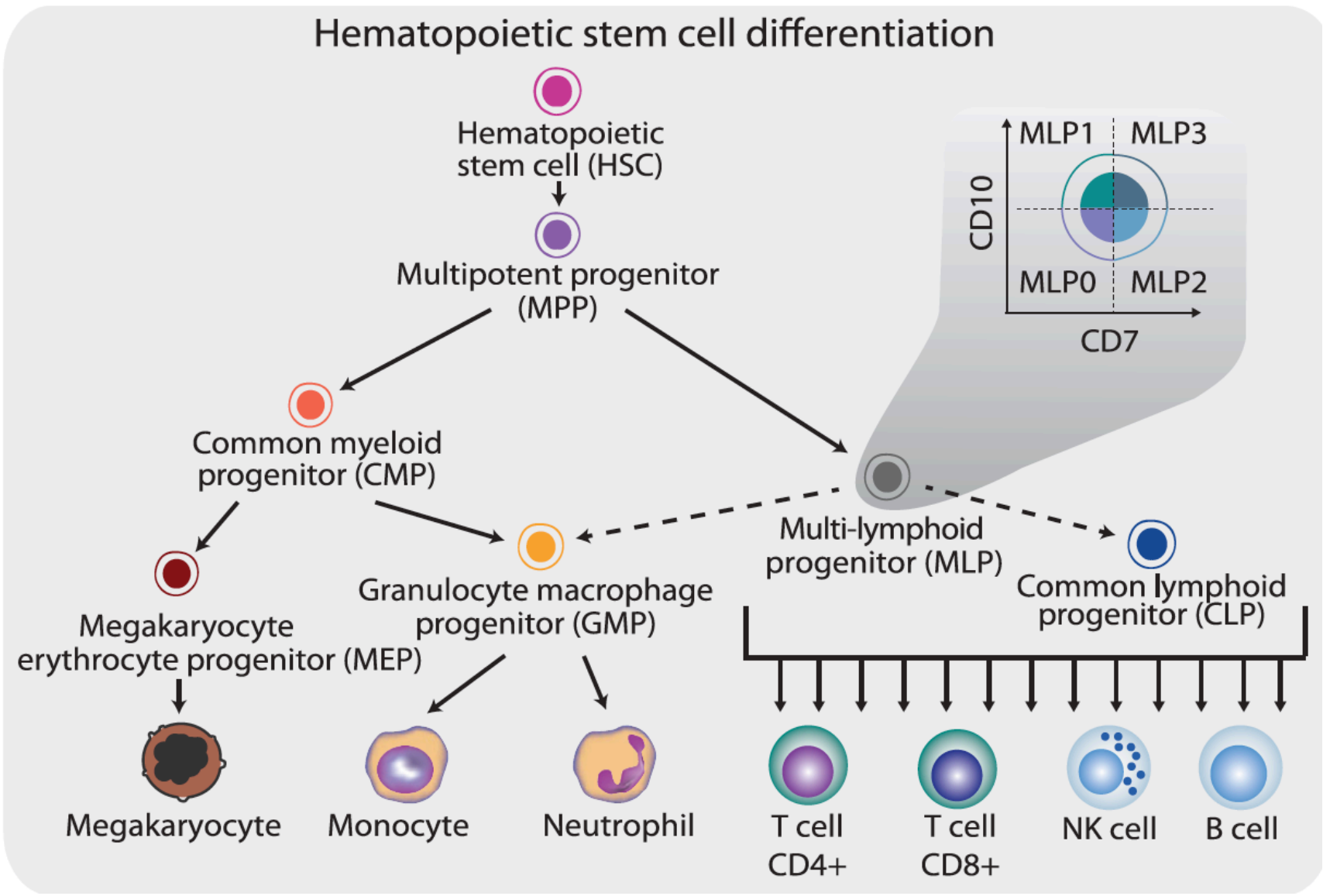
Team Bloodies: Data-driven analysis of the potential candidate transcription factors in hematopoietic stem cell differentiation into multiple progenitor compartments.

Wang, Fangwu; Hoque, Rawnak; Paul, Somdeb; Cavalla, Annie
University of British Columbia, Statistical Methods for High Dimensional Biology

INTRODUCTION

Human hematopoietic stem cells (HSCs) are capable of regenerating the lifelong production of all types of mature blood cells, which is the basis of curative HSC transplantation therapies for numerous hematologic malignancies and gene therapy protocols. Understanding the mechanisms regulating the self-renewal and lineage restriction of HSCs has great clinical value. HSC is thought to acquire multi-step lineage restriction through going down multiple progenitor populations, during which process the myeloid vs. lymphoid binary decision is made with subsequent progeny restricted to either fate. However, recent evidence showed conversion between "myeloid-committed" progenitors and "lymphoid-committed" progenitors, suggesting a more fluid program of hematopoietic cell differentiation[1].

Recent studies implied that the chromatin state, especially of enhancers, foreshadows transcriptional programs in differentiated cells. Transcription factors (TFs) and their interactions within gene regulatory regions are central to the cell fate determination.[2,3] Farlik M. *et al* recently generated DNA methylation profiles with corresponding RNA-seq data for HSC and six progenitor populations, highlighting the important role of DNA methylation in cell differentiation[4].


Hematopoietic stem cell differentiation
Hematopoietic stem cell (HSC)
Multipotent progenitor (MPP)
Common myeloid progenitor (CMP)
Multi-lymphoid progenitor (MLP)
Megakaryocyte erythrocyte progenitor (MEP)
Granulocyte macrophage progenitor (GMP)
Common lymphoid progenitor (CLP)
MLP1
MLP3
MLP0
MLP2
CD10
CD7
Megakaryocyte
Monocyte
Neutrophil
T cell CD4+
T cell CD8+
NK cell
B cell

OBJECTIVES

Aim: Identify TFs potentially responsible for the cell differentiation program in a data-driven approach.

Hypothesis

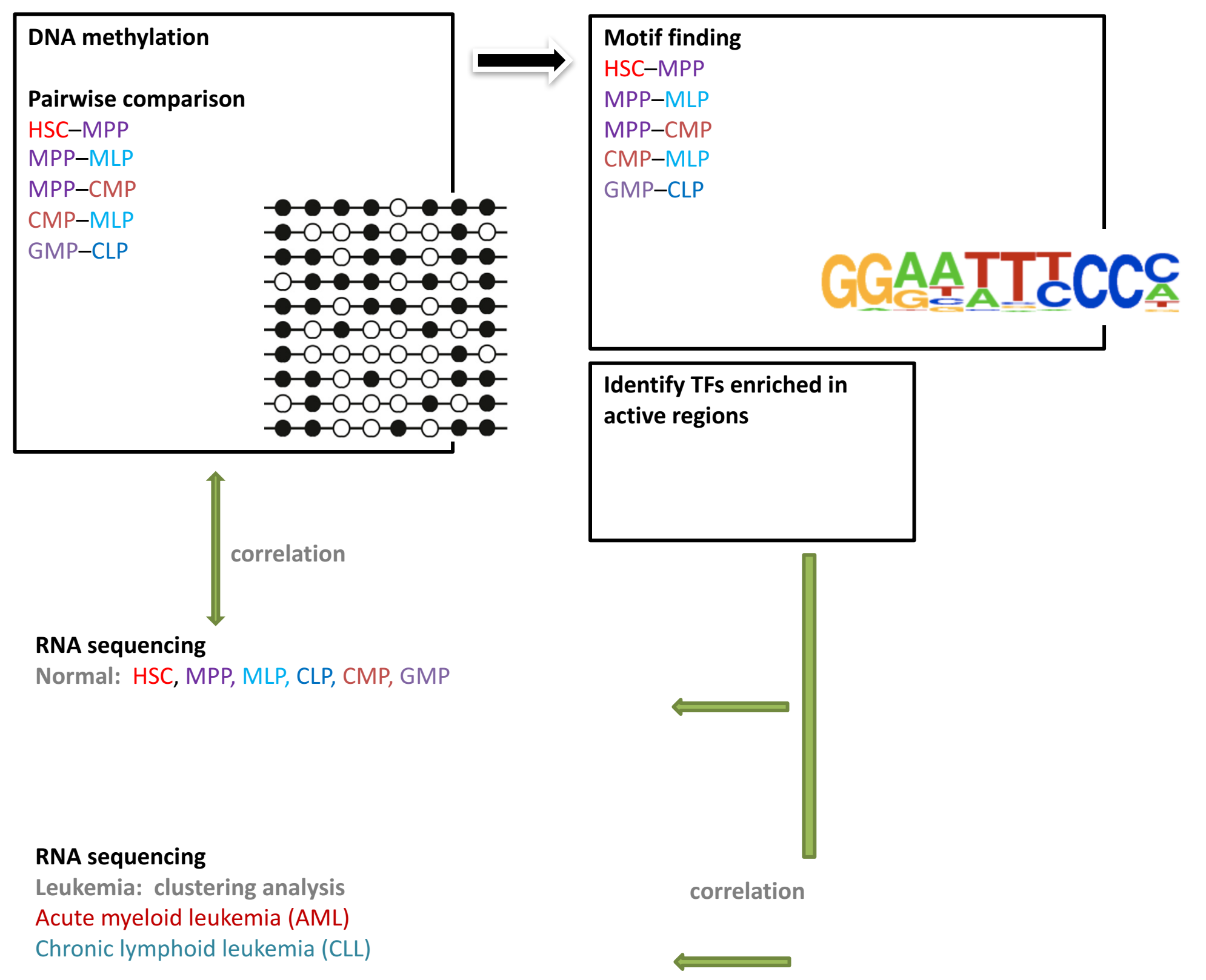
I. DNA methylation status reflects cell identity of closely related progenitor populations during differentiation.

II. Active DNA methylation regions (hypomethylated promoters, enhancers) are associated with the binding of active TFs that regulates the cell fate in the particular progenitor population.

METHODS

We chose the big dataset generated by the Farlik M. *et al* publication with matched DNA methylation and RNA-seq data[4]. The authors previously characterized the differentially methylated regions with TF binding events based on a ChIP-seq database from cell lines of all tissue origins, focusing exclusively on promoter regions.

Different strategy from the published paper: To more rigorously identify TFs with a potential function in cell differentiation, we annotated DNA methylation using both promoters and enhancers. The enhancer regions were defined from two hematopoietic cell lines.



DNA methylation

Pairwise comparison
HSC–MPP
MPP–MLP
MPP–CMP
CMP–MLP
GMP–CLP

Motif finding
HSC–MPP
MPP–MLP
MPP–CMP
CMP–MLP
GMP–CLP

Identify TFs enriched in active regions

correlation

RNA sequencing
Normal: HSC, MPP, MLP, CLP, CMP, GMP

RNA sequencing
Leukemia: clustering analysis
Acute myeloid leukemia (AML)
Chronic lymphoid leukemia (CLL)

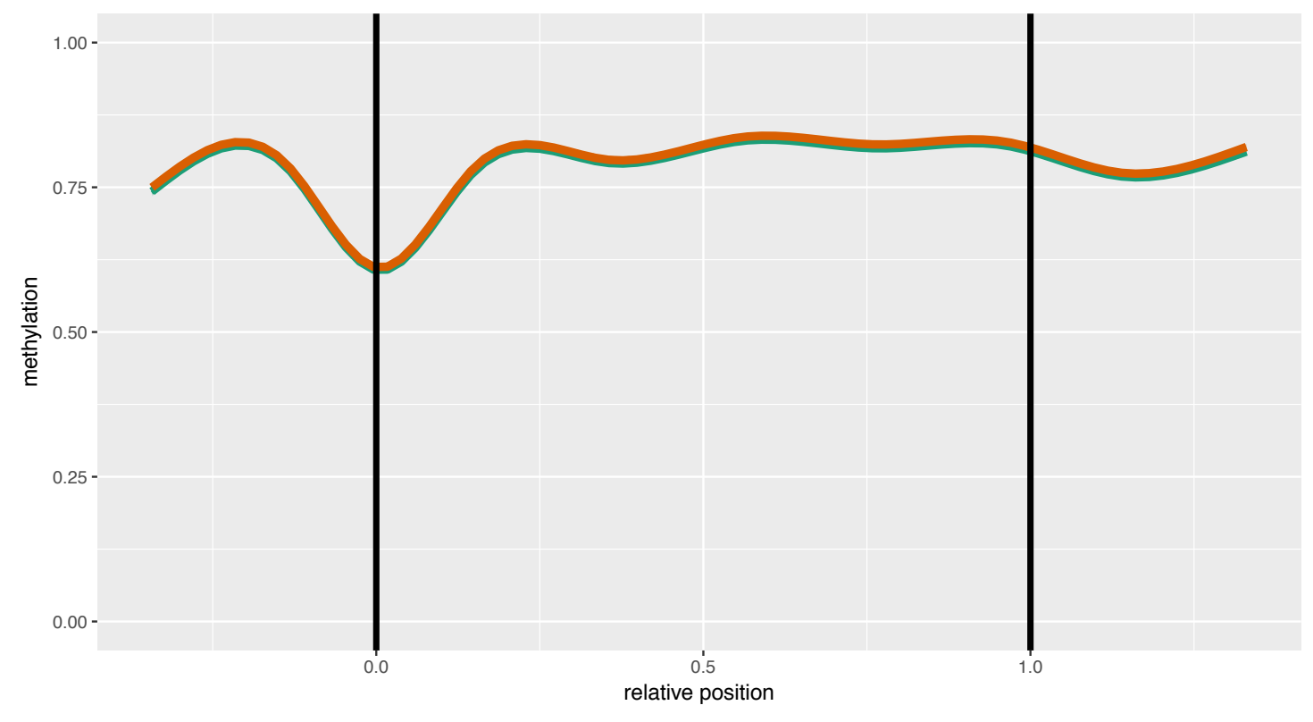
correlation

RESULTS

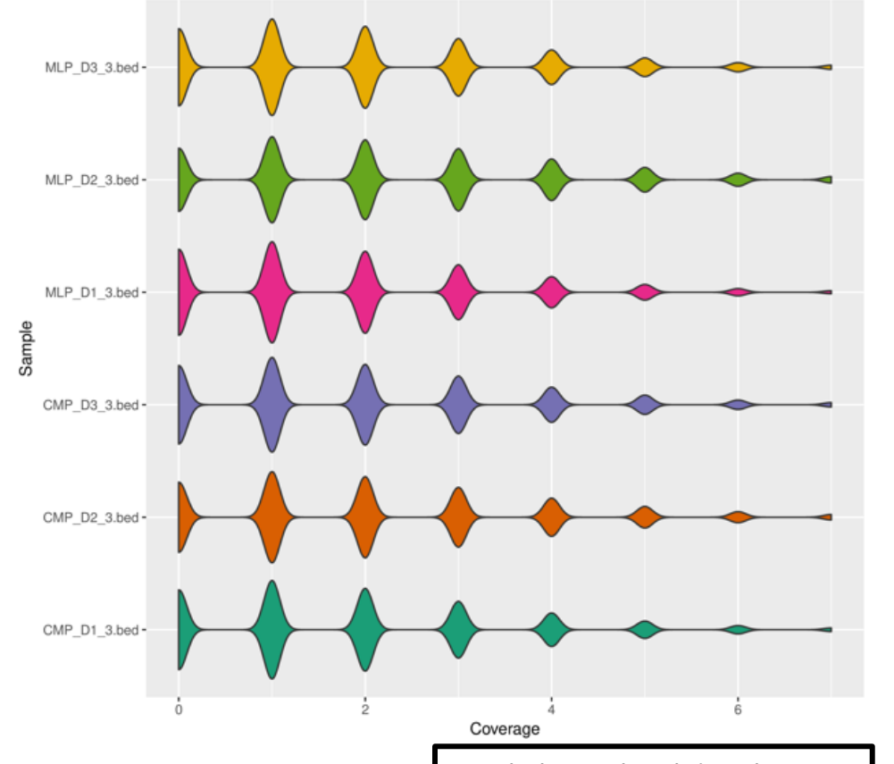
1. DNA methylation analysis on promoters and enhancers

Generation of merged data from technical replicates

To increase the overall coverage, we added up the reads from technical replicates (3 aliquots with 50, 50, 1000 cells from the same donor condition) using bedtools, which resulted in 3 merged datasets (3 biological replicates) for each cell type.



The coverage distribution across the cell types



Methylation level distribution across "Gene" region type

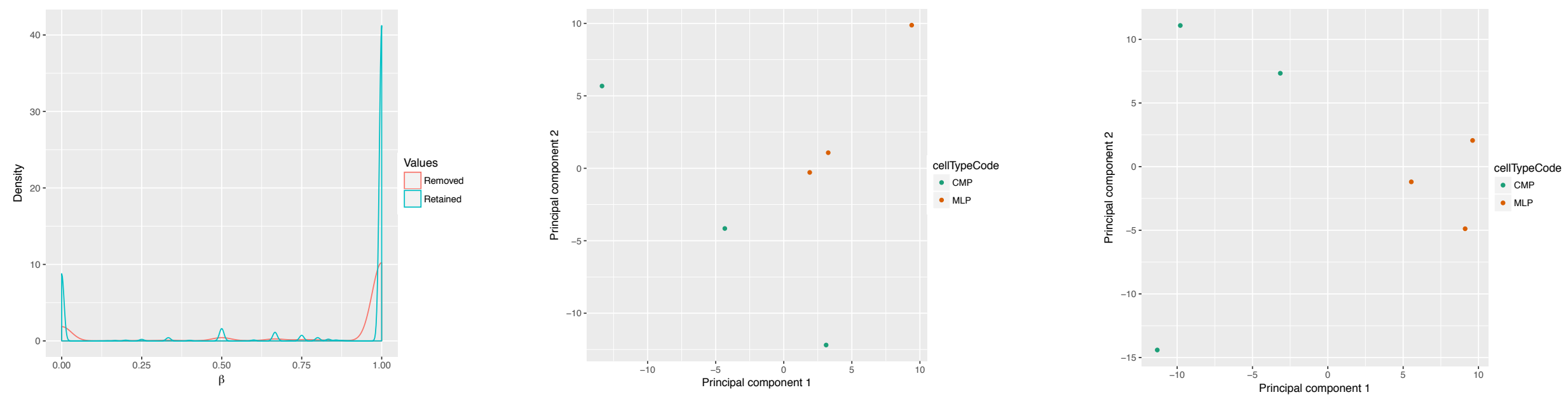
Before merging	
fileName	sites_covgMean
HSC_50_D7_1_R1	**1.060955918**
HSC_50_D7_2_R1	**1.147679985**
HSC_1000_D7_R1	**1.107996393**
After merging	
fileName	sites_covgMean
HSC_D7_3.bed	**1.318091381**

Comparison of mean coverage before and after merging

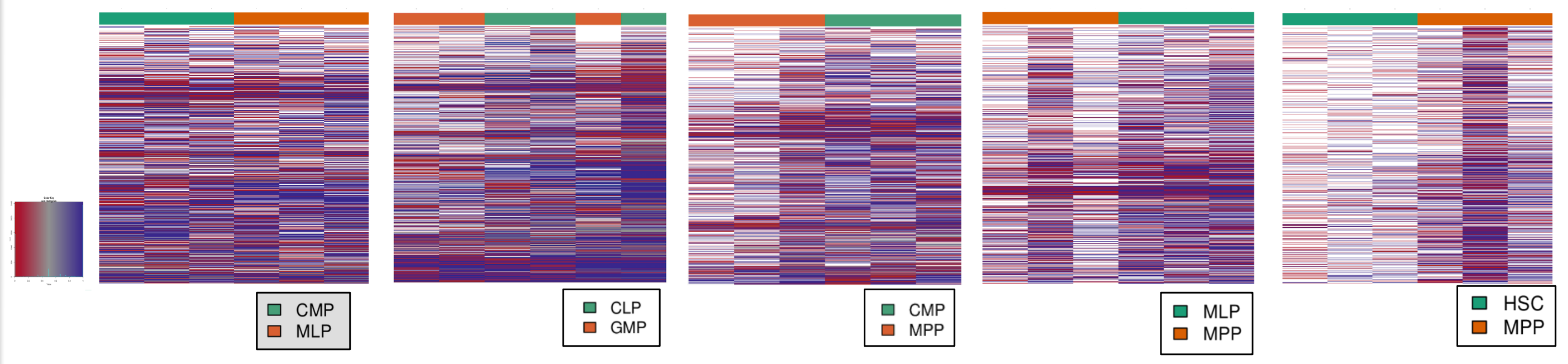
The regional methylation profiles were congruent in cell types CMP and MLP for an entire gene region, which illustrates that methylation is low at the transcription start site (TSS) and high in the latter part of the gene.

RnBeads data import and quality control

- The beta value is a ratio of the number of methylated sites to the total number of sites (methylated + unmethylated). It tells us about the proportion of methylation sites in our genomic regions.
- On filtering, we obtained the following beta value densities. Most beta value densities were retained.



Above is a CMP–MLP PCA analysis. Two promoters are shown on the left and one of the enhancer annotations used from ENSEMBL is on the right. There is distinct clustering of the samples, but the PC2 calculated for cell type CMP seems to have high variance in the data.



Hierarchical clustering of samples based on methylation values in enhancer regions. The heatmap displays only selected sites/regions with the highest variance across all samples.

Defining promoter and enhancer regions

- Promoters: promoter annotation is a built-in region.type option in RnBeads. A promoter is defined by Ensembl gene definitions as the region spanning 1,500 bases upstream and 500 bases downstream of the TSS.
- Enhancers: enhancer annotation tracks are downloaded from UCSC Table Browser and added to RnBeads as custom annotations. Enhancer definition is highly cell type-specific, so we loaded two enhancer datasets (Genome segments-ChromHMM) defined in the closest cell types to hematopoietic progenitor populations that were available. One is from K562, a myeloid leukemia cell line, and the other is GM12878, a B-lymphoblastoid cell line, representing features of myeloid and lymphoid lineages, respectively.

Differential methylation analysis

Pairwise cell types (for promoters)	No. of differentially methylated genes
CMP–MLP	162
GMP–CLP	80
MPP–CMP	313
MPP–MLP	303
HSC–MPP	492

Number of differentially methylated genes with a cut-off of difference of 0.4 and FDR 0.1

We filtered to obtain genes which had a high significance (low false discovery rate (FDR)) and a high differential methylation between the two cell types to help us understand how the different lineages were divided.

Conclusion: There seem to be a large number of genes which have a drastic difference in the DNA methylation in cell types that are directly related, but a small population of genes for cell types were cross-branch.

Correlation between RNA-seq data and DNA methylation data
Low methylation list: genes with relatively low methylation on promoters in one population
High expression list: genes with >2-fold higher expression level in one population

Is there positive correlation (high proportion of overlapped genes) between high expression and low methylation on the promoter?

HSC–MPP		
overlap (% meth list)	High_exp_HSC	High_exp_MPP
Low_meth_HSC	67 (**17%**)	88 (**23%**)
Low_meth_MPP	18 (**15%**)	28 (**23%**)

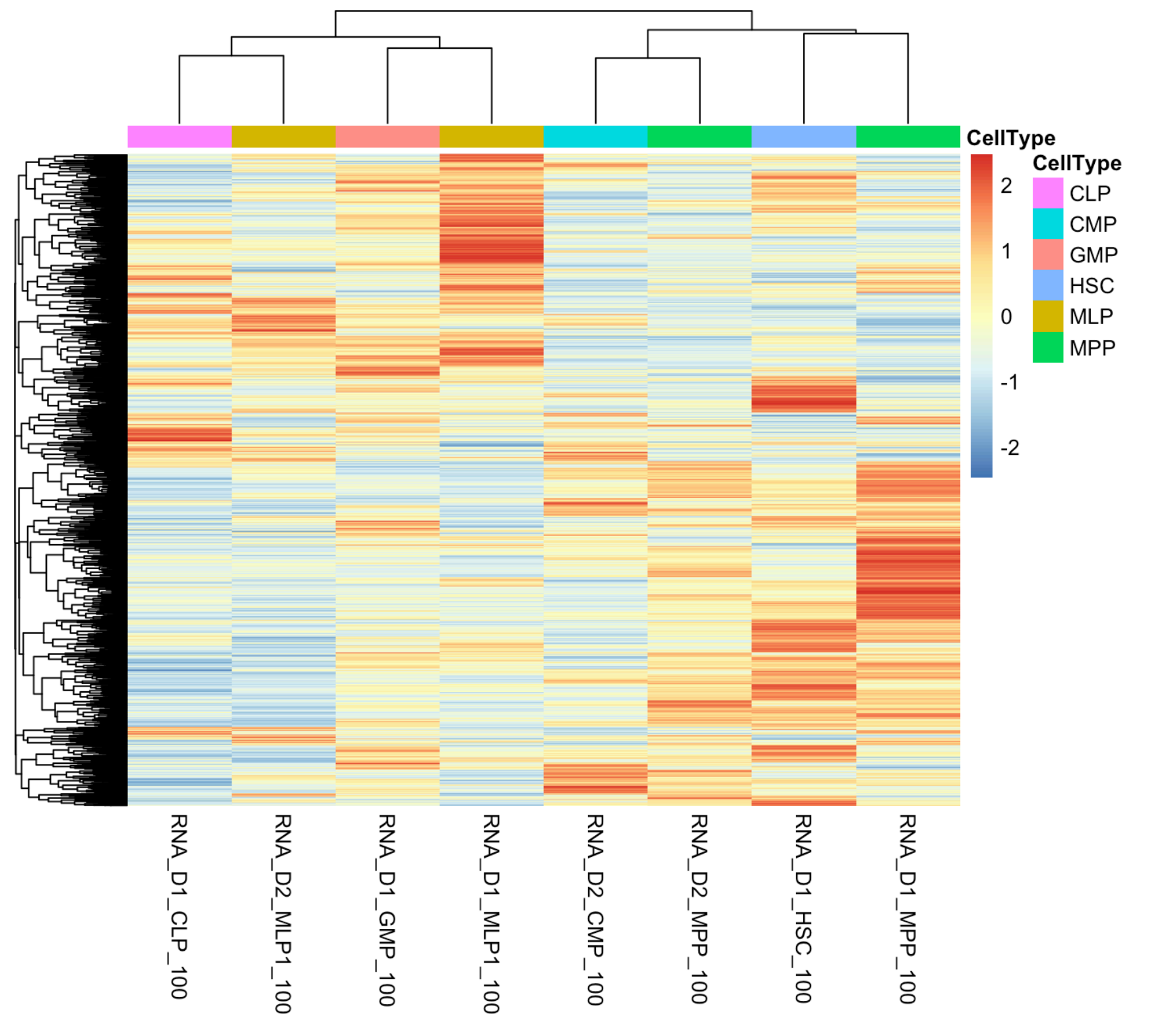
CMP–MLP		
overlap (% meth list)	High_exp_CMP	High_exp_MLP
Low_meth_CMP	10 (**11%**)	18 (**20%**)
Low_meth_MLP	13 (**18%**)	13 (**18%**)

There is no clear correlation from analyzing the overlap between the list of genes with high expression and the list of genes with low methylation promoter status.

2. RNA-seq processing to find genes with >2-fold changes

Data inspection

Filtering: removed one sample outlier, removed 2178 low-count transcripts from all 13 samples
Normalization on library size by DEseq "sizefactor" function



Clustering based on top 2000 genes

- Most primitive cells HSC and MPP are clustered together
- Lymphoid lineages (CLP and MLP) are clustered together
- GMP is intermixed in the lymphoid group, consistent with recent evidence

Generated lists of genes with >2-fold change in each comparison of six progenitor populations

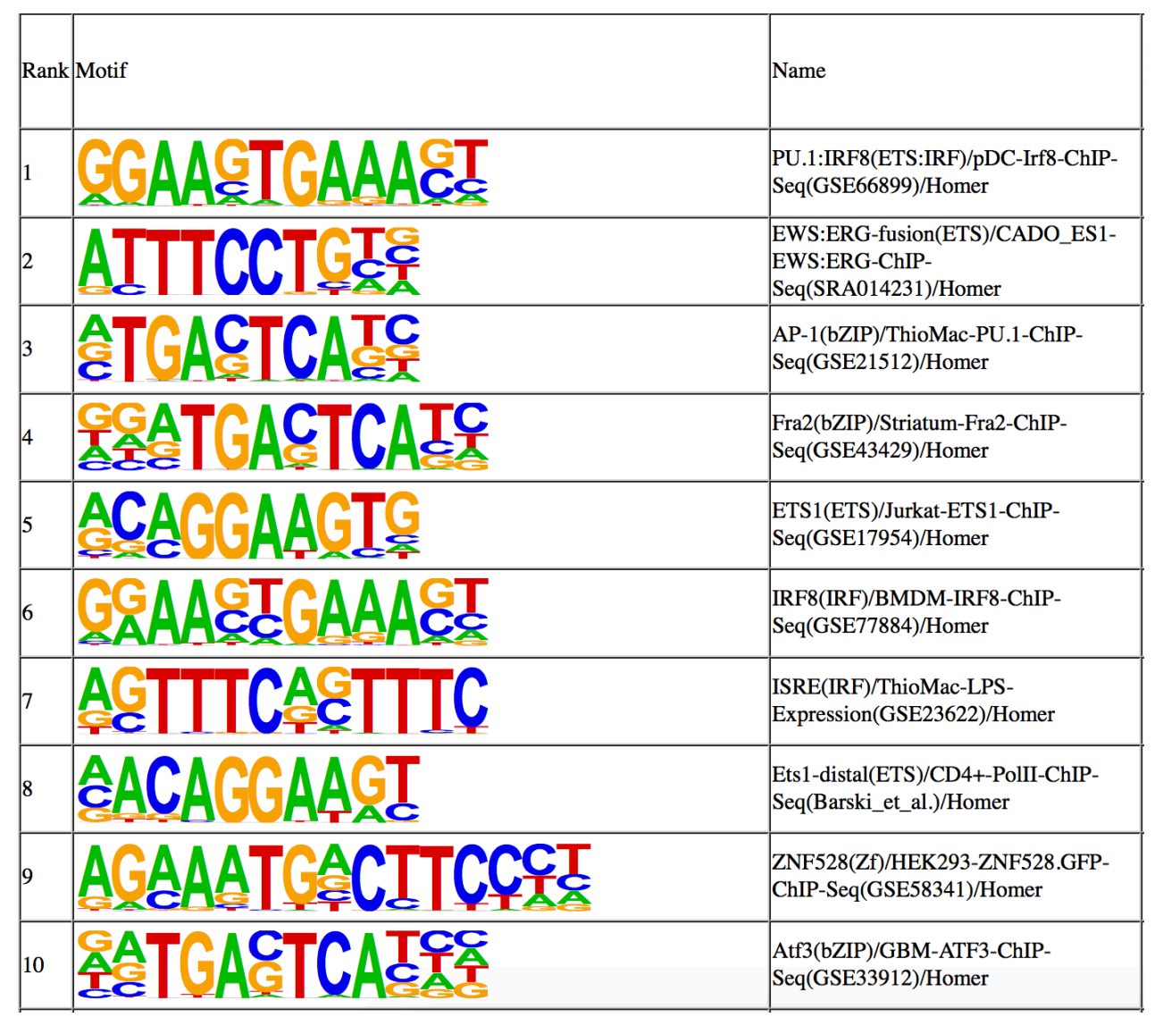
3. Finding TF motifs in differentially methylated regions (promoters and enhancers)

Input: for each pair, methylation analysis created a pair of files indicating hypomethylated regions in each cell
Output: enriched sequences likely to have biological significance
Regulatory regions are not direction-sensitive, so we made a positive and a negative instance of each region and ran HOMER (Hypergeometric Optimization of Motif EnRichment) — a suite of motif discovery tools — with hg19 reference genome and used the knownResults output.

Caveats

- No TF database is reliable since TFs have a high level of redundancy and regions of nonspecific binding
- Owing to the finite amount of data and many degrees of freedom in a motif probability matrix, it is easy to find a motif with a seemingly significant p-value
- Most of the returned results from promoter sequences had very low p-values

Top 10 Homer known motif results found for CMP when compared against MLP



Rank	Motif	Name
1	GGAAGTGAAAGT	PU.1:IRF8(ETS:IRF)/pDC-Irf8-ChIP-Seq(GSE66899)/Homer
2	ATTTCCTGTG	EWS:ERG-fusion(ETS)/CADO_ES1-EWS:ERG-ChIP-Seq(SRA014231)/Homer
3	ATGACTCATC	AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer
4	GGATGACTCATC	Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer
5	ACAGGAAGTG	ETS1(ETS)/Jurkat-ETS1-ChIP-Seq(GSE17954)/Homer
6	GGAAGTGAAAGT	IRF8(IRF)/BMDM-IRF8-ChIP-Seq(GSE77884)/Homer
7	AGTTTCAGTTTC	ISRE(IRF)/ThioMac-LPS-Expression(GSE23622)/Homer
8	AACAGGAAGT	Ets1-distal(ETS)/CD4+-PolII-ChIP-Seq(Barski_et_al.)/Homer
9	AGAAATGACTTCCGT	ZNF528(Zf)/HEK293-ZNF528.GFP-ChIP-Seq(GSE58341)/Homer
10	GATGACTCATCC	Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer

Interesting TFs with known functions in HSCs and differentiation found in enhancer regions
myeloid-/lymphoid-/self-renewal-associated TFs highlighted

HSC–MPP
Active in HSC: GATA1, RUNX1, GATA2, RUNX2, FLI1, PU.1, OCT4, GATA4
Active in MPP: GATA1, RUNX1, GATA2, GATA3, HOXB4, PAX5

CMP–MLP
Active in CMP: PU.1, HOXA9, FLI1, GATA3, CEBP, STAT1
Active in MPP: STAT3, RUNX1, MYC

GMP–CLP
Active in GMP: PAX5, HOXB4, CEBP
Active in CLP: STAT1, STAT5, HOXB13, STAT3, RUNX2, PRDM1, HLF

STAT signaling transducer/TFs are enriched in lymphoid lineages (MLP, CLP) compared to myeloid lineages

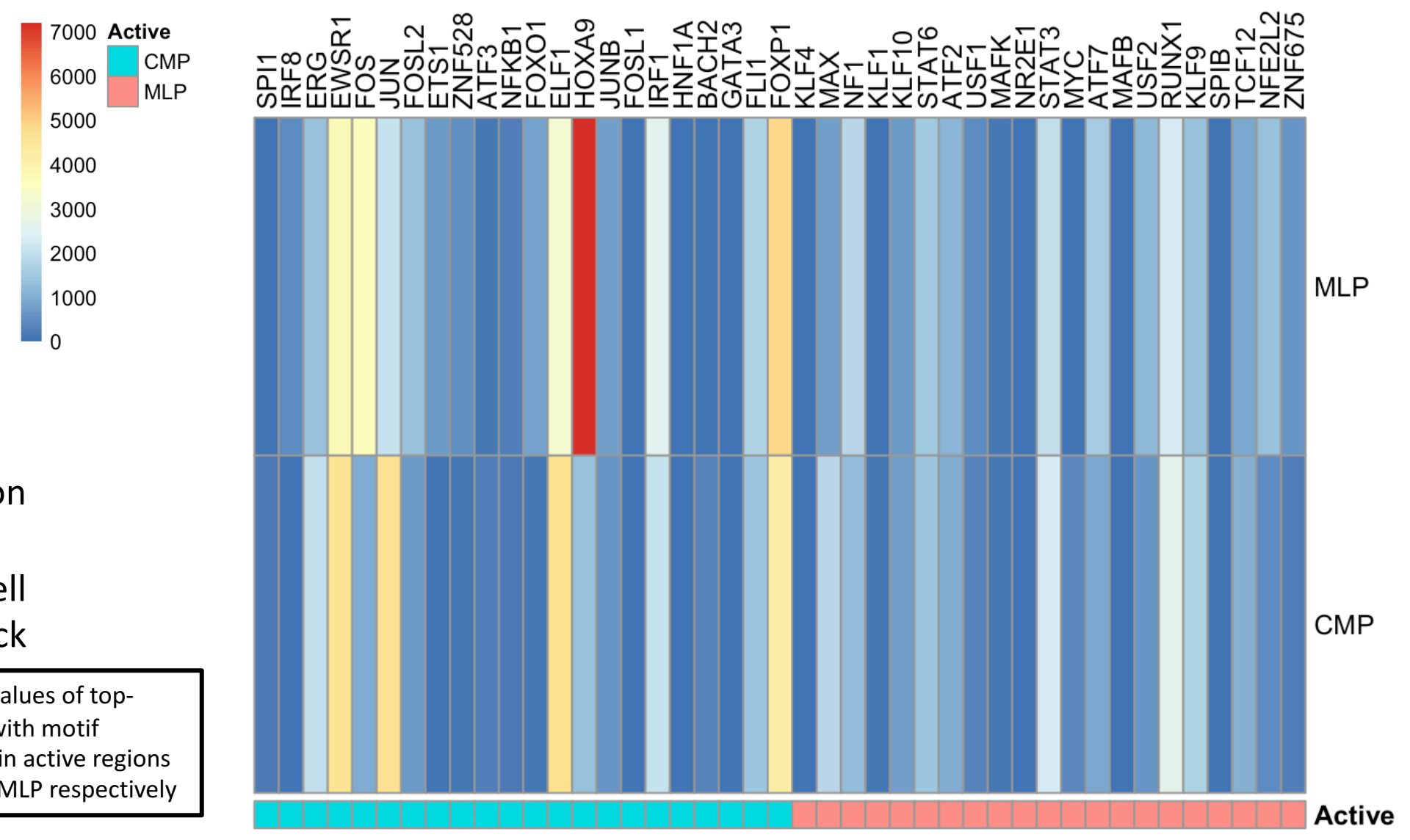
4. Inspecting the expression of TFs identified from motif-finding analysis

Are TFs enriched in the active regions highly expressed in corresponding cell types?

- Match TF names with ENSG gene symbol names
- Extract expression data from normal RNA-seq data
- TFs are identified in the enhancers with differential methylation from the myeloid–lymphoid comparison (CMP–MLP)

Comments

- Different sets of TFs are identified in the enhancers with differential methylation from the CMP–MLP comparison
- The cell type-specific TFs do not show clear pattern of high expression in the cell type where they are supposed to play an active role. However, owing to the lack of replicates, we cannot draw concrete conclusions.



Expression values of top-ranked TFs with motif enrichment in active regions in CMP and MLP respectively

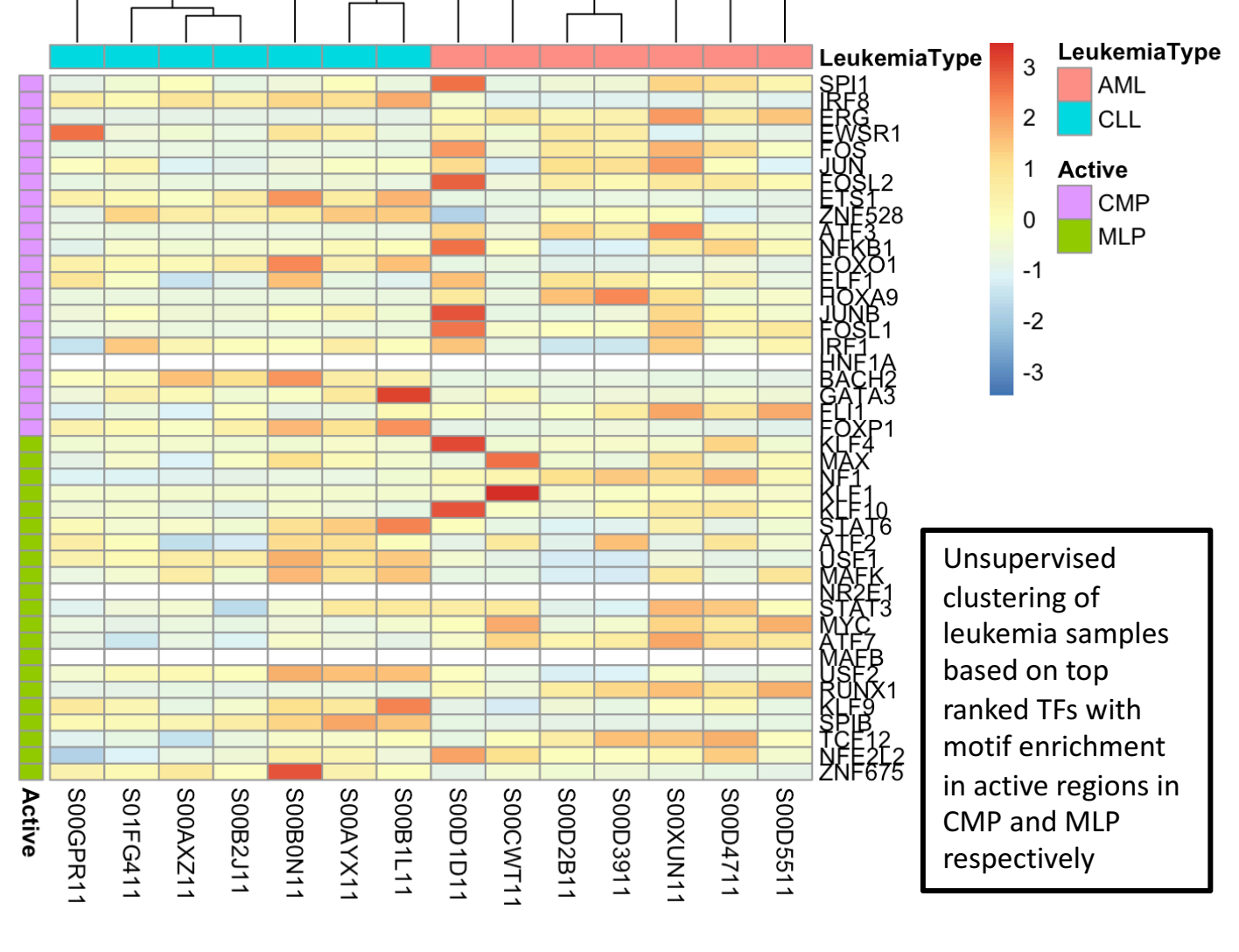
5. Clustering of two types of leukemia patient samples based on recognized TFs

Normalization and differential gene expression (not shown) were conducted using limma.

- Inspect the expression of TFs generated on enhancers from myeloid–lymphoid progenitor comparison (CMP–MLP) of differential DNA methylation
- Are TFs identified from CMP and MLP active regions able to cluster leukemia samples based on molecular subtype (AML versus CLL)?
- Do AML and CLL retain a signature of myeloid- and lymphoid-associated genes that reflects the cell of origin?
- Is this signature reflected by the TFs we identified from normal myeloid and lymphoid progenitors?

Comments

- Leukemia samples that belong to the same cancer type are clustered together using expression values of TFs identified from CMP/MLP active regions
- TFs predicted to be active in the CMP population show higher expression in AML samples, consistent with a myeloid cell program.
- TFs predicted to be active in the MLP population show slight tendency of higher expression in CLL samples, indicating a lymphoid cell program.



Unsupervised clustering of leukemia samples based on top ranked TFs with motif enrichment in active regions in CMP and MLP respectively

CONCLUSIONS

- There are differences in DNA methylation profiles based on cell types between hematopoietic progenitor populations
- DNA methylation on promoter regions shows no clear correlation with the level of gene expression, possibly due to the poor quality of the methylation data (low coverage) and RNA-seq data (replicates not available for some populations)
- *In silico* TF enrichment analysis in differentially methylated loci has identified TFs known to have a particular role in population differentiation
- The speculated active TF status is correlated with the expression level of the TF in the population in the leukemia samples; leukemia samples that belong to the same cancer type are clustered together using expression values of TFs identified from CMP and MLP active regions.

REFERENCES

1. Miller, P.H., Knapp, D.J., Eaves, C.J. (2013). Heterogeneity in hematopoietic stem cell populations implications for transplantation. Curr. Opin. Hematol. 20(4):257-64

2. Notta, F., Zandi, S., Takayama, N., Dobson, S., Gan, O.I., Wilson, G., Kaufmann, K.B., McLeod, J., Laurenti, E., Dunant, C.F., McPherson, J.D., Stein, L.D., Dror, Y., Dick, J.E. (2016). Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science 351(6269):aab2116

3. Göttgens, B. (2015). Regulatory network control of blood stem cells. Blood 125(17):2614-20

4. Farlik, M., Halbritter, F., Müller, F., Choudry ,F.A., Ebert, P., Klughammer, J., Farrow, S., Santoro, A., Ciaurro, V., Mathur, A., Uppal, R., Stunnenberg, H.G., Ouwehand, W.H., Laurenti, E., Lengauer, T., Frontini, M., Bock, C. (2016). DNA methylation dynamics of human hematopoietic stem cell differentiation. Cell Stem Cell 19(6):808-822

5. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38(4):576-589