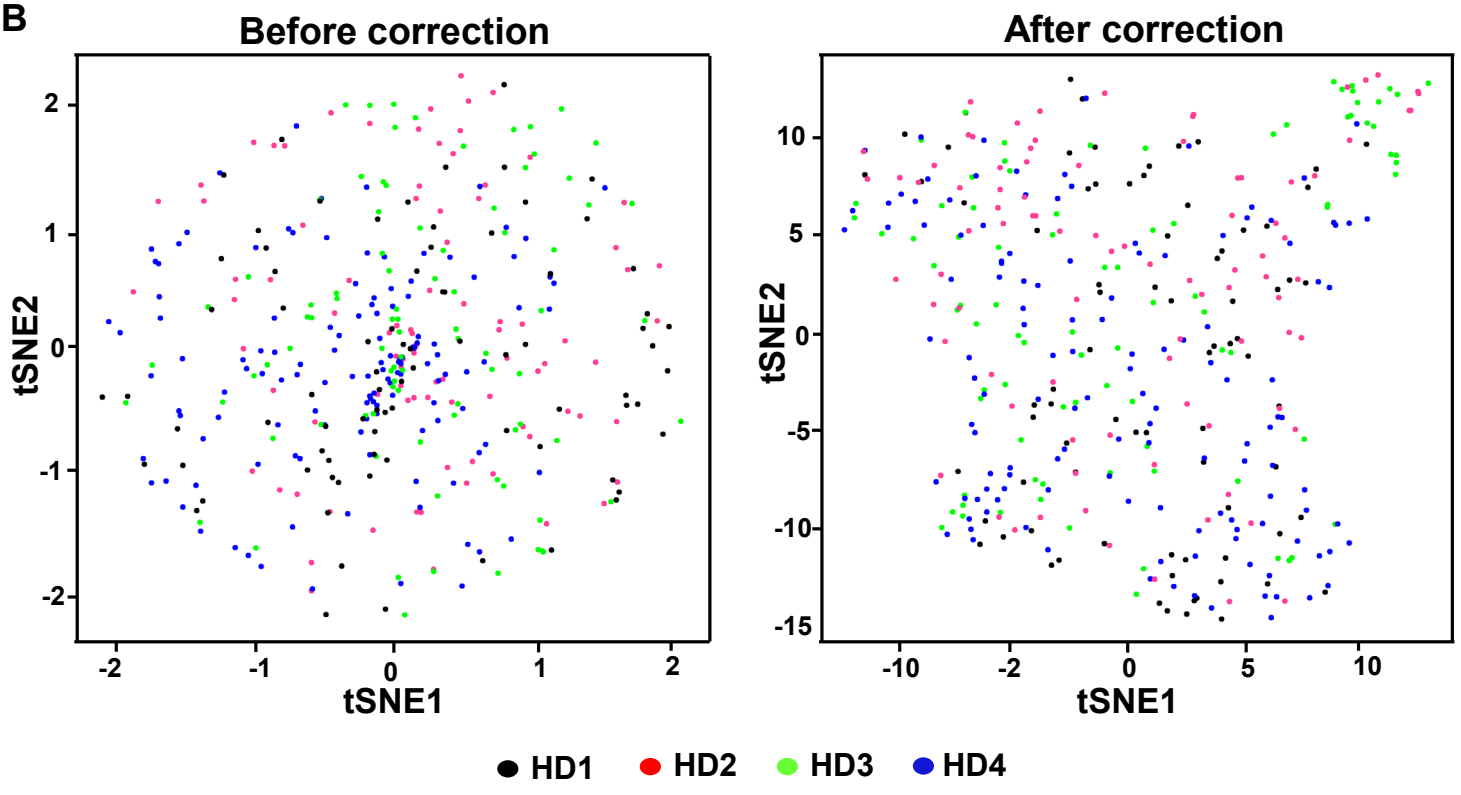

A
B
Expression Ratio
lncRNAs
mRNAs
C
CDF
Number of exons
mRNAs
Known lncRNAs
Predicted lncRNAs
D
CDF
Length of transcripts
mRNAs
Known lncRNAs
Predicted lncRNAs
E
CDF
Species conservation score
mRNAs
Known lncRNAs
Predicted lncRNAs
F
Density
Distance to TSS
Predicted lnRNAs
Known lncRNAs
G
CDF
Co-expression score
lncRNA-coding pairs
Coding pairs
Random pairs

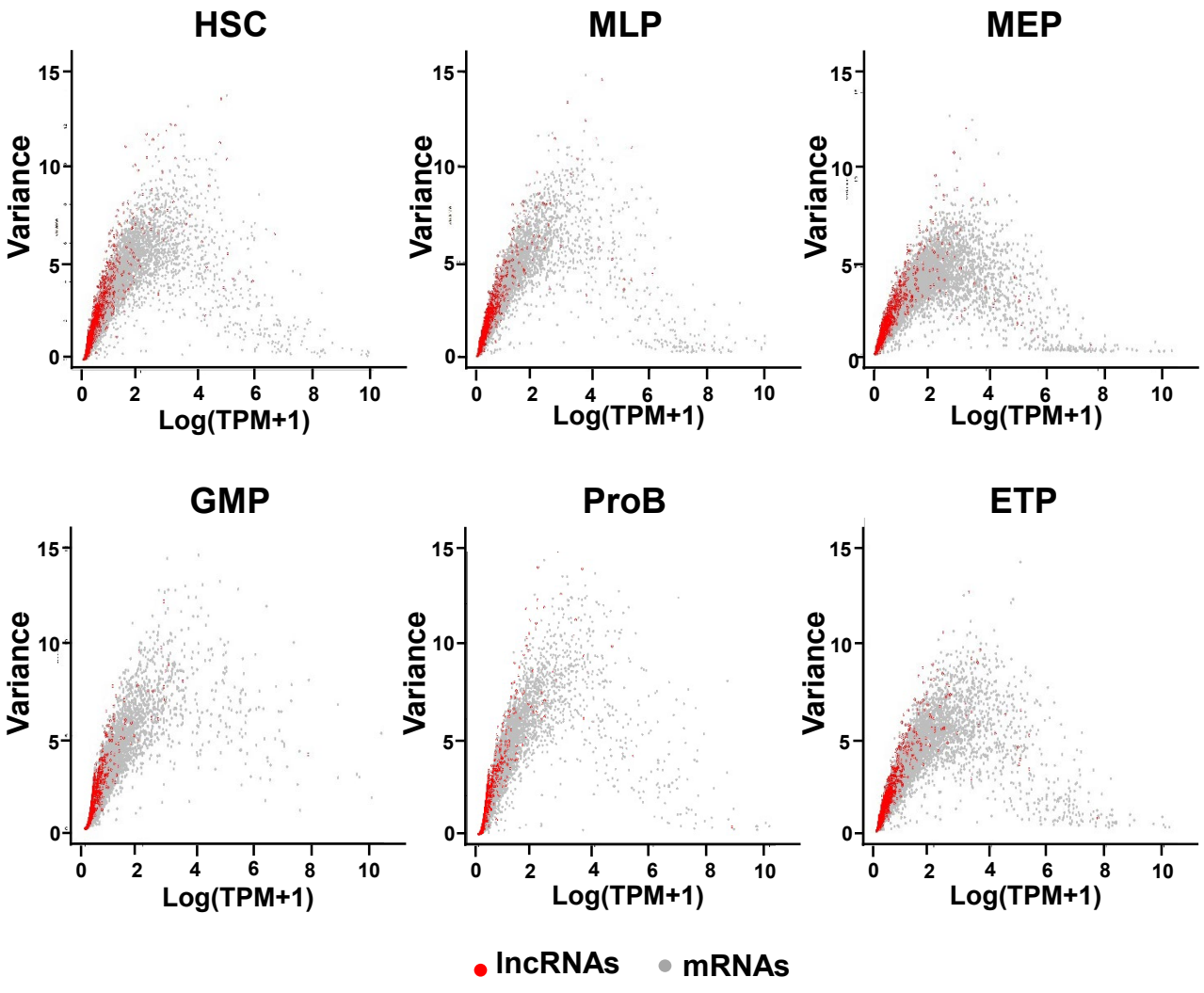

A Gene A
30
15
25
20
10
Bulk (20)
30
15
20
25
10
Mean: 20
Median: 20
SD: 7.906
SEM: 3.536
B Gene B
45
35
5
5
10
Bulk (20)
45
35
5
5
10
Mean: 20
Median: 10
SD: 18.71
SEM: 8.367

A Global transcriptional interrelation of lncRNAs (Healthy 1 - 4)

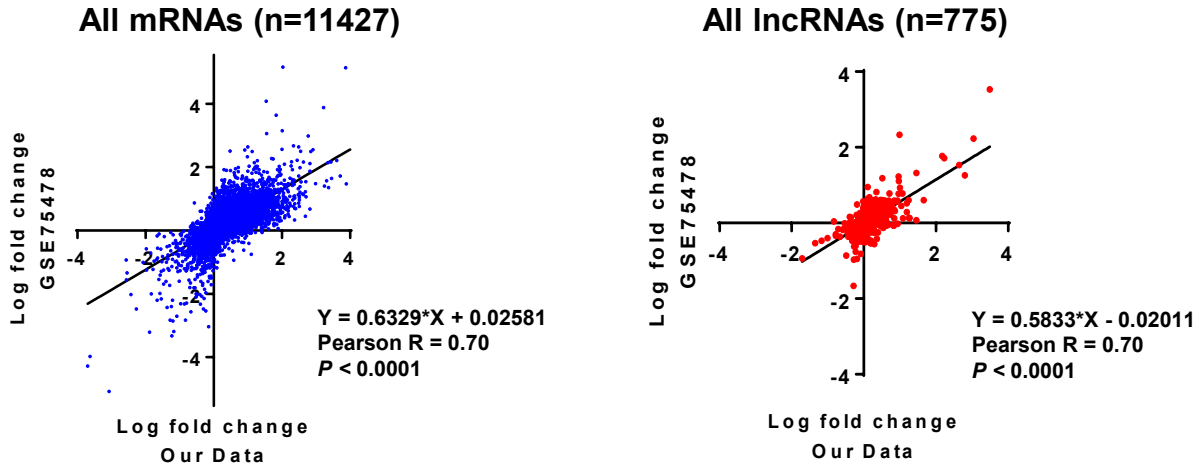


B

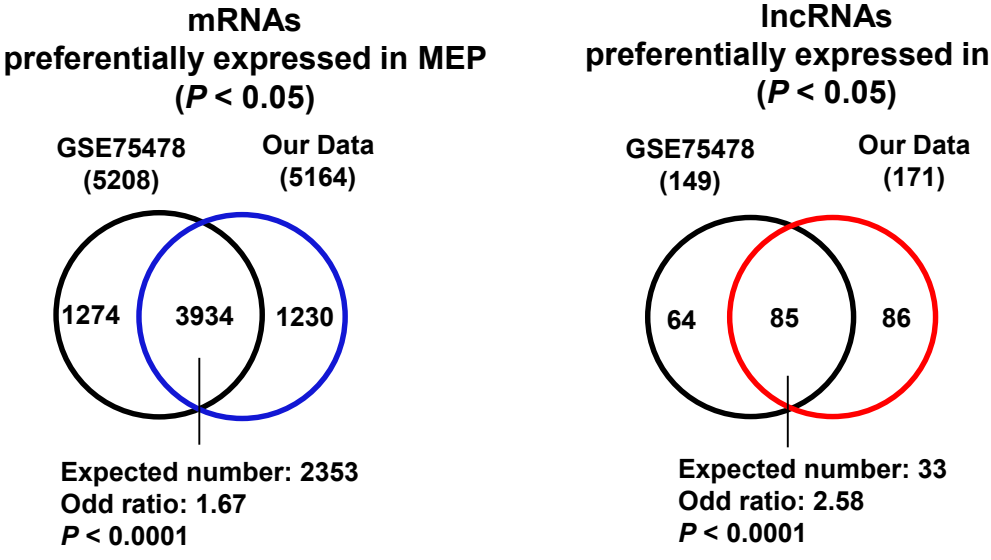



Variance among cells: lncRNA vs. mRNA
HSC
MLP
MEP
GMP
ProB
ETP
Variance
Log(TPM+1)
lncRNAs
mRNAs

A Fold change of gene expression in MEPs vs. non-MEPs

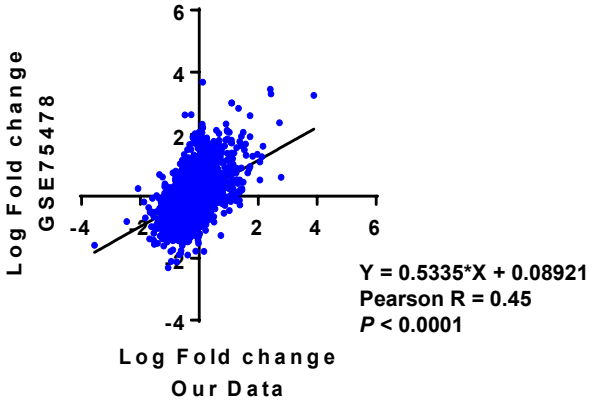


B Preferentially expressed genes in MEPs

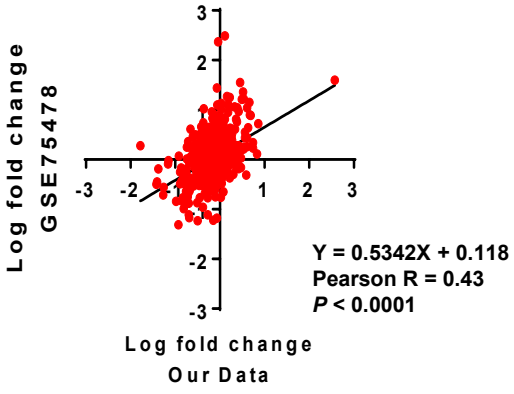


A Fold change of gene expression in HSCs vs. non-HSCs

All mRNAs (n = 11427)

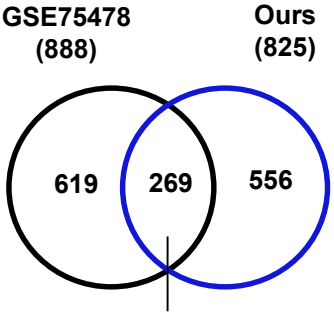


All lncRNAs (n = 775)



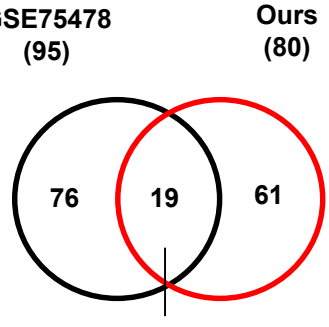
B Preferentially expressed genes in HSCs

mRNAs preferentially expressed in HSC (*P* < 0.05)

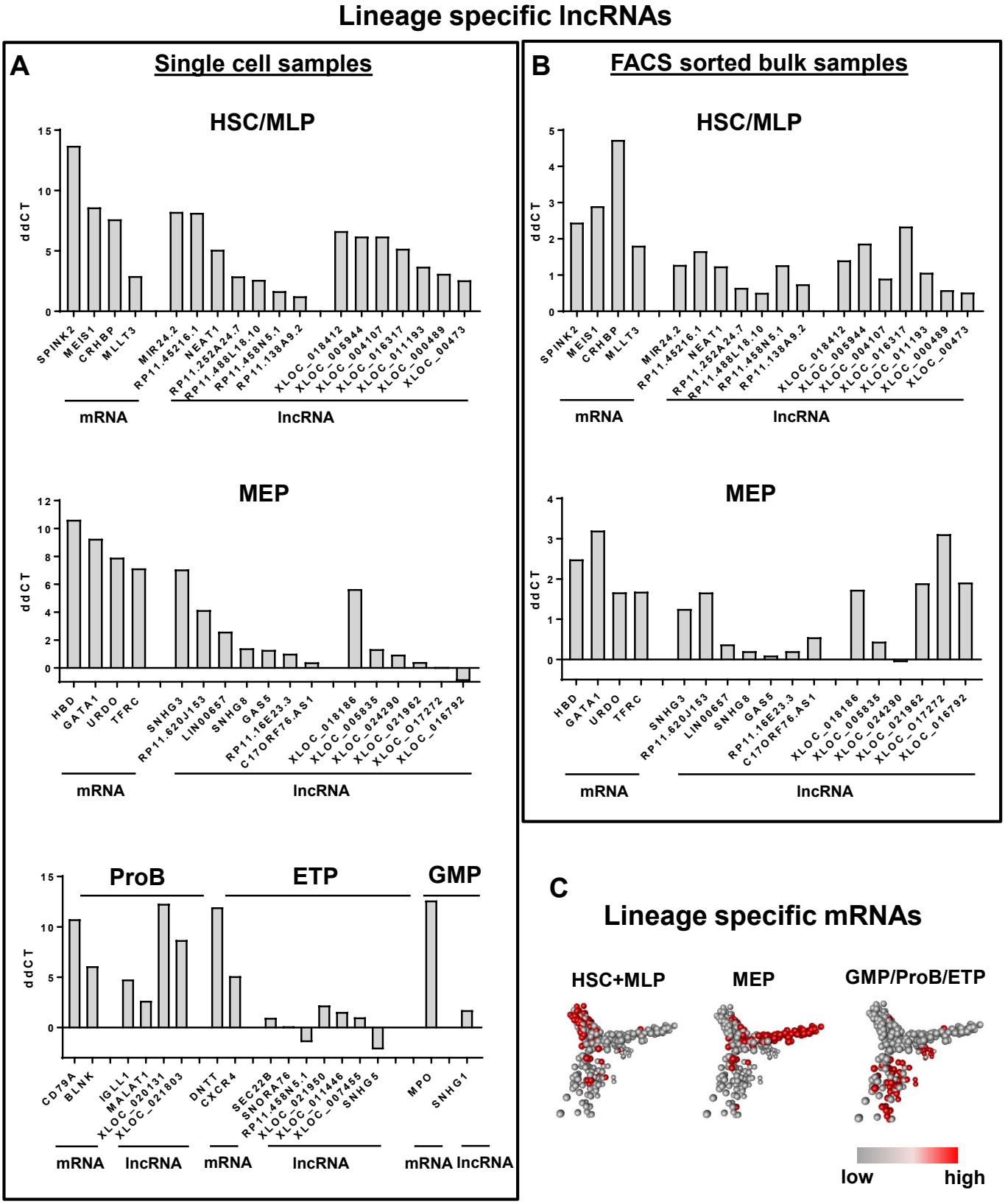


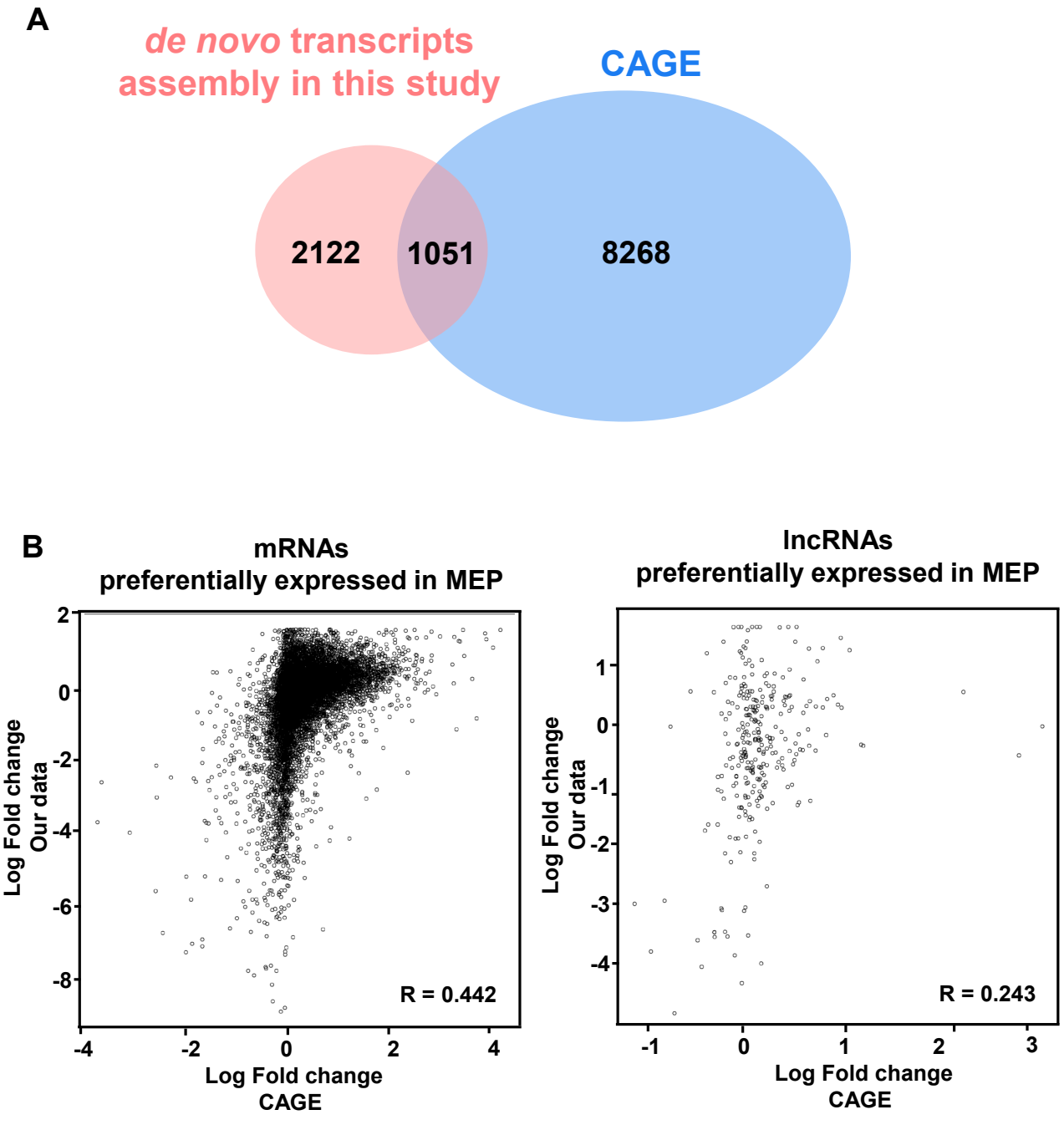
Expected number: 64
Odd ratio: 4.20
P < 0.0001

lncRNAs preferentially expressed in HSC (*P* < 0.05)



Expected number: 10
Odd ratio: 1.90
P < 0.0001


Lineage specific lncRNAs
A
Single cell samples
HSC/MLP
ddCT
SPINK2
MEIS1
CRHBP
MLLT3
MIR24.2
RP11.45216.1
NEAT1
RP11.252A24.7
RP11.488L18.10
RP11.458N5.1
RP11.138A9.2
XLOC_018412
XLOC_005944
XLOC_004107
XLOC_016317
XLOC_011193
XLOC_000489
XLOC_00473
mRNA
lncRNA
MEP
HBD
GATA1
URDO
TFRC
SNHG3
RP11.620J153
LIN00657
SNHG8
GAS5
RP11.16E23.3
C17ORF76.AS1
XLOC_018186
XLOC_005835
XLOC_024290
XLOC_021962
XLOC_017272
XLOC_016792
ProB
ETP
GMP
CD79A
BLNK
IGLL1
MALAT1
XLOC_020131
XLOC_021803
DNTT
CXCR4
SEC22B
SNORA76
RP11.458N5.1
XLOC_021950
XLOC_011446
XLOC_007455
SNHG5
MPO
SNHG1
B
FACS sorted bulk samples
C
Lineage specific mRNAs
HSC+MLP
MEP
GMP/ProB/ETP
low
high


A
de novo transcripts assembly in this study
CAGE
2122
1051
8268
B
mRNAs preferentially expressed in MEP
Log Fold change Our data
Log Fold change CAGE
R = 0.442
lncRNAs preferentially expressed in MEP
Log Fold change Our data
Log Fold change CAGE
R = 0.243