Title: An integrative view of the regulatory and transcriptional landscapes in mouse myeloid hematopoiesis

Old title: Transcriptomics and chromatin landscapes in hematopoietic stem and myeloid lineages

Authors (number and order is NOT settled):

Cheryl A. Keller

Elisabeth Heuston

Belinda Giardine

Guanjue Xiang

Lin An

Amber Miller

Berthold Gottgens

Feng Yue

David Bodine

Yu Zhang

Ross C. Hardison

Cheryl Keller <cak142@psu.edu>, "Heuston, Elisabeth (NIH/NHGRI) [F]" <elisabeth.heuston@nih.gov>, "Belinda M. Giardine" <giardine@bx.psu.edu>, Guanjue Xiang <gzx103@psu.edu>, Lin An <anlin00008@gmail.com>, Amber Miller <aer104@psu.edu>, Bertie Gottgens <bg200@cam.ac.uk>, fyue@HMC.PSU.EDU, "Bodine, David (NIH/NHGRI) [E]" <tedyaz@nhgri.nih.gov>, Yu Zhang <yuzhang@stat.psu.edu>, Ross Hardison <ross@bx.psu.edu>

**Major points**

- New datasets are described for ATAC-seq and RNA-seq in stem, multilineage, and maturing myeloid hematopoietic cells from mouse bone marrow.

- Chromatin landscapes and transcriptome profiles reveal similar relationships among hematopoietic stem cells (HSCs), multilineage myeloid progenitors (CMP, GMP, and MEP), and maturing monolineage cells (CFU-Mk, Mk, CFU-E, ERY).

- HSCs and multilineage progenitors are very similar in their global chromatin landscape and transcriptome profiles. Larger changes are seen during maturation after commitment.

- An integrative view of the ATAC-seq and histone modification profiles was determined using the 2-D genome segmentation approach IDEASs. This provides a compact and concise display of function-associated states along the chromosomes of each cell type.

- ATAC-seq peaks have been annotated in multiple ways to provide insights into roles of candidate regulatory elements (cREs).

(a) One annotation approach assigns each ATAC-seq peak to a chromatin state determined by IDEAS.

(b) Another shows overlap with transcription factor occupancy data across mouse hematopoietic cell types, and another tracks the the history of appearance and loss of each peak region in each cell type.

These annotations constitute a versatile and comprehensive collection of cREs that can be used in many ways in further studies of gene regulation in hematopoiesis.

- The RNA-seq data obtained from the same cell populations are valuable for many studies. Signal tracks and estimates of transcript levels are provided in multiple formats, from browsers to graphical displays.

- The correlation of transcript levels for each gene and signal strengths for ATAC-seq peaks, varying across cell types, was used to assign cREs to likely target genes, providing a comprehensive set of locus models for regulation across myelopoiesis. These assignments were evaluated and refined by comparisons to TADs based on Hi-C data.

- The data on transcript levels across cell types was also used to evaluate the efficacy of current epigenomic data to explain levels of gene expression.

- Can we put the RNA-seq data on the differentiation axes defined by single cell transcriptomics?

**Datasets used**

- ATAC-seq data on a purified population containing hematopoietic stem cells (LSKs), multilineage myeloid progenitors (CMP, GMP, and MEP), and maturing monolineage cells (CFU-Mk, Mk, CFU-E, ERY, GRN, MONO), all isolated by FACS from adult mouse bone marrow (Hardison and Bodine ENCODE3 data)

- RNA-seq data on the same cell types, determined as ribo-depleted total RNA by two different protocols: TotalScript and ScriptSeq (all cell types in both protocols) (Hardison and Bodine ENCODE3 data)

- Histone modification data from Amit lab (iChIP data, published in Science 2014)

- Histone modification data from Bodine lab

- new CTCF ChIP-seq data in HSCs (and other progenitors?)

- Two-dimensional (along chromosomes and across cell types) segmentations into epigenetic states using IDEAS (Zhang et al NAR 2016, 2017).

**Introduction**

The production of many distinct blood cell types from a common stem cell (hematopoiesis) is critically important in human health. It has been studied intensively in humans and mouse, the latter serving as a good model for many aspects of hematopoiesis. The classic model for lineage choice was inferred from detailed, elegant studies of the capacity of marker-purified populations to form mature cells of the various hematopoietic lineages. More recent analyses of single cells have revealed heterogeneity in each of these populations, and in some cases they uncovered a strong bias in multilineage progenitors toward a single cell type (e.g. MEPs biases toward ERY). While many details remain to be determined, all the work strongly supports the key role of gene regulation in carrying out this differentiation program.

As studies continue into the relationships and conversions among the various progenitor cells, comprehensive epigenomic and transcriptomic data can be used to provide important resources for better understanding of the event occurring in – and regulating - hematopoietic differentiation. Considerable resources are already available (Lara-Anastasio et al Amit lab 2014; Sherm Weissman’s lab 2016 for mouse; Buenrostro, Green etc 2016 for human). However some important lineages are not well represented, such as megakaryocytes (MK), and some of the previous data has light coverage of the genome. We have generated new RNA-seq data to provide more complete coverage of the transcriptomes and nuclease accessibility (ATAC-seq) to complement previous studies and provide a more complete examination of the myeloid lineages, emphasizing MK and erythroid lineages. We apply a new integrative method (IDEAS) to more precisely define chromatin states based on multiple epigenetic features across a myelo-erythroid series starting with hematopoietic stem cells and extending to MK and erythroblasts (ERY).

By combining the new data with recently determined maps of transcription factor (TF) occupancy and chromatin interaction frequencies, we integrated all this information into specific predictions of cREs, gene targets, and views of the activity landscapes of epigenomes at a range of scales, leading to important insights into the ability of epigenomic features to explain differences in gene expression.

**Results**

**New data (Fig. 1)**

Describe data, quality, and replicate concordance

RNA-seq from total RNA

ATAC-seq: describe how the collection of 211,815 replicated peaks was made (only replicated peaks within a cell type or population are retained *if* a replicate is available). Note that this is smaller than the set of 260,868 peaks that were used earlier (found in at least two cell type replicates – not necessarily the same cell type – if replicate is available).

New histone modification (H3K27ac, H3K4me1) data in ERY and MK (Hueston et al.), plus already published data (H3K27ac) in G1E and G1E-ER4+E2.

More data are being mapped onto mm10:

BRD2, BRD3, BRD4 in G1E and G1E-ER4

RAD21 in G1E-ER4

POL2 in G1E and G1E-ER4

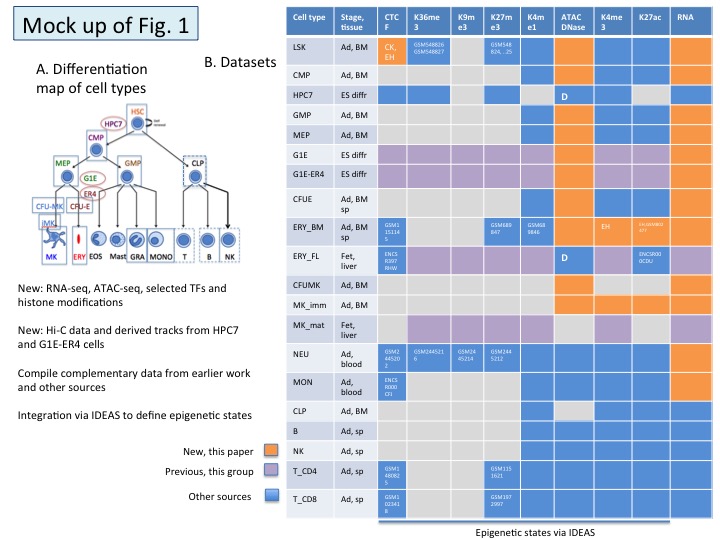
CTCF in LSK, HPC7, NEU, T\_CD4, T\_CD8

(see Excel file BRDs\_RAD21\_POL2\_etc\_for\_mm10.xlsx)

Hi-C related tracks

Hi-C data are released with previous publications, but we provide informative tracks based on those data.

HPC7 cells, G!E-ER4 cells



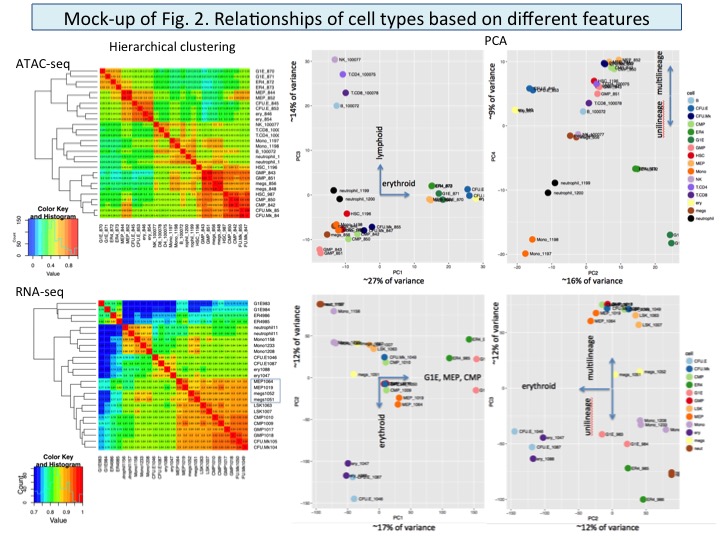
**Similar relationships among cell types are observed based on RNA-seq, ATAC-seq, and chromatin state profiles (Fig. 2)**

- Chromatin landscapes and transcriptome profiles reveal similar relationships among LSKs (which include hematopoietic stem cells (HSCs) and MPPs), multilineage myeloid progenitors (CMP, GMP, and MEP), and maturing monolineage cells (CFU-Mk, Mk, CFU-E, ERY).

- HSCs and multilineage progenitors are very similar in their global chromatin landscape and transcriptome profiles.

- Guanjue Xiang and Yu Zhang estimated distances between cell types using the IDEAS segmentation as the input data (using an earlier segmentation result on only 4 marks). This could be another, complementary approach that brings out the benefit of integration

- We need to do the clustering and annotations on the same set of ATAC-seq peaks. I suggest the 211,815 replicated peaks.



**Integrative analysis via IDEAS (Fig. 3)**

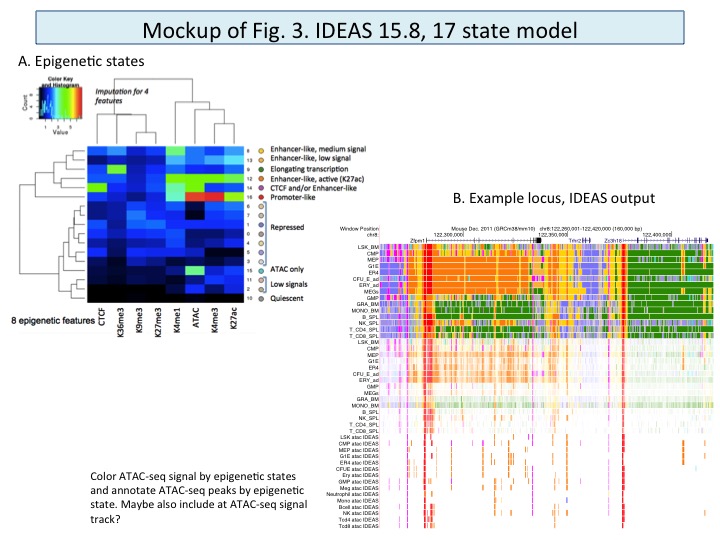
We applied the novel Integrative and Discriminative Epigenome Annotation System (IDEAS) to find chromatin states in each cell type; these chromatin states correspond to coherent sets of the features such as DNA accessibility and histone modifications. IDEAS performs a two-dimenational segmentation along chromosomes and across cell types, thereby maintaining information about features at specific genomic positions and local similarities in the epigenomes of different cell types. The resulting segmentation has high consistency across cell types, thereby increasing its value for comparisons and discrimination across cell ypes. IDEAS also uses the full range of signal strength for each feature to avoid the limitations of binarization during the segmentation.

Currently using ssegmentation 15.8 (15 cell types, 8 features, four with extensive imputation) on mm10.

We will do this on a more complete set of data once the data are on mm10.

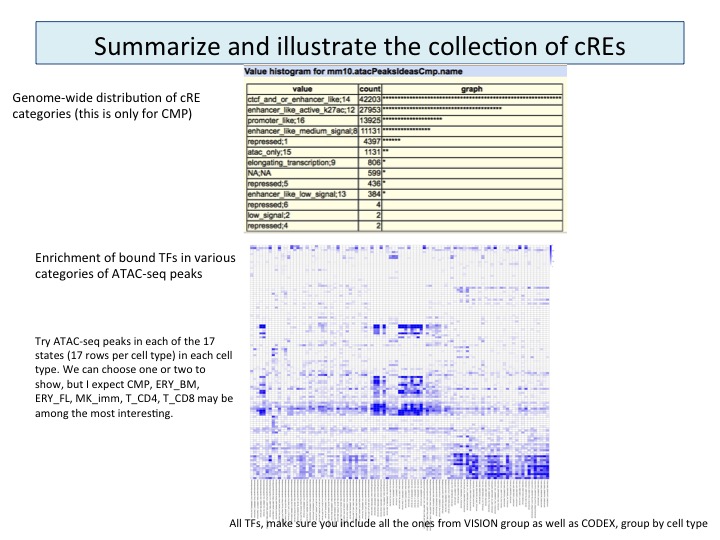
We are examining normalization using NISC prior to input into IDEAS.

ATAC-seq peaks and signal tracks colored by epigenetic state.



**Annotate ATAC-seq peaks by chromatin states and overlap with TFs (Fig. 4)**

Almost all REs are accessible to nucleases in the nuclei of cells in which they are active, but not all nuclease sensitive sites are REs. We used the extensive information on epigenetic states and TF occupancy to determine which ATAC-seq peaks are candidate REs (cREs), and possible roles for each. This provides the most extensive collection of cREs in mouse hematopoietic cells to date (true?). It encompasses \_\_% of the cREs from Lara-Anastasio and \_\_% of our curated collection of erythroid regulatory elements.



*Calculation of the enrichment of a transcription factor (TF) ChIP-seq peak of a specific cell type in a certain index set*

CODEX database: 97 samples, 10 cell types

TF ChIP-seq peak Enrichment = ( foreground + 10 ) / ( background + 10 )

Foreground = The number of TF ChIP-seq peaks that intersect with the DNA intervals of an index set (1059 peaks in the Index set in previous example)

Background = The number of TF ChIP-seq peaks we would expect to intersect with the DNA intervals of a Index set if the TF ChIP-seq peak is randomly distributed throughout all 153,636 DNA intervals

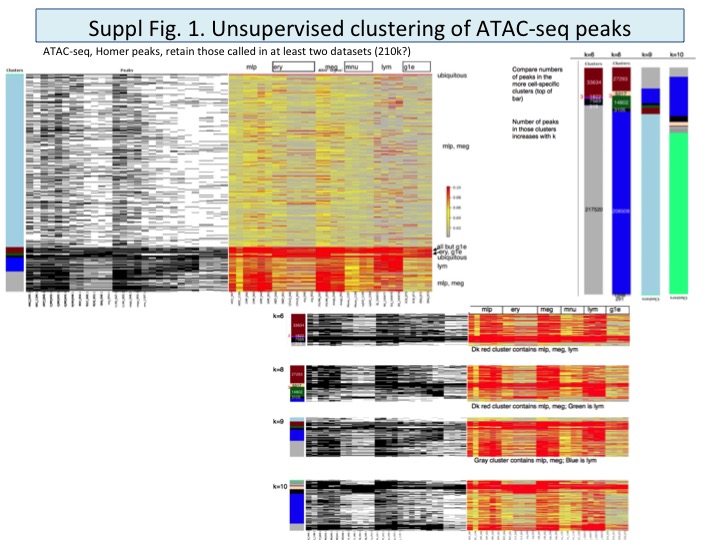
Plus 10 is to increase robustness.

example 1: if Foreground=5, background=1, enrichment=5 (without plus 10)

example 2: if Foreground=5, background=1, enrichment=1.36 (with plus 10)

**Group ATAC-seq peaks by occurence across all cell types (Fig. 5)**

The data on occurrence of ATAC-seq peaks across multiple lineages also provides information on their potential roles in a variety of biological contexts. We explored conventional clustering approaches and also developed a simple but informative and flexible indexing approach to capture the occurrence and history of ATAC-seq peaks.



*Indexing approach*:

Filter ATAC-seq peaks to retain those called in replicates

211,815 out of original 260,868

Apply rule when replicates available – Amit peaks are single determination

For each DNA interval in the peak set, assign a 1 if it has been called as a peak in a given cell type, 0 if otherwise

This gives a vector with 16 entries, one for each cell type

Format: 1\_1\_0\_1\_0\_0\_0\_1\_1\_0\_0\_0\_1\_1\_0\_0

This index can be considered a cell-type code

All DNA intervals with the same index are grouped together

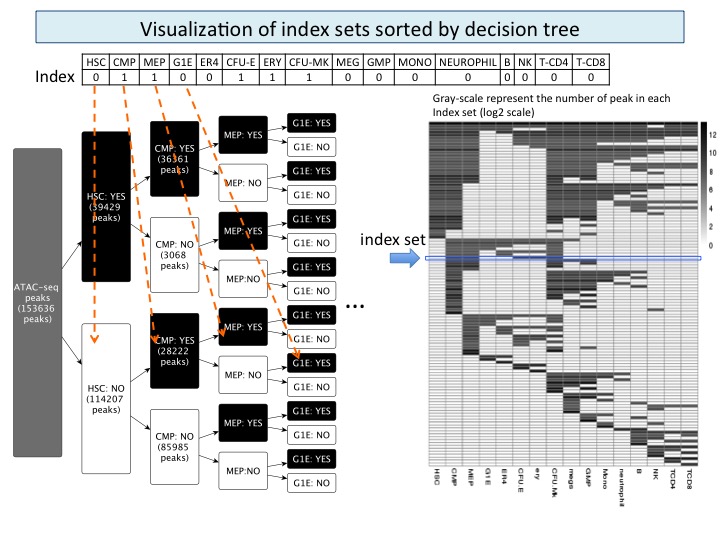
Refer to these as an index set

Distinctive features

Not generate by a clustering algorithm

Assignments is independent of the order in which cell types are considered

Retain an explicit connection between the index set and occurrence in a cell type



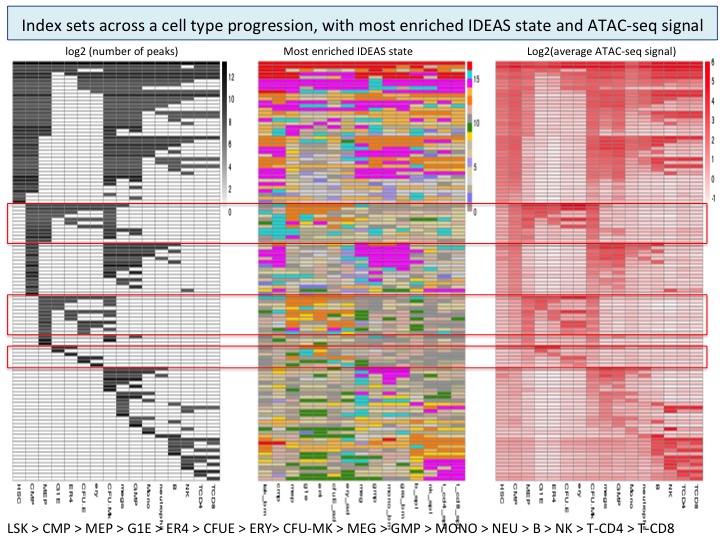
*Functional correlations:*

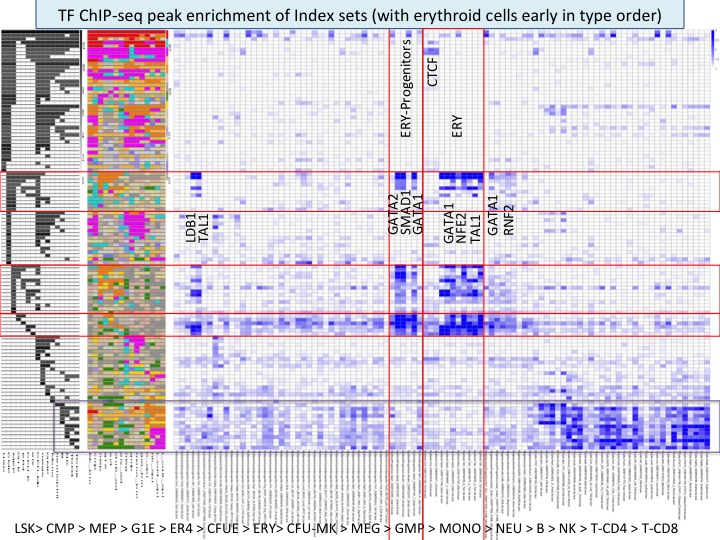
Are ubiquitous peaks enriched for promoters? For what categories of genes? widely expressed?

Cell-type restricted (in a few related cell types)

Cell-type specific (only one cell type – do we ever see this?)

Are there groups of cells within which we see a substantial fraction of cREs that are only in that group?





**Assign cREs to target genes (Fig. 6)**

- Use correlations across cell types between occurrence of ATAC-seq peaks and expression of genes to assign each peak to all correlated genes, then filter for presence on the same chromosome, presence in the same TAD.

- Group all cREs in a given TAD with all expressed genes in that TAD. Maybe put a distance limit on them;

Compare the sets.

Any better appraoches?

*Need a figure*

**How well do the cREs explain expression across cell types? Or some other evaluation of how useful these integrated results on the data are. (Fig. 7)**

Possibilities:

Use analyses like Yu Zhang did for the IDEAS vs chromHMM comparison.

Regress expression level of each gene in each cell type on the ATAC-seq signal intensity of each cRE assigned to it. Can this be done in a way that incorporates the different epigenetic states from IDEAS?

*Need a figure*

**Views of landscapes at different scales** (*Is this too much for this paper, or is it a way to add novelty?*) **(Fig. 8?)**

At tens of Mb, you see lively zones and quiescent (low signal) zones, which correlate with A and B compartments. Quiescent zones are pretty constant across hematopoietic cell types (test this globally).

At about 1 Mb, you see regions enriched for TFs, BRDs, expression. Do these correspond to TADs? They are surrounded by quiescent zones.

The regions with known regulatory activity are about 100-200 kb (globin genes, etc). They have been called sub-TADs. Can we annotate them as well? Do we see a distinctive signal in the segmentation patterns? Has this been done genome-wide?

**Discussion**

These data are highly valuable.

Our work is novel in the diversity of data that is integrated (ATAC-seq, histone mods, TFs, Hi-C data) the approach to segmentation by epigenetic states (IDEAS), the methods prediction and annotation of cREs, ...

The expected shifts in chromatin landscapes and gene expression are observed, and they are particularly widespread in erythroid cells (Weintraub and Groudine were right - no surprise there). However, compared to the global patterns of chromatin landscapes and gene expression profiles, the lineage-specific changes are remarkably limited. Thus major cell fate decisions (other than going erythroid) are occurring in a context of global regulatory similarity. So what is really driving the fate decisions?

Discuss using our resources at different scales??

**Methods**

Datasets used

- ATAC-seq data on hematopoietic stem cells (HSCs), multilineage myeloid progenitors (CMP, GMP, and MEP), and maturing monolineage cells (CFU-Mk, Mk, CFU-E, ERY), all isolated by FACS from adult mouse bone marrow (Hardison and Bodine ENCODE3 data)

*ATAC-seq in replicates from VISION project*

Bodine lab and Hardison lab

Public but not yet published (ENCODE project)

Single determinations from Amit lab (Lara-Astiaso et al 2014 Science)

Uniform mapping and data processing in Hardison lab

Signal tracks were very similar for cell populations assayed in both VISION and Amit lab

Peak calls from Homer (shorter intervals; Fseq “peaks” ATAC-sensitive regions

Amit data used only for cell types not interrogated in VISION (mature lymphoid cells)

>100k peaks called per cell type and replicate

Union and merge to join peak intervals found in multiple cell datasets

Filter to retain peaks either

(1) found in at least two cell type replicates (not necessarily the same cell type) if replicates available or

(2) called in a cell type with no replicate available

Result: 260,868 peaks across the 16 cell types

Keep replicates separate

Peak calls (binary)

Signal strength (ATAC) in each cell type

- RNA-seq data on the same cell types, determined as ribo-depleted total RNA by two different protocols: TotalScript and ScriptSeq (all cell types in both protocols) (Hardison and Bodine ENCODE3 data)

- Histone modification data from Amit lab (iChIP data, published in Science 2014)

- Two-dimensional chromatin segmentations (along chromosomes and across cell types) using these data. The method is Yu Zhang’s IDEAS (published in NAR 2016).