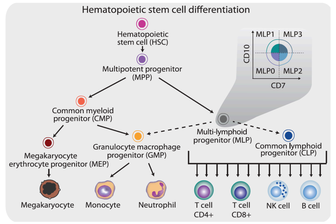
Data-driven analysis of the potential candidate transcription factors in hematopoietic stem cell differentiation into multiple progenitor compartments

Wang, Fangwu; Hoque, Rawnak; Paul, Somdeb; Cavalla, Anna
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 the interconversion 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]



OBJECTIVES

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

Hypothesis:

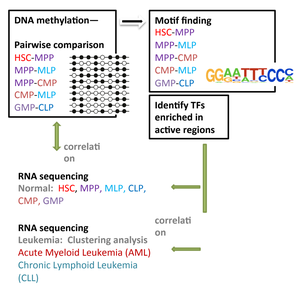
I. DNA methylation status reflects cell identity of closely related progenitor populations during differentiation.

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 from 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 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.



RESULTS

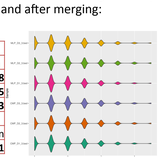
1. DNA methylation analysis on promoters and enhancers

Generation of merged data from technical replicates

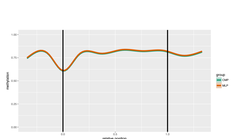
To increase the overall coverage, we added up the reads from technical replicates (3 aliquots with 50, 50, 1000 cells from the same donor condition) using bedtools which resulted in 3 merged data (3 biological replicates) for each cell type.

Comparison of mean coverage before and after merging:

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



Coverage distribution across cell types

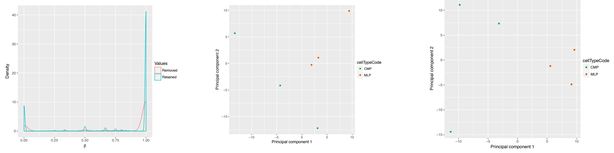


Methylation level distribution across “Gene” region type

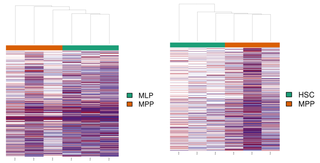
The regional methylation profiles in both cell types CMP and MLP were congruent for an entire gene region which illustrates that methylation is low in the transcription start site (TSS) whereas high in the later part of the gene.

RnBeads data import and quality control

- The beta value is a ratio of the number of methylation 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. There were far fewer beta value densities that were removed from analysis than those that were retained.



Above is a PCA analysis between two cell types (CMP–MLP) using the principal components PC1 and PC2 (for promoters on the left and one of the enhancer annotations used from ENSEMBL, on the right). There is distinct clustering among the samples, but the PC2 calculated for CMP seems to have a lot of variance in the data.



Hierarchical clustering of samples based on methylation values on enhancer regions. The heatmaps display 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. Under Ensembl gene definitions, a promoter is defined as the region spanning 1,500 bases upstream and 500 bases downstream of the transcription start site.
- 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 we can find to hematopoietic progenitor populations. 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

Number of differentially methylated regions (DMR) with a cut-off of difference of 0.4 and FDR of 0.1.

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

Reason behind the threshold values?

We targeted those genes which had a good significance (low false discovery rate) and wanted genes with a high differential methylation between the two cell types so as to understand how the difference in methylation would help in the expression of the genes in the pairs of cell types, which would aid in understanding how the different lineages were divided.

Conclusion: There seems 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 that were cross branch. The correlation with gene expression follows in the next section.

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 (more proportion of overlapped genes) between high expression and low methylation on the promoter?

HSC-MPP comparison		
overlap (% meth list)	High_exp_HSC	High_exp_MPP
Low_meth_HSC	82 (**22%**)	94 (**25%**)
Low_meth_MPP	23 (**19%**)	29 (**24%**)

CMP-MLP comparison		
overlap (% meth list)	High_exp_CMP	High_exp_MLP
Low_meth_CMP	11 (**12%**)	22 (**24%**)
Low_meth_MLP	15 (**21%**)	14 (**19%**)

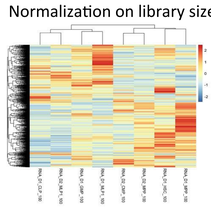
There is no clear correlation from analyzing the overlap between the list of genes with high expression and the list of genes with low methylation promoter status.

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 top2000 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.

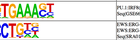
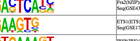
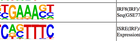
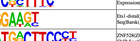
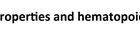
Generate lists of genes with 2-fold change in each comparison of 6 progenitor populations.

Sample clustering based on 2000 most abundantly expressed transcripts

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

For each comparison, our methylation analysis created a pair of files indicating hypomethylated regions in each cell type in comparison to its partner, and thus which promoter regions were more accessible and likely to be in use. Since regulatory regions are not direction-sensitive, 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 the hg19 reference genome and used the knownResults output.

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. Therefore motifs should be ignored when the results start becoming very different from one another (in terms of sequence) yet have similar p-values. In addition, high quality motifs usually appear multiple times in the list with different offsets. CLP cells, in comparison to GMP cells, had the motif GCATCGGA. However, most of the results we obtained had very low p-values.

Rank	Motif	Name
1		PU.1:IRF8(ETS:IRF)/pDC-Irf8-ChIP-Seq|GSE66499|Homer
2		EWS:ERG-fusion(ETS)/CADO_ES1-EWS:ERG-ChIP-Seq|SRA014231|Homer
3		AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq|GSE21512|Homer
4		Fra2(bZIP)/Striatum-Fra2-ChIP-Seq|GSE43429|Homer
5		ETS1(ETS)/Jurkat-ETS1-ChIP-Seq|GSE17954|Homer
6		IRF8(IRF)/BMDM-IRF8-ChIP-Seq|GSE77884|Homer
7		ISRE(IRF)/ThioMac-LPS-Expression|GSE23622|Homer
8		Ets1-distal(ETS)/CD4+-Ets1-ChIP-Seq|Banchereau_et_al.|Homer
9		ZNF526(ZF)/HEK293-ZNF526.GFP-ChIP-Seq|GSE30341|Homer
10		Atf3(bZIP)/GBM-ATF3-ChIP-Seq|GSE33912|Homer

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

Interesting TFs with known functions in HSC stem cell properties and hematopoietic cell differentiation recognized from enhancer regions

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

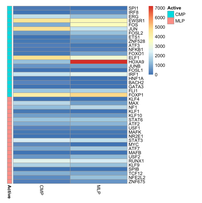
(**myeloid**/**lymphoid**/**self-renewal** associated TFs were highlighted in different colors)

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

3. Inspect the expression of TFs identified from motif finding analysis

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

- Match TFs 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 a clear pattern of high expression in the cell type where they are supposed to play an active role. However, due to the poor quality of data (no replicates), we cannot draw robust conclusions from this observation.

Expression values of 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 within 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 (low coverage) and RNA-seq data (replicate not available for some populations).
- *In silico* findings of transcription factor enrichment in differential methylation loci have identified TFs known to have a particular role in the differentiation of the population.
- 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/MLP active regions.

REFERENCES

1. Miller PH, Knapp DJ, Eaves CJ. Heterogeneity in hematopoietic stem cell populations implications for transplantation. Curr Opin Hematol. 2013 Jul;20(4):257-64

2. Notta F, Zandi S, Takayama N, et al, Distinct routes of lineage development reshape the human blood hierarchy across ontogeny.Science. 2016 Jan 8;351(6269):aab2116.

3. Göttgens B. Regulatory Network Control of Blood Stem Cells. Blood. 2015 Apr 23;125(17):2614-20.

4. Farlik M, Halbritter F, Müller F, et al. DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. Cell Stem Cell. 2016 Dec 1;19(6):808-822

5. Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589. PMID: 20513432