frontiers in Genetics

ORIGINAL RESEARCH
published: 01 October 2021
doi: 10.3389/fgene.2021.707117

Tree-Based Co-Clustering Identifies Chromatin Accessibility Patterns Associated With Hematopoietic Lineage Structure

*Thomas B. George, Nathaniel K. Strawn and Sivan Leviyang**

Department of Mathematics and Statistics, Georgetown University, Washington, DC, United States

Chromatin accessibility, as measured by ATACseq, varies between hematopoietic cell types in different lineages of the hematopoietic differentiation tree, e.g. T cells vs. B cells, but methods that associate variation in chromatin accessibility to the lineage structure of the differentiation tree are lacking. Using an ATACseq dataset recently published by the ImmGen consortium, we construct associations between chromatin accessibility and hematopoietic cell types using a novel co-clustering approach that accounts for the structure of the hematopoietic, differentiation tree. Under a model in which all loci and cell types within a co-cluster have a shared accessibility state, we show that roughly 80% of cell type associated accessibility variation can be captured through 12 cell type clusters and 20 genomic locus clusters, with the cell type clusters reflecting coherent components of the differentiation tree. Using publicly available ChIPseq datasets, we show that our clustering reflects transcription factor binding patterns with implications for regulation across cell types. We show that traditional methods such as hierarchical and kmeans clusterings lead to cell type clusters that are more dispersed on the tree than our tree-based algorithm. We provide a python package, chromcocluster, that implements the algorithms presented.

Keywords: chromatin accessibility, hematopoiesis, clustering, tree (graphs), ATACseq, epigenetics

OPEN ACCESS

Edited by:
Marcelo R. S. Briones, Federal University of São Paulo, Brazil

Reviewed by:
Congting Ye, Xiamen University, China
Vladimir B. Teif, University of Essex, United Kingdom

****Correspondence:***
Sivan Leviyang
Sivan.Leviyang@georgetown.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *09 May 2021*
Accepted: *14 September 2021*
Published: *01 October 2021*

Citation:
George TB, Strawn NK and Leviyang S (2021) Tree-Based Co-Clustering Identifies Chromatin Accessibility Patterns Associated With Hematopoietic Lineage Structure. Front. Genet. 12:707117. doi: 10.3389/fgene.2021.707117

1 INTRODUCTION

The development of the ATACseq technique over the past decade has spurred a broad investigation of chromatin accessibility across cell types (Buenrostro et al., 2013; Klemm et al., 2019). In particular, chromatin accessibility has been intensively studied using ATACseq across many hematopoietic cells types (Lara-Astiaso et al., 2014; Corces et al., 2016; Scott-Browne et al., 2016; Lau et al., 2018; Calderon et al., 2019; Yoshida et al., 2019; Sun and Barreiro, 2020; Xiang et al., 2020). Hematopoiesis, which involves the differentiation of a single hematopoietic stem cell into the different blood cell types, is well characterized and the lineages through which the differentiation occurs can be described by a differentiation tree (Seita and Weissman, 2010). Chromatin accessibility has been shown to vary across different hematopoietic cell type lineages, and these differences have been shown to be essential to cell differentiation and cell function, e.g. (Heinz et al., 2010; Shea et al., 2010; Song et al., 2011; Huang et al., 2016; Lau et al., 2018; Sun and Barreiro, 2020).

While many studies have shown differences of accessibility across hematopoietic cell types, we lack a quantitative description of how variation in accessibility across the hematopoietic cell types

reflects the form of the differentiation tree. In this context, many questions remain unanswered. Are most accessible genomic loci accessible only in a specific, cell type or are a significant number of loci accessible jointly across a particular collections of cell types (e.g. all T cells)? If many loci are accessible across a particular collection of cell types, do those cell types form connected components of the differentiation tree, or are they dispersed? More generally, how do we quantitatively find and describe associations between chromatin accessibility and the differentiation tree? And finally, if such associations exist, how do they shape cellular regulation? Answering these questions would provide a context in which to analyze the chromatin accessibility of particular hematopoietic cell types and to understand differences in cellular regulation across the hematopoietic cell types.

Here, we address these questions using a recently completed ImmGen (Shay and Kang, 2013) consortium dataset published by Yoshida et al. (Yoshida et al., 2019). Yoshida et al. characterized accessibility through bulk ATACseq across 90 murine cell types, including 78 immune cell types descending from bone marrow derived, hematopoietic stem cells. The availability of bulk ATACseq across such a large number of cell types using consistent protocols provides a novel opportunity to investigate accessibility patterns in hematopoietic cell types. Typically, in ATACseq studies across multiple cell types, the ATACseq workflow ends with the formation of an accessibility matrix M, with the rows of M corresponding to genomic loci and the columns corresponding to cell types. M can be binary, reflecting a call of accessible or not-accessible for a particular genomic locus in a particular cell type or can take a range of values, for example if the height of the ATACseq peak is used to quantify accessibility. Describing chromatin accessibility across cell types can then be framed as describing the structure of M. In our context, we are interested in understanding how the structure of M, which is built from the Yoshida et al. ATACseq dataset, reflects the hematopoietic differentiation tree.

In a general setting, the most common way to describe the structure of a matrix, M, is to construct another matrix, $\tilde{M}$, with some simple form that is a good approximation of M. Standard approaches, such as the svd, are difficult to interpret, and have not been commonly used in the context of genomics data. Starting with gene expression datasets in the early 2000s (Eisen et al., 1999; Perou et al., 2000; Saelens et al., 2018) and extending to current ATACseq datasets, clustering has been the most common approach to describing M.

In the case of chromatin accessibility datasets, a common clustering analysis involves clustering of the columns (i.e. cell types), typically by dimension reduction followed by k-means or by hierarchical clustering, which identifies cell types with similar chromatin accessibility patterns across the genome, e.g. (Cusanovich et al., 2018; Collins et al., 2019; Sciumè et al., 2020). Column clustering has the advantage of decomposing the cell types of the differentiation tree into distinct clusters that can then be analyzed. However, cell type clustering provides little information about the overall structure of M which typically has many more rows than columns. Row (i.e. locus) based clustering, which identifies loci with similar cell type accessibility patterns, is also common, e.g. (Song et al., 2011; Lara-Astiaso et al., 2014; Scott-Browne et al., 2016; Lau et al., 2018; Yoshida et al., 2019). Lara-Astiaso et al. (2014) used k-means to row cluster a dataset involving 16 hematopoietic cell types. They noted that loci in different row clusters were accessible across different cell types (see their **Figure 1**). For example, in one of their locus clusters, the loci were accessible in stem cells, but not in other cell types. Yoshida et al. used t-SNE to project rows (i.e. loci) onto 2-d and then identified loci that cluster in the 2-d space and are either accessible across all cell types or are accessible within a single cell type. Both these examples reflect an association between M and the differentiation tree, but restriction to row clustering limits investigation of the association.

Biclustering, which involves specifying a paired cluster of rows (i.e. loci) and columns (i.e. cell types), received significant attention during the 2000s in the context of gene expression data (Cheng and Church, 2000; Lazzeroni and Owen, 2002; Shabalin et al., 2016). Biclustering is particularly effective at finding substructure within M and multiple biclusters can be found and used to construct an approximating matrix $\tilde{M}$, but the resulting approximating matrix $\tilde{M}$ can be difficult to interpret and its computation is often unstable (Pontes et al., 2015).

Here, we take a middle ground between row or column based clustering and biclustering by considering the structure of M through co-clustering. By co-clustering, we mean selecting row clusters and column clusters whose pairings provide a grid-like structure to the approximating matrix $\tilde{M}$. While in bi-clustering a particular row cluster is paired with a particular column cluster, so that rows in different clusters can have their columns clustered in different ways, in co-clustering every row cluster is paired with every column cluster, so that every row has its columns clustered in the same way. In some sense, our work is an extension of the row clustering results of Lara-Astasio et al. (discussed above) which suggested that the structure of M can be well approximated by co-clustering. Our approach is to first row cluster, using the well-known Louvain algorithm, (Blondel et al., 2008), which allows for scaling to large number of loci, a typical situation in chromatin studies. We then column cluster. But importantly, we develop a novel clustering algorithm that restricts column clustering to clusters that are composed of coherent components of the hematopoietic differentiation tree. This co-clustering provides a simple and biologically meaningful structure to $\tilde{M}$ in which a particular column cluster is a coherent hematopoietic phenotype and the overall structure of M can be viewed through the accessibility of these hematopoietic phenotypes across multiple locus clusters. Further, the construction allows us to characterize the variance in M that is associated with the differentiation tree.

Previous authors have considered clustering loci in the context of a cell type network such as the hematopoietic differentiation tree, but typically with the goal of annotating loci (Biesinger et al., 2013; Sohn et al., 2015; Zhang et al., 2016). For example, treeHMM (Biesinger et al., 2013) infers a hidden state at each genomic locus for each of the cell types through a hidden Markov model, with the hidden state serving as an annotation of the locus. From our perspective, these methods serve to construct a matrix

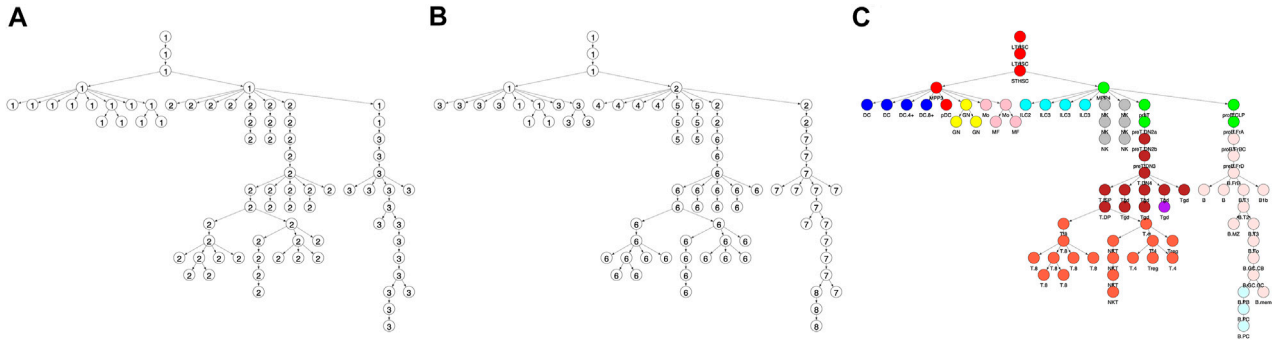

FIGURE 1 | Locus accessibility across locus clusters reveals cell type specific patterns. Shown is the fraction of loci that were accessible within a locus (i.e. row) cluster across each of the cell types in the differentiation tree. Trees correspond to locus cluster **(A)** 5, **(B)** 7, and **(C)** 9. Within each tree, the size of a vertex provides the fraction of loci in the cluster that were accessible for the corresponding cell type. For example, cluster 5 had 5,947 loci with 87 and 2% accessible in the DC.4+ (red vertex) and ILC2 cell types, respectively, as shown in panel A (**Figure 4** provides the same data, but in a heatmap format).

M-which is the starting point of our analysis-with the value of M being the hidden state of the HMM across cell types and loci. In contrast to our setting, in which we just have ATACseq data, these methods allow for multiple assays for each cell type-for example ChIPseqs of different histone modifications-in which case constructing M is complex.

We show that roughly 1/2 of accessible loci in the Yoshida et al. dataset are accessible in only one or 2 cell types in the differentiation tree. Putting aside these cell type specific loci, we show that the other loci fall into roughly 20 locus clusters. Each of these locus clusters can be characterized by cell types in which the accessibility is relatively high and cell types that are relatively low, and the cell types with high accessibility compose a coherent component of the differentiation tree. We show that with 12 cell type clusters (i.e. column clusters) that decompose the differentiation tree, we can capture roughly 80% of the cell type specific variation in M. We also investigate transcription factors (TFs) in the context of this co-clustering, showing that the co-clustered structure of M is reflected in the motif and binding patterns of TF across loci and cell types.

2 MATERIALS AND METHODS

In a Python package available for download, chromcocluster, we have implemented the algorithms described here. The package also includes all files needed to reproduce the particular clusterings of the Yoshida et al. dataset that we present here. The package can be accessed through PyPI at https://pypi.org/project/chromcocluster/ or through Github at https://github.com/SLeviyang/chromcocluster.

2.1 Construction of the M Matrix

We downloaded the Yoshida fastq files from GEO accession GSE100738. We used the standard ENCODE ATACseq workflow to call peaks at a particular IDR. We collected all peaks called by the ATACseq workflow across all cell types. Each peak was associated with a locus on the murine mm10 genome centered at the peak summit and extended 250 base pairs up and down stream, forming a 500 base pair window. To each window, we associated the quality score of the peak summit that defined it. Then for each chromosome, moving $5'$–$3'$, we sequentially evaluated the loci and formed a master list of loci. We did this in a greedy manner. When we encountered a locus that did not intersect with a locus already in our master list, we gathered the current locus and, moving $5'$–$3'$, we gathered all subsequent loci not in the master list that intersected with the current locus. From these gathered loci, we selected the locus with the highest quality score and added it to the list. Then we moved on to the next locus that did not intersect with the added locus. Previous authors used a similar approach (Lara-Astiaso et al., 2014; Corces et al., 2016; Yoshida et al., 2019). Given the master list of loci, we then formed M. A locus contained a peak for a given cell type if any of the cell types peaks intersected with the locus.

2.2 Locus Clustering

As an input graph to the Louvain algorithm, we let each row of M be a node and placed edges between rows (i.e. nodes) at a particular FDR. Each row consisted of 78 ones and zeros. Given two rows with n and N ones, respectively, we let s be the number of columns in which the two rows shared a 1. We assumed that s had a hypergeometric distribution (78 balls, N white balls and n draws), which corresponds to a null in which we permute the column of one of the rows. We then calculated the p-value cutoff that would lead to the particular FDR given the matrix M. Once the graph was constructed, we used the Python sklearn package implementation of the Louvain algorithm to perform the clustering.

2.3 Cell Type Clustering

To precisely define the notion of a cell type cluster that *respects* the differentiation tree, let $\mathcal{U}$ be a set of vertices (i.e. cell types or columns of M) on the differentiation tree. Let $\mathcal{V}_j$ for $j = 1, 2, \ldots, \ell$ be the partition of $\mathcal{U}$ into its ℓ connected components. If $\ell = 1$, then $\mathcal{U}$ is connected and *respects* the tree. If $\ell > 1$, let $r_1, r_2, \ldots, r_\ell$ be the roots of the connected components $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_\ell$, respectively. Then $\mathcal{U}$ *respects* the tree if there is a single node p in the differentiation tree for which $r_1, r_2, \ldots, r_\ell$ are all children. Note that the r_j need not be all the children of p.

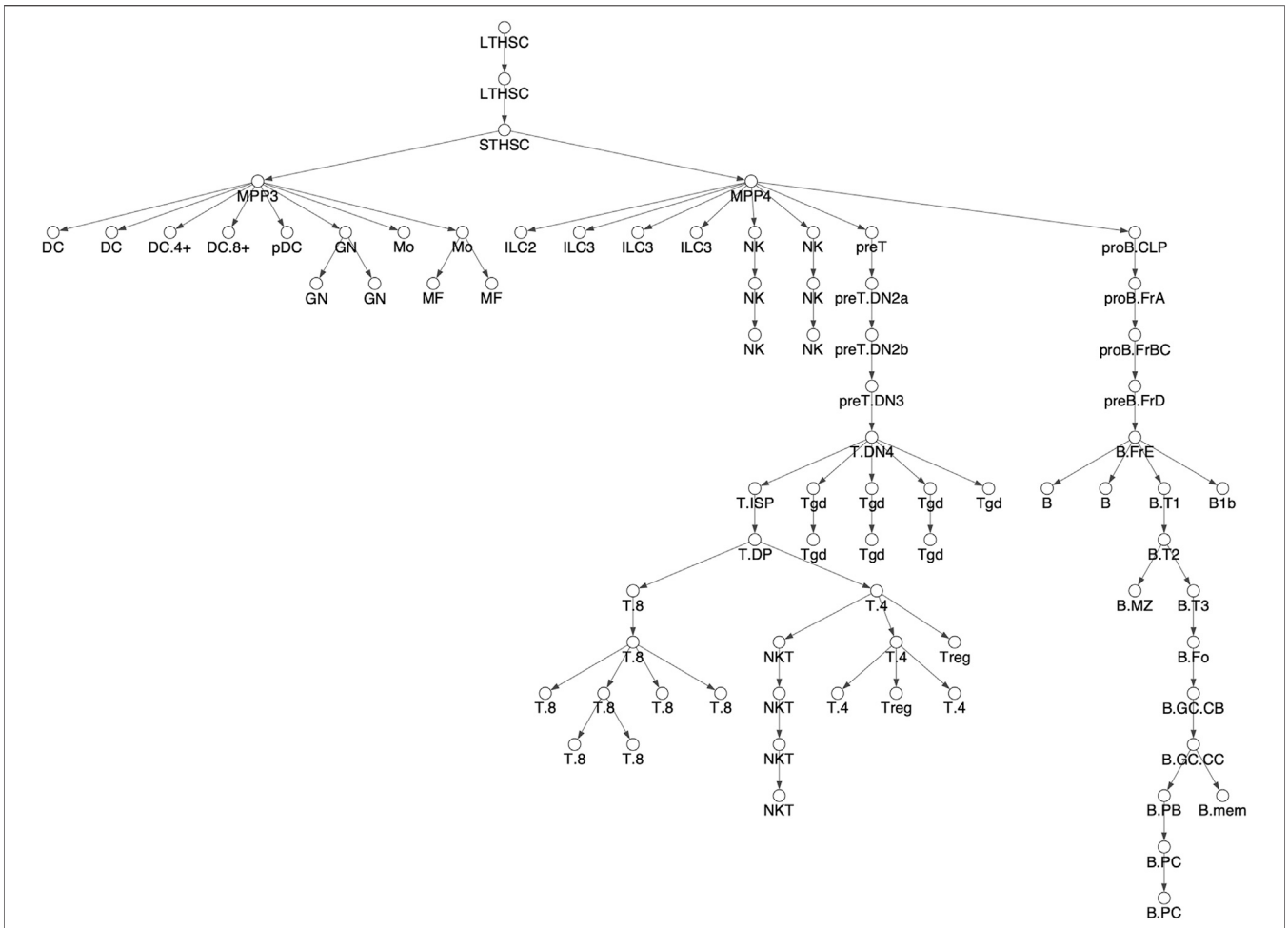

FIGURE 2 | Differentiation tree. We considered 78 hematopoietic cell types and assumed the shown tree structure. The tree structure is the same as **Figure 3A** in Yoshida et al. Cell types names have been shortened for readability, see **Supplementary Figure S1** for full names corresponding to Yoshida et al.

Choosing a partitioning of the differentiation tree into k clusters that *respect* the tree is equivalent to selecting $k - 1$ nodes p_i for $i = 1, 2, \ldots, k - 1$ and for each p_i selecting ℓ_i nodes that are children of p_i: $r_{i,1}, r_{i,2}, \ldots, r_{i,\ell_i}$. Importantly, the p_i need not be distinct, but the $r_{i,j}$ must all be unique. This allows a particular parent node to associated with more than one cluster. For example, in **Figure 2C**, the MPP4 cell type is a parent node to two clusters, the grey NK cluster and the aqua ILC cluster. In this example, p_1 and p_2 could both be the MPP4 cell type, the children of p_1, i.e. $r_{1,1}, r_{1,2}, \ldots, r_{1,4}$, would be the four ILC nodes and the children of p_2, $r_{2,1}$ and $r_{2,2}$, would be the two NK nodes that are children of MPP4.

Given a particular choice for the p_i and $r_{i,j}$, the differentiation tree can be partitioned into disjoint clusters U_i for $i = 1, 2, \ldots, k$ that respect the tree as follows.

1. Remove the edges in the differentiation tree between each p_i and its child nodes $r_{i,j}$ for $i = 1, 2, \ldots, k{-}1$ and $j = 1, 2, \ldots, \ell_i$. This will form a connected component $V_{i,j}$ for each $i = 1, 2, \ldots, k{-}1$ and $j = 1, 2, \ldots, \ell_i$ and an additional connected component V_k that has the root of the differentiation tree as its root.
2. For $i = 1, 2, \ldots, k{-}1$, the cluster U_i is formed by the union of the connected components $V_{i,j}$ for $j = 1, 2, \ldots, \ell_i$. And the cluster U_k is simply the connected component V_k.

Note that each cluster U_i respects the tree because it is formed by connected components with roots $r_{i,1}, r_{i,2}, \ldots, r_{i,\ell_i}$ that are children of a single parent node p_i.

Cell type clustering partitions the column indices (i.e. cell types) into the k sets $U_1, U_2, \ldots, U_k$. As described above, the Louvain algorithm forms row clusters, let those be $W_1, W_2, \ldots, W_n$. Then each pair U_i, W_j specifies a co-cluster composed of all elements of M with row index in W_j and column index in U_i and we build an approximation to M, $\tilde{M}^{(k)}$, which has the same dimensions as M but for which elements in a co-cluster are replaced by the co-cluster mean. We measure fit through the sum of squared differences between M and $\tilde{M}^{(k)}$, i.e., the squared Frobenius norm of the difference between these two matrices, $\|M - \tilde{M}^{(k)}\|_2^2$.

For a given clustering, we refer to the set of p_i as cut vertices (since we cut the edges emanating from them to form our clusters) and for a given p_i, we refer to the vertices $r_{i,j}$ for

$j = 1, 2, \ldots, \ell_i$ as the cut group of the cut vertex p_i. Note that a cut vertex can have multiple cut groups, as noted in the example above.

To minimize $\|M - \tilde{M}^{(k)}\|^2$, we take an iterative approach. First, we construct an initial, random clustering. We choose the cut vertices $p_1, p_2, \ldots, p_{k-1}$ randomly without replacement from the non-leaf vertices of the differentiation tree and for each p_i we choose a single cut group composed of all of its children. We then iteratively attempt to improve the fit of the clustering through two types of modifications:

1. Cut group modifications:

We choose a cut vertex v and enumerate all modifications to the cut groups of v that involve adding a single child of v to a cut group, removing a vertex from a cut group, or transferring a vertex between cut groups if v has more than 1 cut group. We update our clustering to the best fit modification, if it is lower than the current fit.

2. Cut vertex modifications:

We choose a cut vertex v and a vertex $v' \neq v$. We then consider deleting a cut group of v and adding a cut group to v'. The new cut group in v' is chosen by selecting a single child of v' that is not in a cut group, and we consider all the modifications formed by the different children of v'. We update our clustering to the best fit modification, if it is lower than the current fit.

We iteratively improve the fit by applying the cut group modification in a cycle through all cut vertices v, followed by applying the cut vertex modifications in a cycle through all combinations of pairs v, v' where v is a cut vertex. We stop the algorithm when we complete both these cycles without improvement in fit. Since the optimization is non-convex, we applied this iteration using 20 different initial clusterings to determine the clusterings presented in the Results.

2.4 ANOVA Decomposition

For notational convenience, let M and $\tilde{M}^{(k)}$ be the matrices M and $\tilde{M}^{(k)}$ restricted to rows in a particular locus cluster. Then a standard ANOVA analysis decomposes the variation of M into a portion predicted by $\tilde{M}^{(k)}$ and a residual portion. The associated R-squared is given by,

$$\mathrm{R}^2_{\mathrm{total}} = 1 - \frac{\|M - \tilde{M}^{(k)}\|_2^2}{\|M - \mu\|^2},$$

where μ is the mean of the entries of M. $\mathrm{R}^2_{\mathrm{total}}$ is exactly the standard R-squared of a linear predictor. The means of the columns of M are not affected by column clustering. With this in mind, we consider the prediction of the column means of M, which we write as $M_{\cdot,i}$ by $\tilde{M}^{(k)}$. Then the portion of variation of $M_{\cdot,i}$ predicted by $\tilde{M}^{(k)}$ is given by the R-squared expression,

$$\mathrm{R}^2_{\text{cell type}} = 1 - \frac{\|M_{\cdot,i} - \tilde{M}^{(k)}\|_2^2}{\|M_{\cdot,i} - \mu\|_2^2}.$$

We calculate R-squared over the whole matrix M by averaging the R-squared values over the locus clusters.

2.5 TF Motif Analysis

Most of our TF analysis followed the workflow described in Schep et al. in (Schep et al., 2017). We downloaded motif descriptions using the R package chromVARmotifs and then used the R package motifmatchR to call motifs on the DNA sequences spanned by our loci at a p-value of 5E-6. This gave us a binary matrix, A, analogous to M except that columns corresponded to motifs and a 0 and 1 value in an entry corresponded to the absence or presence of a motif at a locus, respectively.

To determine motif enrichment for a co-cluster, for each cell type in the co-cluster, we calculated the fraction of accessible loci that contained the motif and then averaged over all cell types in the cluster. This gave us a raw co-cluster score r. We computed an analogous raw null score n using all loci and cell types not in the co-cluster. Finally, we computed an enrichment score,

$$\text{enrichment score} = \frac{r - n}{r + n}. \quad (1)$$

Schep et al. computed a similar score, except that they normalize r–n by a variance term. We found that the variance term was often small, leading to statistical instability, so instead we normalized by $r + n$. We then permuted columns to find a cutoff to our enrichment score that gave a 0.05 FDR. We called a motif as enriched for a co-cluster if its enrichment score exceeded the cutoff.

2.6 ChIPseq Workflow

We downloaded fastq files from Revilla-I-Domingo et al. (2012), with GEO accessions GSM932921, GSM932922 GSM932925, and GSM932926. We downloaded the fastq files from Wagner et al. (2020) and Istaces et al. (2019) with GEO accessions GSM3900380, GSM3900381 and GSM3559328, GSM3559327, respectively. We aligned the fastq to the murine mm10 genome using bowtie2 (-X1000 was the only non-standard flag) (Langmead et al., 2009), filtered for poorly aligned reads and duplicates using samtools and Picard, and called peaks using MACS2 (Zhang et al., 2008) using all standard settings in the macs2 callpeak executable except with the quality cutoff set to 0.10 (-q 0.10). We called a locus from our ATACseq analysis as containing a ChIPseq peak if the ChIPseq peak summit was within the locus 500 base pair window.

3 RESULTS

We downloaded the ATACseq data of Yoshida et al. from the NCBI GEO database (Barrett et al., 2013). The Yoshida et al. dataset includes ATACseq libraries across 90 cell types, but we restricted our attention to the 78 cell types derived from adult, bone marrow derived, hematopoietic stem cells, i.e. HSCs (Seita and Weissman, 2010). The other 12 cell type, which do not derive from HSCs, include 5 stromal cell types and seven embryonic

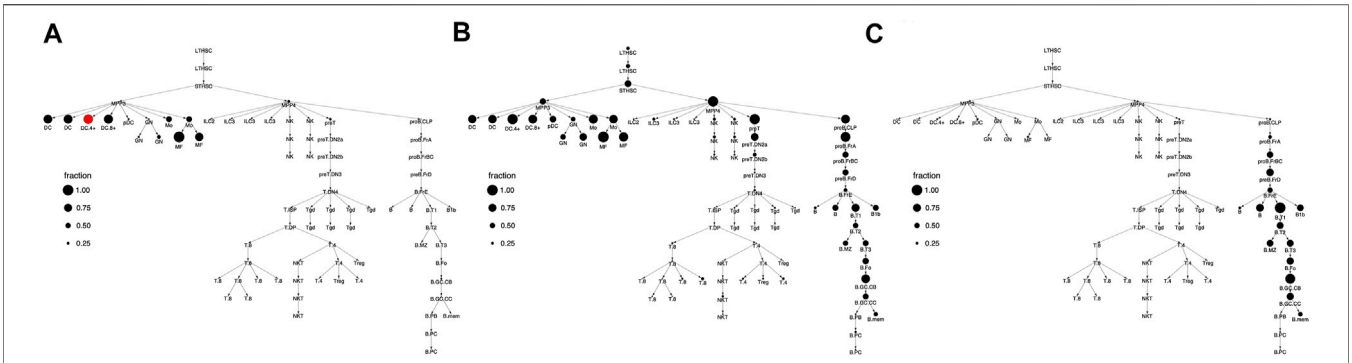

FIGURE 3 | Cell type clusterings that respect the hematopoietic differentiation tree. Shown are the column (i.e. cell type) clusters for **(A)** $k = 3$ **(B)** $k = 8$, and **(C)** $k = 12$. In panel C, clusters are specified by color and this color scheme is used throughout the text. In panels A and B, clusters are specified by number.

yolk sak derived, macrophage cell types. The differentiation tree we assume is shown in **Figure 3** and is exactly that of Yoshida et al., as shown in their **Figure 3A**. (In our **Figure 3**, cell names are shortened for readability. See **Supplementary Figure S1** for full names corresponding to Yoshida et al.). We applied the ENCODE ATACseq pipeline (Davis et al., 2018) to the Yoshida et al. ATACseq samples for each of the cell types, resulting in a collection of peaks representing accessible loci for each cell type. Then, following the approach of previous authors, e.g. (Lara-Astiaso et al., 2014; Corces et al., 2016; Yoshida et al., 2019), we constructed a master list of loci, composed of non-overlapping 500 base pair windows that intersected with every ATACseq peak across all cell types. We then formed the chromatin accessibility matrix M with the rows corresponding to each of the loci in the master list and the columns corresponding to the 78 cell types. An entry of M was 1 if the cell type had a peak that intersected with the corresponding 500 base pair locus, otherwise the entry was 0.

An important issue in constructing M is the sensitivity, specificity trade-off in calling locus accessibility. We modulated this tradeoff by choosing different IDR values (Li et al., 2011) in the ENCODE pipeline. The IDR is similar to an FDR and is used to determine reproducible peak calls. We considered IDR values of 0.01, 0.05, 0.10 and 0.15 which led to roughly 159, 246, 331, and 424 thousand loci which were accessible, respectively, in one or more of our 78 cell types. As comparison, Yoshida et al. considered roughly 512 thousand loci over their 90 cell types. We defined a cell specific locus as a locus which was accessible in 2 or less cell types; we found that 38, 43, 51, and 57% of the loci for the respective IDR values were cell specific. The increasing level of cell specific loci with increasing IDR may reflect increasing noise. Alternatively, cell specific accessible loci may have lower levels of accessibility, leading to their being called only at larger IDR values. Regardless of the specific IDR, the number of cell specific and non-cell specific loci both constituted a substantial fraction of the total accessible loci. Below we present results for an IDR of 0.01, taking a conservative approach to calling accessibility. Under this IDR, 94% of our loci intersected with the midpoint of a Yoshida et al. locus. Our results were essentially unchanged using other IDR, see Materials and Methods for further details.

TABLE 1 | The effect of edge FDR on locus clustering. To apply the Louvain clustering algorithm, we constructed a graph in which loci were represented by nodes and edges between nodes represented loci with similar accessibility patterns. We placed edges between two nodes at different FDR. Shown are the percent of loci that were connected to another locus (connected loci), the number of clusters with more than 30 loci (number of large clusters), and the fraction of loci that fell within a large cluster (loci in large clusters).

FDR	Connected loci	Number of large clusters	Loci in large clusters
0.0001	57%	36	87%
0.001	72	20	97
0.01	89	21	98
0.05	97	17	99
0.10	99	16	99

3.1 Locus Clustering

We clustered the rows (i.e. loci) of the accessibility matrix M using the Louvain algorithm (Blondel et al., 2008). Importantly, we only clustered the rows corresponding to non-cell specific loci, of which there were 98,848 under 0.01 IDR. The Louvain algorithm takes a graph (i.e nodes and edges) as input and clusters the nodes to maximize a measure of community structure. In our setting, the rows (i.e. loci) of M form the nodes. To form edges, we placed an edge between two nodes if the corresponding rows had a statistically significant number of columns with equal entries (i.e.the two loci shared accessibility states across a statistically significant number of cell types). We clustered nodes with edges placed at an FDR of 0.0001, 0.001, 0.01, 0.05, and 0.10, respectively. As shown in **Table 1**, the fraction of nodes connected by an edge to some other node fell as edge FDR was lowered, reflecting the existence of loci with cell type accessibility patterns that did not closely match any other loci. As the table further shows, the Louvain algorithm formed between 16 and 21 clusters with 30 or more nodes across all FDR, except in the case of an FDR of 0.0001. The results for an edge FDR of 0.0001 suggest an overly conservative approach in placing edges, leading to too many clusters and many isolated nodes. Below, we present results for the edge FDR of 0.001, choosing a relatively conservative value as we did for the IDR. Our results are essentially unchanged using the larger edge FDR values, see Materials and Methods for details.

TABLE 2 | Row clusters. Twenty row clusters generated by the Louvain clustering contained 95% of the non-cell specific loci. Shown are the number of loci in each cluster (size), the fraction of cluster loci within 500 base pair (prox, i.e. proximal) and more than 3,000 base pair (dist, i.e. distal) of a transcription start site (TSS), and the fraction of the cluster's proximal loci and distal loci that were in the FANTOM database of promoters (prom) and active enhancers (enhan). Row clusters 0 and 3 had loci largely proximal to TSS, while all other clusters were largely composed of loci distal to TSS. In most clusters, greater than 70% of the proximal loci and 10–40% of distal loci were labeled as promoters and active enhancers, respectively.

Cluster	Size	Location		Fantom		Cluster	Size	Location		Fantom	
		Prox	Dist	Prom	Enhan			Prox	Dist	Prom	Enhan
0	5,960	0.92	0.06	0.81	0.42	10	3,449	0.09	0.86	0.73	0.13
1	9,289	0.08	0.85	0.67	0.18	11	2,453	0.05	0.85	0.76	0.25
2	6,931	0.06	0.88	0.70	0.09	12	2,152	0.34	0.59	0.78	0.38
3	6,293	0.60	0.35	0.78	0.43	13	704	0.03	0.93	0.50	0.05
4	5,979	0.19	0.75	0.71	0.30	14	673	0.02	0.91	0.71	0.16
5	5,947	0.05	0.87	0.66	0.17	15	441	0.05	0.91	0.55	0.12
6	5,549	0.10	0.83	0.77	0.23	16	304	0.04	0.87	0.77	0.28
7	5,172	0.14	0.78	0.75	0.30	17	292	0.02	0.93	0.20	0.11
8	3,689	0.04	0.88	0.53	0.10	18	173	0.04	0.91	0.86	0.08
9	3,557	0.04	0.88	0.61	0.12	19	45	0.02	0.93	0.00	0.17

At an edge FDR of 0.001, the Louvain algorithm produced 100s of clusters, but the top 20 clusters included 95% of the nodes in the graph with the remaining clusters all containing less than 30 nodes and most containing only 2 nodes. The smaller clusters could reflect noise in calling loci and edges or they could reflect loci with uncommon patterns of accessibility. **Table 2** shows the number of loci in each of the largest 20 locus clusters. Of the loci in these 20 clusters, 97% intersected with the midpoint of a Yoshida et al. locus. The first 13 clusters have greater than 2000 loci and the rest of the clusters have 100s of loci, except for cluster 19 which has 45 loci. **Figure 4** shows the fraction of loci in each row cluster that are accessible within each of the 78 cell types.

Notably, clusters 0 and 3 contained loci that are accessible in almost all cell types. As shown in **Table 2**, most of the loci in these clusters are proximal to transcription start sites (TSS), which we define as within 500 base pair of a TSS. Most of the loci in the other locus clusters, accessible in only a subset of the cell types, were distal to TSS, which we defined as greater than 3,000 base pair from a TSS. Yoshida et al. noted a similar pattern, with loci close to TSS (what they term TSS OCR) accessible across most cell types and loci far from TSS (what they term DE OCR) accessible in only certain cell types.

To investigate locus functionality, we downloaded the list of mouse promoters and active enhancers maintained by the FANTOM consortium (FANTOM5 version) (Lizio et al., 2015). As shown in **Table 2**, across almost all clusters, greater than 70% of the proximal loci were identified as promoters by FANTOM. For distal loci, roughly between 10 and 40% were identified as active enhancers by FANTOM. The lower fraction of distal loci identified by FANTOM may reflect cell types not sampled by FANTOM, non-active enhancers, or accessible loci that are not enhancers. However, both proximal and distal loci were statistically enriched for loci identified by FANTOM, supporting a functional role for the clustered loci.

The clustering reveals clear associations between cell phenotype and accessibility. As an example, **Figure 1** shows the fraction of loci in row clusters 5, 7, and 9 called as accessible for each cell type. The figure gives the same data as rows 5, 7, and 9 of the heatmap in **Figure 4**, but in the context of the differentiation tree. In row cluster 5, roughly 50–80% of loci are accessible in macrophages and DCs, while in other cell types accessibility of these loci is less than 2%. In row cluster 7, between 40 and 80% of loci are accessible in macrophages, dendritic cells, and most B cells, while in other cell types less than 10% of loci are accessible. In row cluster 9, all B cells except pro-B cells and plasma B cells have between 60 and 95% of loci as accessible, while in all other cell types less than 10% are accessible. These three row clusters reflect a decomposition of accessible loci in macrophages, DC and B cells into loci that are accessible only in macrophages and DC, only in B cells and jointly. Some of the smaller row clusters, which reflect a more specific cell phenotype, have a more definitive separation of accessibility across cell type. For example, all 704 loci in cluster 14 are accessible in ILC3 cell types and inaccessible in all other cell types. The less definitive separation in cell type accessibility we see in clusters 5, 7, and 9 may reflect experimental noise in the ATACseq workflow, but particular loci within a cell type may also vary in their accessibility over time due to unstable positioning of nucleosomes or due to a variation of cell state (Natoli et al., 2011; Wang et al., 2012).

3.2 Cell Type Clustering

With the row clustering fixed, we next applied a column (i.e. cell type) clustering. For column clustering, we chose the number of column clusters as a particular value, k, and produced a column clustering for each $k = 2, 3, \ldots, 12$. Since one of our main motivations was to quantify the degree to which accessibility associates with the differentiation tree, we restricted column clusters to reflect the structure of the tree through a novel algorithm.

Given a graph (i.e. nodes and edges), a set of nodes is said to be connected if there is a path along the tree connecting every pair of nodes in the set. Importantly, restricting cell type clusters to connected components did not give good results. As an example, consider accessibility in locus (i.e. row) cluster 5, as shown in **Figure 1**. Only 5% of loci in the cluster are accessible in the MPP3 progenitor cell type, but 80% the loci are accessible in macrophage and DC, which are children of the MMP3 cell type. However, for pDC (plasmacytoid DC) and GN

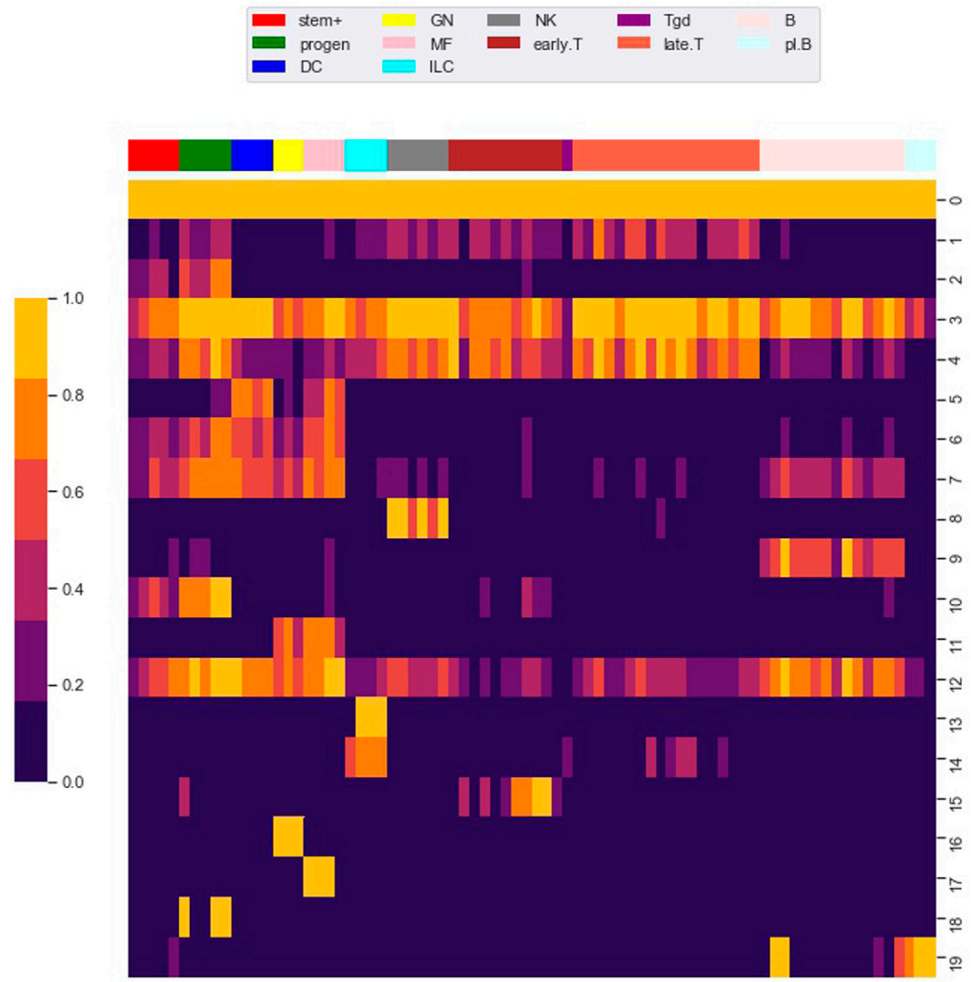

FIGURE 4 | Row clustering using the louvain algorithm. We used the Louvain algorithm to cluster 98,848 rows (i.e. loci) of the *M* matrix, giving 20 main row (i.e. locus) clusters. Each row in the heat map corresponds to a row cluster and each column to a cell type. The number of loci in each row cluster is given in **Table 2**. Heatmap colors show the fraction of loci within a row cluster that were accessible for the particular cell type. Column groupings by cell type are based on our column clustering for $k = 12$ discussed below.

(neutrophil), which are also children of MMP3, the loci are relatively inaccessible (<30%). To account for this dynamic, we say a set of nodes *respects* the differentiation tree if

1. The nodes form a single connected component or,
2. The nodes form multiple connected components, but the root nodes of the connected components are all children of a common parent node.

As an example, consider the clustering shown by the node coloring in **Figure 2C**. The rose colored cluster is composed of T.4 and T.8 cell types, each of which forms a separate connected component of the tree. But the root nodes of the T.4 and T.8 connected components are both children of the T.DP cell type, so that the cluster *respects* the tree. As another example, consider the grey colored cluster composed of NK cells. The cluster is composed of two connected components, but the two roots of the connected components are both children of the MPP4 cell type, so that cluster *respects* the tree. See Methods for further details.

For a given k, our clustering algorithm selects the cell type clustering that respects the tree with the best co-clustering approximation of M, see Methods for algorithm details. **Figure 2** shows the column clustering produced for $k = 3$, 8, 12 on the differentiation tree. When $k = 3$, our clustering splits the differentiation tree into a stem/progenitor cell and myeloid cluster, a B cell cluster, and a T cell cluster. This clustering shows that the general division of immune cell types into myeloid, B, and T phenotypes is reflected in accessibility differences. Further, stem cell and progenitor cell types are most similar to myeloid cell types in their accessibility. Increasing to $k = 8$, splits stem cells and progenitor cells into separate clusters, puts the NK and ILC cell types into separate clusters, introduces a cluster containing plasma and memory B cells, and splits the myeloid compartment. Interestingly, neutrophils are grouped with stem cells, while DC, monocytes, and macrophages are split into their

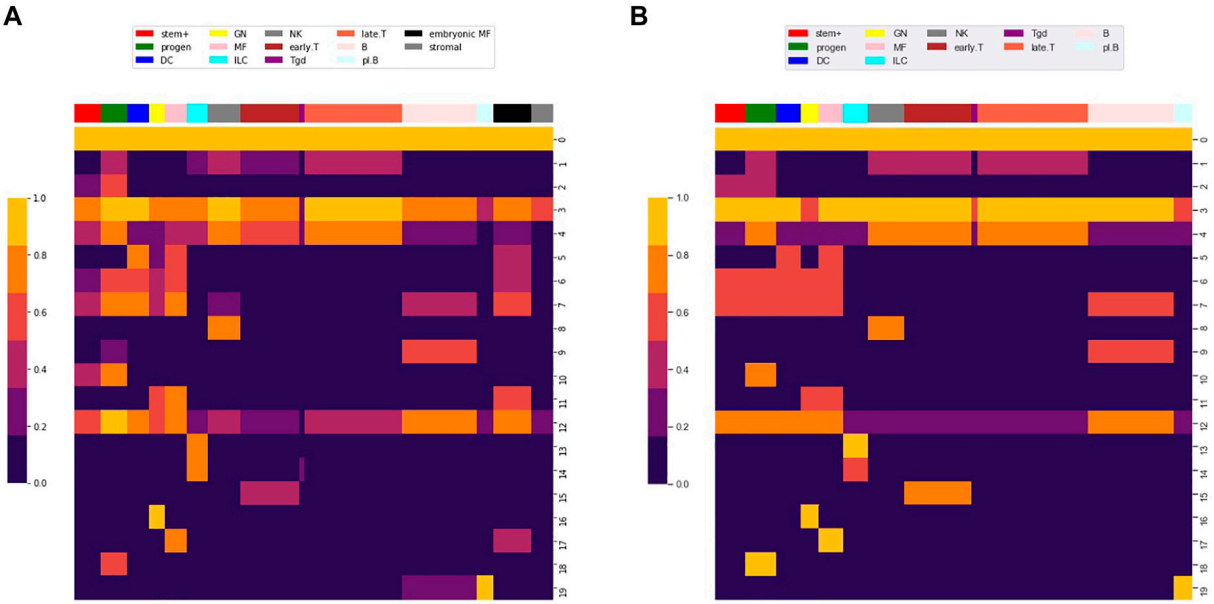

FIGURE 5 | Co-clustered approximation of the accessibility matrix *M*. Heatmaps of **(A)** $\tilde{M}^{(12)}$ and **(B)** $\tilde{M}^{(12)}_{i/a}$. Panel A also includes cluster means for embryonic macrophages (MF) and stromal cells, which are not part of $\tilde{M}^{(12)}$. The co-clustered matrices $\tilde{M}^{(12)}$ and $\tilde{M}^{(12)}_{i/a}$ approximate *M*. Compare the heatmaps of this figure to the heatmap of *M* given in **Figure 4**. Only locus clusters 0 and 3 are accessible in stromal cells, while embryonic MF have similar accessibility patterns to bone marrow derived MF.

own compartment. By $k = 12$, the T cell compartment is split into a CD8 and CD4 cluster and an early T cell cluster.

3.3 Co-Clustered Approximation of the Accessibility Matrix

For a particular k value, the combination of row and column clusters divided M into a grid of $20 \times k$ co-clusters. While M is a binary matrix, we built the co-clustered approximation of M, $\tilde{M}^{(k)}$, by setting all the entries in a co-cluster to the mean value of the co-cluster entries. As an analogy, in the kmeans algorithm, many data points are approximated by the k means associated with the k clusters. Similarly. $\tilde{M}^{(k)}$ approximates the many entries of M through the means of the co-clusters. **Figure 5A** visualizes $\tilde{M}^{(k)}$ for $k = 12$.

As an application of the clustering, we compared the accessibility patterns of stromal and embryonic macrophage cell types to the 78 cell types included in the differentiation tree. **Figure 5A** shows the mean values for stromal and embryonic cell types across the 20 locus clusters. Locus cluster 0 and 3 are accessible in stromal cells, suggesting that these clusters are composed of globally accessible promoters. In contrast, for stromal cells no other cluster has a significant level of accessibility, while embryonic macrophages have an accessibility pattern similar to bone marrow derived macrophages.

In constructing the matrix $\tilde{M}^{(k)}$, we allowed each co-cluster to take on a different value. As a more restrictive model, we assumed that a cell type within a particular row (i.e. locus) cluster can be either in an relatively inaccessible 1) or accessible 1) state, rather than in a continuum of accessibility states. Biologically, this more restrictive model supports a single regulatory mechanism shaping the accessibility structure of the loci in each locus cluster. To examine the impact of this model, we built a matrix $\tilde{M}^{(k)}_{i/a}$ which had the same co-clusters as $\tilde{M}^{(k)}$, but with co-clusters sharing the same locus cluster restricted to have one of two values, representing either an inaccessible or accessible state. **Figure 5B** visualizes $\tilde{M}^{(k)}_{i/a}$ for $k = 12$.

We applied an ANOVA analysis to calculate the variation of M captured by the co-clustered matrices $\tilde{M}^{(k)}$ and $\tilde{M}^{(k)}_{i/a}$. We calculated two R-squared values, R^2_{total} and $R^2_{\text{cell type}}$, for the fraction of the total variation in M captured by the co-clustering and the fraction of cell type (i.e. column) associated variation in M captured by the cell type clustering, respectively; see Methods for details. If each cell type was in a separate cluster, then $R^2_{\text{cell type}}$ would equal 1 and if all cell types were in a single cluster then $R^2_{\text{cell type}}$ would equal 0. Since we constructed our column clusters to respect the differentiation tree, we used $R^2_{\text{cell type}}$ as a measure of the association between accessibility and the structure of the differentiation tree. **Figure 6** shows the R-squared values for different values of k. Also included in the figure are co-clusterings in which cell types were clustered using hierarchical and kmeans clustering, respectively. We used the R-squared values for these two commonly used clustering methods as a baseline against which to compare our clustering approach, which is constrained by the tree.

As seen in **Figure 6A**, the fraction of variation captured by the co-clustering varied between 0.20 and 0.40 as the number of cell type clusters k rose from 2 to 12. In contrast, as shown in **Figure 6B**, the fraction of cell type associated variation

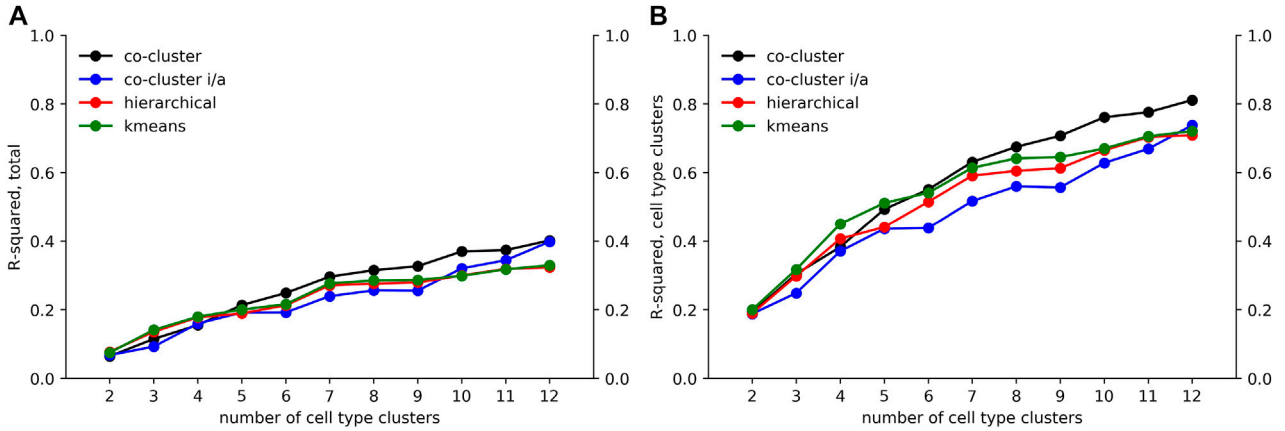

FIGURE 6 | Decomposition of accessibility variation. Using an ANOVA analysis, we calculated an R-squared value giving the fraction of **(A)** total variation and **(B)** cell type associated variation in M captured by our co-clustered approximation matrices $\tilde{M}^{(k)}$ and $\tilde{M}^{(k)}_{i/a}$ as well as co-clustering in which columns were clustered using hierarchical and k-means methods without taking the differentiation tree into account.

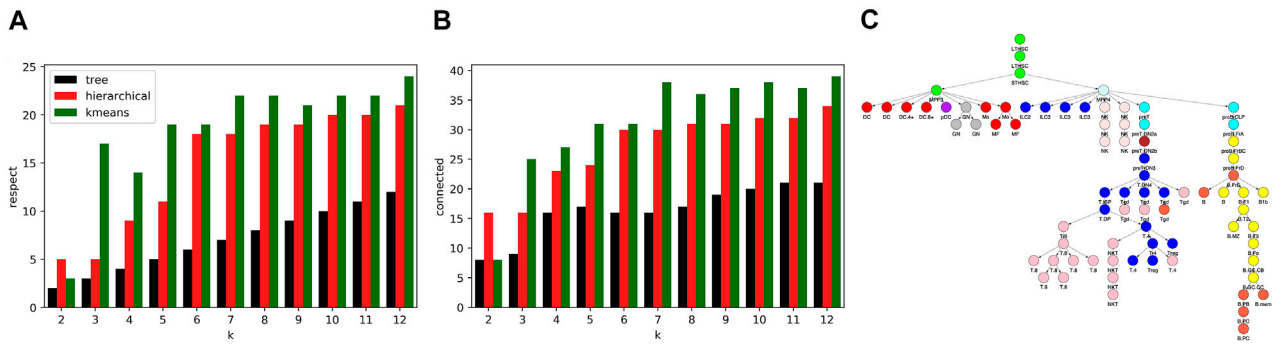

FIGURE 7 | Comparison of cell type clustering across methods. To compare the form of clusters between our tree based method and hierarchical and kmeans methods, we decomposed clusters into components that respected the tree and that were connected. Shown are the sum across clusters of components that **(A)** respect the tree and **(B)** are connected relative to the tree. The number of components generated by our method is less across all values of k. **(C)** Shown are the clusters formed using hierarchical clustering for $k = 12$. As an example of the component counts, the blue cluster is formed from 3 components that respect the tree and six connected components. Note that the clusters are more dispersed on the tree than the clusters of **Figure 4C**.

captured by the cell type clusters rose from 0.20 to 0.80, meaning that our cell type clustering captured most of the association between accessibility and cell type, at least for the higher k values. The large fraction of the total variation in accessibility not captured by our co-clustering (roughly 0.6 for $k = 12$) is associated with the row clustering. As discussed above and reflected in **Figure 4**, within a row (i.e. locus) cluster, cell types tended to have either a high or low fraction of loci in an accessible state, but the high and low fraction were often intermediate, e.g. 0.50 and 0.10, instead of extreme, e.g. 1.0 and 0. This within cell type variation could reflect noise in the ATACseq workflow, stochasticity in the accessibility state of the loci, or row clusters that are too broad.

The R^2_{total} and $R^2_{\text{cell type}}$ values based on the $\tilde{M}^{(k)}$ co-clusterings were similar to values based on $\tilde{M}^{(k)}_{i/a}$ and hierarchical and kmeans column clusterings. The similarity of the R-squared values between $\tilde{M}^{(k)}$ and $\tilde{M}^{(k)}_{i/a}$ provides support for viewing accessibility within a particular locus (i.e. row) cluster as falling into one of two states for all the cell types. The similarity of the R-squared values based on hierarchical and kmeans clusterings to our co-clusterings demonstrates that there is not a significant component of cell type variation that does not respect the differentiation tree.

While the R^2 values of our algorithm and kmeans and hierarchical clustering were similar, the cell type clusters differed substantially in form. To characterize cluster form, we decomposed each cell type cluster into components that respected the tree and connected components and then summed the total components of each decomposition across all clusters. As seen in **Figures 7A,B**, our algorithm generated clusters that could be decomposed into less components that respect the tree and less connected components across all k values. Kmeans and hierarchical clustering led to dispersed clusters. As an example, **Figure 7C** shows the hierarchical clustering for $k =$

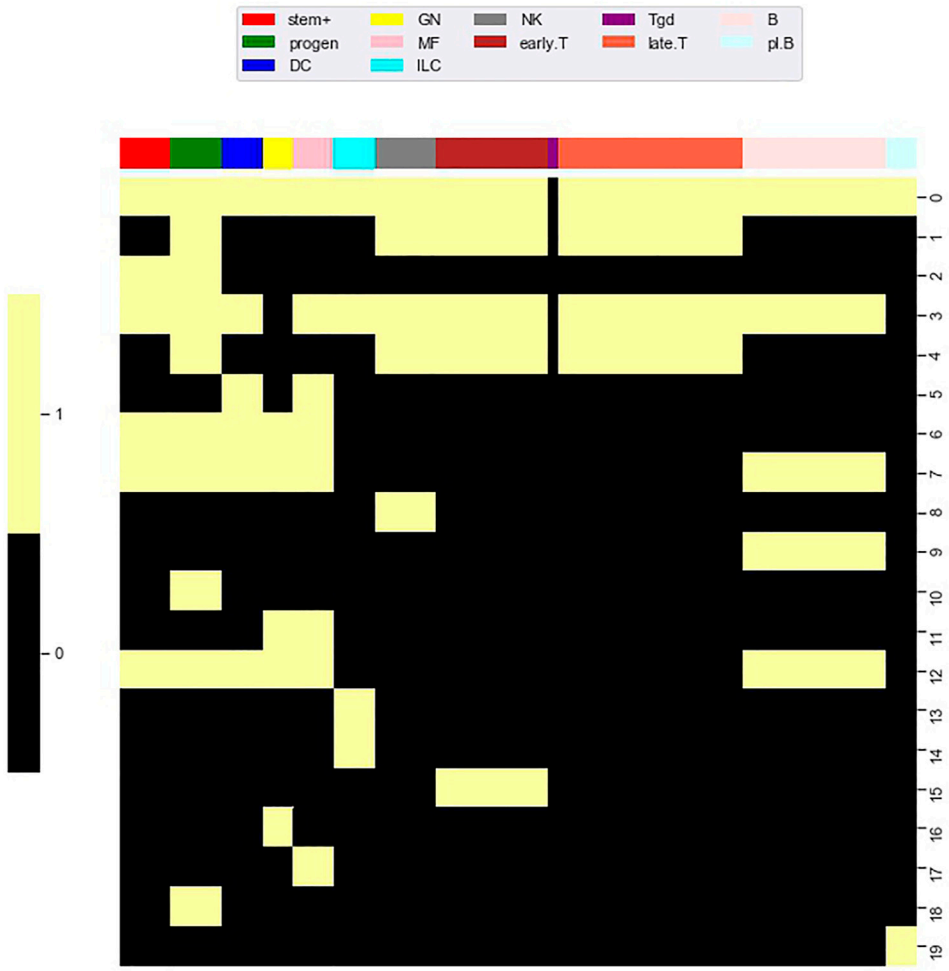

FIGURE 8 | Accessibility co-clusters. To investigate enrichment of TF motifs, we defined a collection of accessibility co-clusters formed through the combination of locus clusters and all cell types in the accessible state of $\tilde{M}_{i/a}^{(12)}$. In the heatmap, accessibility co-clusters are the yellow blocks within each of the 20 rows, compare to $\tilde{M}_{i/a}^{(k)}$ in **Figure 2B**. As an example, accessibility co-cluster 5 is formed by the loci in locus cluster 5 and the DC and macrophage cell type clusters.

12. Importantly, the higher number of components in the cluster decompositions for hierarchical and kmeans clustering did not lead to better fits, as the R^2 results show.

3.4 TF Motif Enrichment Across Co-Clusters

Using methods introduced by Schep et al. (2017), Yoshida et al. computed an accessibility score measuring the enrichment of a TF motif on accessible loci within a particular cell type relative to the presence of the TF motif on accessible loci across all cell types. From the perspective of a matrix analysis, this is a column (i.e. cell type) based approach because enrichment is assessed over all accessible loci for a single cell type. To adapt the method of Schep et al. to co-clusters, we defined 20 co-clusters formed by the combination of each locus cluster and the cell types in the accessible state of $\tilde{M}_{i/a}^{(12)}$ for that locus cluster, see **Figure 8**. We refer to these co-clusters as accessibility co-clusters. There is one accessibility co-cluster for each locus cluster. For example, accessibility co-cluster 9 is formed by the loci in locus cluster 9 and the cell types in our B cell, cell type cluster while accessibility co-cluster 5 is formed by the loci in locus cluster 5 and the DC and macrophage cell type clusters. For each accessibility co-cluster, we adapted the method of Schep et al. by considering enrichment of a TF motif over accessible loci in the accessibility co-cluster against a background of accessible loci over all other loci and cell types, see Materials and Methods for computational details.

Importantly, since loci near TSS tended to be accessible across all cell types, we restricted our analysis to loci greater than 3,000 base pair from a TSS. This had the advantage of restricting our analysis to regulatory features specific to putative enhancers, which are likely different than regulatory features specific to promoters. For TF motifs, we used the 76 TF motifs identified by Yoshida et al. as significantly associated with accessibility (see their **Supplementary Table S5**). Of these TF motifs, we found that 43 were statistically enriched in at least one of the 20 accessibility co-clusters (FDR 0.05).

Figure 9 shows enrichment across the 43 significant motifs and the 20 accessibility co-clusters. Of the 43 TF motifs that we found to be enriched in at least one accessibility co-cluster, the

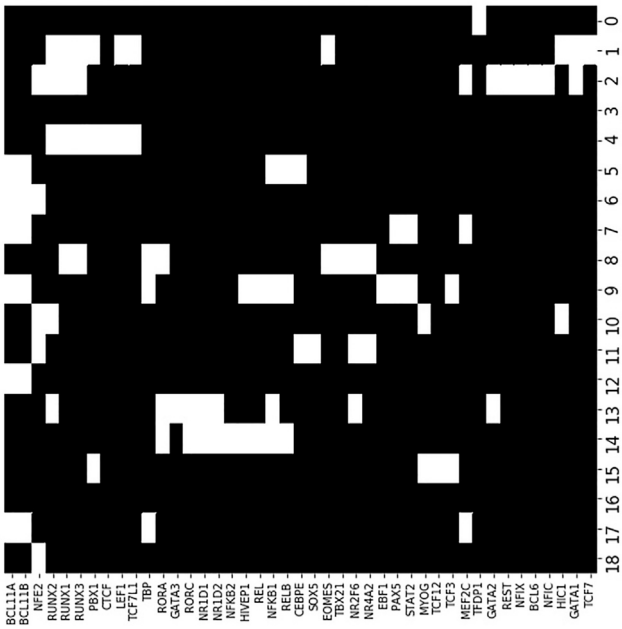

FIGURE 9 | TF motif enrichment associates with particular accessibility co-clusters. We evaluated enrichment of each TF motif (columns) within each accessibility co-cluster (rows). At an FDR of 0.05, we found 43 TF motif, co-cluster pairings that were enriched (white coloring). As a particular example, the BCL11A motif (far left column) was enriched in six accessibility co-clusters, in particular co-cluster 9. Accessibility co-cluster 9 is formed by loci in row cluster 9 and B cell types. The co-cluster is visualized in **Figure 8** by the yellow block within row 9.

seven motifs BCL11A, BCL11B, NFE2, NFKB1, RUNX1, RUNX2, RUNX3 were enriched in more than 3 accessibility co-clusters. The TF BCL11A is an instructive example. Yoshida et al. found BCL11A motifs enriched in accessible loci in B cells and myeloid cell types (see their **Figure 5F**). Similarly, we find BCL11A to be enriched in four accessibility co-clusters: the co-cluster formed by locus cluster 5 and myeloid cell types; the co-cluster formed by locus cluster 7 and stem cells, myeloid cells and B cells; the co-cluster formed by locus cluster 9 and B cells; and the co-cluster formed by locus cluster 12 and myeloid cells and B cells. Our enrichment analysis extends the results of Yoshida et al. by decomposing the cell type enrichment of BCL11A. For example, the enrichment of BCL11A in accessible B-cell loci reflects enrichment in loci that are jointly accessible across myeloid and B cells, but also in loci that are accessible only within B cells or myeloid cells, respectively.

There were 13 motifs enriched for a single accessibility co-cluster with the most significant enrichment occurring for TCF12, TBX21, NFKB2, EBF1, and GATA3. EBF1 and GATA3 are instructive examples. Yoshida et al. also generated RNAseq datasets for each of their cell types. Based on these RNAseq datasets, EBF1 is expressed solely in B cells while GATA3 is expressed in ILC, NK, and T cells. Reflecting these expression patterns, EBF1 is known as a master regulator of B cell differentiation (Nechanitzky et al., 2013) and GATA3 is a regulator of T cell differentiation (Ho et al., 2009). We found the EBF1 motif enriched in the co-cluster formed from locus cluster 9 and B cell types, matching the known regulation role of EBF1. In contrast, GATA3 was enriched in the co-cluster formed from locus cluster 13 and ILC3 cell types. This result matched at least part of the expression pattern of GATA3, but did not reflect the regulatory role of GATA3 in T cell differentiation. For GATA3, our enrichment shows that ILC3 cells have accessible loci that are more enriched for the GATA3 motif than T cells, but the regulatory significance of this result is unclear.

The remaining 23 TF were enriched in 2 or 3 co-clusters. The transcription factors PAX5 and EOMES are instructive examples. PAX5 is a master regulator of B cell differentiation (Horcher et al., 2001). We find PAX5 enriched in the co-clusters formed by locus cluster 7 and myeloid and B cells and by locus cluster 9 and B cells. Yoshida et al.'s RNAseq data show that PAX5 is expressed solely in B cells, so PAX5 motifs in locus cluster 7 are not bound by PAX5 in myeloid cell types, but these loci are accessible, suggesting an association with other TFs. EOMES regulates effector NK and T cells (Gordon et al., 2012) and, in line with this regulation, Yoshida et al.'s RNAseq data shows EOMES expressed in NK cells and CD8 T cells. We find EOMES enriched in two accessibility co-clusters, one formed by locus cluster 1 and T cells and one formed locus cluster 8 and NK cells. EOMES motifs are enriched in two locus clusters with loci that are accessible in disjoint cell clusters, T and NK cell types, respectively. In contrast, PAX5 motifs were enriched in co-clusters that spanned multiple cell clusters, e.g. B cells and myeloid cells, over which the loci were jointly accessible.

3.5 Association of TF ChIPseq Peaks and Co-Clusters

To validate and explore the functional consequences of our TF motif analysis, we collected publicly available ChIPseq datasets from the GEO database for PAX5 and EOMES. We used PAX5 ChIPseq data from Revilla-I-Domingo et al. (2012), that sampled pro-B cells and mature B cells, and we used EOMES ChIPseq data from Wagner et al. (2020) and Istaces et al. (2019) that sampled NK cells and CD8 thymocytes, respectively. Briefly, we downloaded fastq files from GEO and applied a standard peak calling workflow to call peaks. We then identified intersections between our master collection of 159 thousand accessible loci and the TF peaks, see Materials and Methods for accessions and workflow details.

Figure 10A shows a scaled count of the number of loci that contained a PAX5 peak across different locus clusters for the mature-B and pro-B cell types. Since locus clusters differ in size, the raw count is not directly informative. Instead, we scaled the raw count by the expected count under the null of equal distribution of peaks across loci. A scaled count above one represents an enrichment of peaks in the locus cluster. Only locus clusters with a scaled count greater than 1 for either the mature B or pro-B ChIPseq datasets are shown. Our TF motif analysis showed enriched motifs in accessibility co-clusters 7 and 9 and, correspondingly, for both the pro-B and mature-B cell ChIPseqs, locus clusters 7 and 9 had enriched counts.

The mature B cell type falls within our B cells, cell type cluster. B cells are present in the accessibility clusters formed by locus

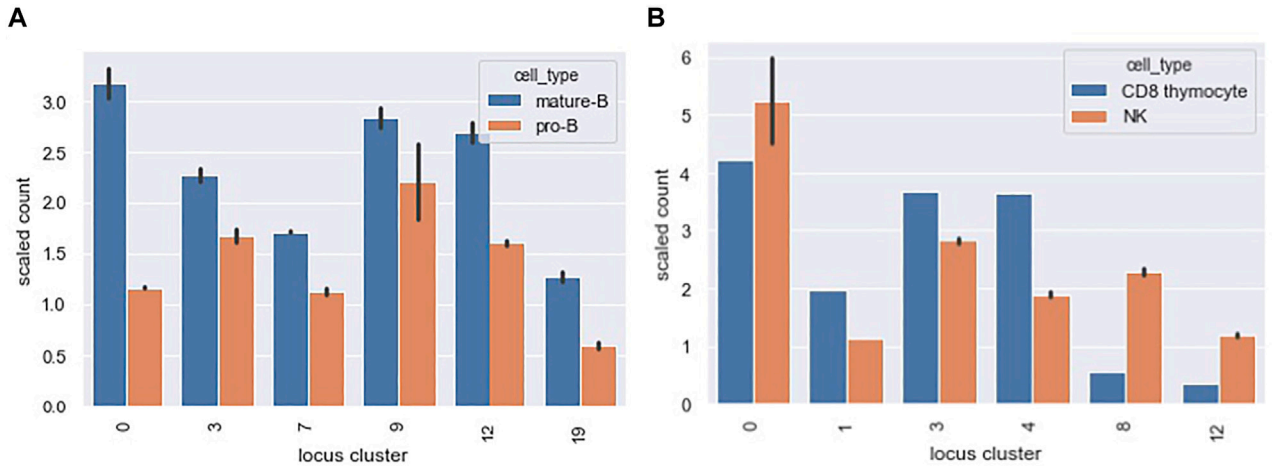

FIGURE 10 | Distribution of PAX5 and EOMES ChIPseq peaks across locus clusters. The bar graphs show a scaled count (*y*-axis) of the number of loci in a locus cluster (*x*-axis) that contained a ChIPseq peak for **(A)** PAX5 and **(B)** EOMES ChIPseq datasets. Scaled counts took into account the different sizes of our 20 locus clusters. A scaled count above 1 represents an enrichment of peaks above an expectation under a uniform null. The locus clusters enriched for PAX5 ChIPseq peaks had a statistically significant association with accessibility in pro-B and mature B cell types and we found an analogous association for EOMES peaks, see text for details.

clusters 0, 3, 7, 9 and 12 (see the B cell columns of **Figure 8**). All of these locus clusters had enriched PAX5 peak counts, and only locus cluster 19 had an enriched count without associated accessibility in B cells. Overall, there was a statistically significant association between PAX5 peaks and accessibility (*p*-value 0.001, hypergeometric test). The pro-B cell type falls within our progenitor (pro) cell type cluster. Progenitor cells are present in 10 of the accessibility clusters and PAX5 peaks had enriched counts in four of these, reflecting a marginally significant association between peaks and accessibility (*p*-value 0.08). Interestingly, in locus cluster 19, PAX5 peak counts were enriched in the mature B cell ChIPseq but not the pro-B cell ChIPseq. We would expect the reverse because accessibility cluster 19 is specific for progenitor cells. This deviation might reflect differences in cell state between the ChIPseq studies and Yoshida et al.

Figure 10B shows analogous results for EOMES ChIPseq peaks in the NK and CD8 thymocyte cell types. Our motif analysis showed enrichment of EOMES in co-clusters formed by locus clusters 1 and 8. Locus cluster 1 and 8 had enriched EOMES ChIPseq peak counts for both NK and CD8 thymocyte cell types and only the NK cell type, respectively. In-line with these results, early T cells - of which CD8 thymocytes are a member - and NK cells are both accessible in locus cluster 1 but only NK cells are accessible in locus cluster 8. Both NK and CD8 thymocytes had a statistically significant association between their accessibility co-clusters and the locus clusters at which EOMES peak counts were enriched (*p*-values 0.0003 and 0.001 respectively). For both PAX5 and EOMES ChIPseq datasets, we also calculated the fraction of loci with peaks that were cell specific. Recall, cell specific loci were accessible in 2 or less of the Yoshida et al. cell types we considered. For PAX5 and EOMES, roughly 15 and 5% of loci with peaks were cell specific, respectively. In contrast, roughly 80 and 70% of loci with peaks fell within one of our 20 locus clusters, demonstrating that accessibility patterns across multiple cell types capture the dominant portion of TF binding, at least for EOMES and PAX5 and for the accessible loci we consider.

4 DISCUSSION

The recently published ATACseq dataset of Yoshida et al. provides a valuable resource through which to investigate patterns of chromatin accessibility across immune cell types. Here, we used this dataset to investigate the degree to which co-clustering of genomic loci and cell types can capture and describe patterns of chromatin accessibility. Some genomic loci were accessible in only 1 or 2 cell types, and we found that roughly half of accessible loci over immune cell types were of this type, in line with previous analyses of datasets encompassing non-hematopoietic cell types (Song et al., 2011). The other half of the accessible loci, which are accessible in multiple cell types, were the focus of our study. We found that essentially all of these loci, >95%, can be grouped into 20 locus clusters. Within each locus cluster, the cell types showed roughly two states of accessibility reflecting a relatively high and low percentage of the loci that were accessible, respectively. For example, in locus cluster 1 which was composed of roughly 9,000 loci, we found that in most T cell types, roughly 60% of the loci were accessible while in most other cell types roughly 0–10% were accessible. The dichotomy between cell types was more extreme in some locus clusters. For example, in locus cluster 8, we found all loci were accessible in NK cells but not in any other cell type. Ideally, in terms of cluster coherence, locus clustering might lead 100% or 0% of loci being accessible within a cell type. Certainly some of the locus cluster incoherence we see results from noise in the ATACseq workflow. But chromatin accessibility is not static, and some portion of the incoherence may reflect stochasticity in nucleosome positioning,

binding of TF complexes, or other dynamic effects (Natoli et al., 2011; Wang et al., 2012). It could also be that more locus clusters would result in greater coherence. Changing the graph we used as input to the Louvain clustering did not lead to more coherent locus clusters, but further work is needed to explore this issue.

Given the locus clusters, we found that a modest number of cell type clusters could capture a large fraction of the variation in accessibility associated with cell types. Using 12 cell type clusters, we were able to capture 80% of the cell type associated variation. Further, the cell type clusters we formed reflected coherent phenotypes as defined by the hematopoietic differentiation tree. When we formed cell type clusters using methods that were insensitive to the differentiation tree, the fraction of variance captured did not improve. Our cell type clustering extends the results of Lara-Astiaso et al. (2014) describing an association between accessibility and hematopoietic cell type.

Ultimately, we characterize chromatin accessibility to better understand cellular regulation. In particular, chromatin accessibility is strongly associated with TF binding (Klemm et al., 2019). Using both TF motif analysis and existing ChIPseq studies, we have shown that TF binding patterns associated with our co-clusters. Importantly, our results show that some TFs act across co-clusters. For example, we found that PAX5 motifs are enriched in two sets of loci. One set is accessible only in B cells while the other is accessible in both B cells and some myeloid cell types. Our ChIPseq analysis confirmed that PAX5 bound to both types of loci in B cells. Myeloid cell types do not express PAX5, at least at homeostasis, but the loci that are accessible in myeloid cell types and that are bound by PAX5 in B cells may regulate myeloid cells through other TFs or under non-homeostatic conditions. We found a different binding pattern for the transcription factor EOMES. In our ChIPseq analysis, we found that EOMES bound loci in NK and T cells, but that loci bound in NK cells were inaccessible in T cells and vice-versa. Our motif analysis suggests that many TF act across several co-clusters, in a manner similar to PAX5 and EOMES. These results suggest that analyzing patterns of chromatin accessibility through co-clustering may be essential in understanding the overlap and divergence of regulation in different cell types.

From a computational viewpoint, our work provided two insights. First, we found that co-clustering, rather than biclustering, provided a relatively stable and scalable means of analyzing ATACseq datasets across many cell types. We initially attempted a biclustering approach but found that solutions depended on starting conditions of the algorithm, that the algorithms did not scale well to the large number of accessible loci, and that interpretation was difficult. Second, we developed a novel graph based clustering algorithm to account for the hematopoietic differentiation tree. In this context, the most significant insight is the form that we assumed for the clusters. Initially, we formed clusters as connected components of the differentiation tree, but we found that the clusters created did not approximate the Yoshida et al. data well. Certain cell types have accessibility patterns that are different than the patterns of their parent cell type and connected components force the parent to be included with the children. Accounting for this effect vastly improved the fit of our clustering and points to the need for clustering approaches that account for the specifics of differentiation biology.

Our analysis involved several computational choices that may affect our results. We made binary calls of whether a locus was accessible or inaccessible. Using a continuous measure may better reflect chromatin accessibility biology and may affect our clustering results. From a computational perspective, we depended on the binary nature of the data to construct the input graph to the Louvain algorithm. We have also not explored an iterative co-clustering approach, e.g (Cheng and Church, 2000). Our two-step clustering of loci followed by cell types makes our approach simple and scalable, but an iteration may lead to better results. Biologically, we are limited to the cell types given in the Yoshida et al. dataset and our assumption of a particular form to the differentiation tree. More generally, our algorithm depends on the input of a differentiation tree, which is not typically available outside of well characterized cell types such as hematopoietic cells. However, tools exist to form differentiation trees from samples across cell types, most commonly from RNAseq data, e.g. monocle, and such tools could be applied upstream of our algorithm and analysis. Further, while our algorithm currently requires a tree structure, extension to arbitrary graphs is possible and represents a direction for future work.

Overall, we have demonstrated a co-clustering approach that quantifies and delineates the association between chromatin accessibility and immune cell type. Our results provide a context in which to assess chromatin accessibility of other immune cell types. With the increased application of single cell ATACseq and the likely generation of even larger bulk ATACseq datasets, computational approaches to characterize chromatin accessibility patterns over an increasingly broad set of hematopoietic cell types will be needed.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

TG and SL performed the analysis and wrote the manuscript. TG, NS, and SL conceived of the algorithms. NS and SL conceived the project.

ACKNOWLEDGMENTS

We thank the ImmGen consortium, without which this work would not have been possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.707117/full#supplementary-material

REFERENCES

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for Functional Genomics Data Sets-Update. *Nucleic Acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193

Biesinger, J., Wang, Y., and Xie, X. (2013). Discovering and Mapping Chromatin States Using a Tree Hidden Markov Model. *BMC Bioinf.* 14 Suppl 5, S4. doi:10.1186/1471-2105-14-S5-S4

Blondel, V., Guillaume, J-L., Lambiotte, R., and Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *J. Stat. Mech. Theor. Exp.* 2008, P10008. doi:10.1088/1742-5468/2008/10/p10008

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position. *Nat. Methods* 10, 1213–1218. doi:10.1038/nmeth.2688

Calderon, D., Nguyen, M. L. T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., et al. (2019). Landscape of Stimulation-Responsive Chromatin across Diverse Human Immune Cells. *Nat. Genet.* 51, 1494–1505. doi:10.1038/s41588-019-0505-9

Cheng, Y., and Church, G. M. (2000). "Biclustering of Expression Data," in ISMB. International Conference on Intelligent Systems for Molecular Biology, August 2000., San Diego, CA, 93–103.

Collins, P. L., Cella, M., Porter, S. I., Li, S., Gurewitz, G. L., Hong, H. S., et al. (2019). Gene Regulatory Programs Conferring Phenotypic Identities to Human NK Cells. *Cell* 176, 348–360.e12. doi:10.1016/j.cell.2018.11.045

Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., et al. (2016). Lineage-specific and Single-Cell Chromatin Accessibility Charts Human Hematopoiesis and Leukemia Evolution. *Nat. Genet.* 48, 1193–1203. doi:10.1038/ng.3646

Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., et al. (2018). A Single-Cell Atlas of *In Vivo* Mammalian Chromatin Accessibility. *Cell* 174, 1309–1324.e18. doi:10.1016/j.cell.2018.06.052

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2018). The Encyclopedia of DNA Elements (ENCODE): Data portal Update. *Nucleic Acids Res.* 46, D794–D801. doi:10.1093/nar/gkx1081

Eisen, M. B., Spellman, P., Brown, P., and Botstein, D. (1999). Cluster Analysis and Display of Genome-Wide Expression Patterns MICHAEL. 96 *PNAS*, 12930–12933.

Gordon, S. M., Chaix, J., Rupp, L. J., Wu, J., Madera, S., Sun, J. C., et al. (2012). The Transcription Factors T-Bet and Eomes Control Key Checkpoints of Natural Killer Cell Maturation. *Immunity* 36, 55–67. doi:10.1016/j.immuni.2011.11.016

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cel.* 38, 576–589. doi:10.1016/j.molcel.2010.05.004

Ho, I.-C., Tai, T.-S., and Pai, S.-Y. (2009). GATA3 and the T-Cell Lineage: Essential Functions before and after T-Helper-2-Cell Differentiation. *Nat. Rev. Immunol.* 9, 125–135. doi:10.1038/nri2476

Horcher, M., Souabni, A., and Busslinger, M. (2001). Pax5/BSAP Maintains the Identity of B Cells in Late B Lymphopoiesis. *Immunity* 14, 779–790. doi:10.1016/s1074-7613(01)00153-4

Huang, J., Liu, X., Li, D., Shao, Z., Cao, H., Zhang, Y., et al. (2016). Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. *Dev. Cel.* 36, 9–23. doi:10.1016/j.devcel.2015.12.014

Istaces, N., Splittgerber, M., Lima Silva, V., Nguyen, M., Thomas, S., Le, A., et al. (2019). EOMES Interacts with RUNX3 and BRG1 to Promote Innate Memory Cell Formation through Epigenetic Reprogramming. *Nat. Commun.* 10, 3306. doi:10.1038/s41467-019-11233-6

Klemm, S. L., Shipony, Z., and Greenleaf, W. J. (2019). Chromatin Accessibility and the Regulatory Epigenome. *Nat. Rev. Genet.* 20, 207–220. doi:10.1038/s41576-018-0089-8

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. *Genome Biol.* 10, R25. doi:10.1186/gb-2009-10-3-r25

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D. A., David, E., et al. (2014). Chromatin State Dynamics during Blood Formation. *Science* 345, 943–949. doi:10.1126/science.1256271

Lau, C. M., Adams, N. M., Geary, C. D., Weizman, O.-E., Rapp, M., Pritykin, Y., et al. (2018). Epigenetic Control of Innate and Adaptive Immune Memory. *Nat. Immunol.* 19, 963–972. doi:10.1038/s41590-018-0176-1

Lazzeroni, L., and Owen, A. (2002). Plaid Models for Gene Expression Data. *Stat. Sin.* 12, 31–46.

Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (2011). Measuring Reproducibility of High-Throughput Experiments. *Ann. Appl. Stat.* 5, 1752–1779. doi:10.1214/11-aoas466

Lizio, M., Harshbarger, J., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., et al. (2015). Gateways to the FANTOM5 Promoter Level Mammalian Expression Atlas. *Genome Biol.* 16, 1–14. doi:10.1186/s13059-014-0560-6

Natoli, G., Ghisletti, S., and Barozzi, I. (2011). The Genomic Landscapes of Inflammation. *Genes Dev.* 25, 101–106. doi:10.1101/gad.2018811

Nechanitzky, R., Akbas, D., Scherer, S., Györy, I., Hoyler, T., Ramamoorthy, S., et al. (2013). Transcription Factor EBF1 Is Essential for the Maintenance of B Cell Identity and Prevention of Alternative Fates in Committed Cells. *Nat. Immunol.* 14, 867–875. doi:10.1038/ni.2641

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular Portraits of Human Breast Tumours. *Nature* 406, 747–752. doi:10.1038/35021093

Pontes, B., Giráldez, R., and Aguilar-Ruiz, J. S. (2015). Biclustering on Expression Data: A Review. *J. Biomed. Inform.* 57, 163–180. doi:10.1016/j.jbi.2015.06.028

Revilla-I-Domingo, R., Bilic, I., Vilagos, B., Tagoh, H., Ebert, A., Tamir, I. M., et al. (2012). The B-Cell Identity Factor Pax5 Regulates Distinct Transcriptional Programmes in Early and Late B Lymphopoiesis. *EMBO J.* 31, 3130–3146. doi:10.1038/emboj.2012.155

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A Comprehensive Evaluation of Module Detection Methods for Gene Expression Data. *Nat. Commun.* 9. 1090. doi:10.1038/s41467-018-03424-4

Schep, A. N., Wu, B., Buenrostro, J. D., and Greenleaf, W. J. (2017). ChromVAR: Inferring Transcription-Factor-Associated Accessibility from Single-Cell Epigenomic Data. *Nat. Methods* 14, 975–978. doi:10.1038/nmeth.4401

Sciumè, G., Mikami, Y., Jankovic, D., Nagashima, H., Villarino, A. V., Morrison, T., et al. (2020). Rapid Enhancer Remodeling and Transcription Factor Repurposing Enable High Magnitude Gene Induction upon Acute Activation of NK Cells. *Immunity* 53, 745–758.e4. doi:10.1016/j.immuni.2020.09.008

Scott-Browne, J. P., López-Moyado, I. F., Trifari, S., Wong, V., Chavez, L., Rao, A., et al. (2016). Dynamic Changes in Chromatin Accessibility Occur in CD8 + T Cells Responding to Viral Infection. *Immunity* 45, 1327–1340. doi:10.1016/j.immuni.2016.10.028

Seita, J., and Weissman, I. L. (2010). Hematopoietic Stem Cell: Self-Renewal versus Differentiation. *Wires Syst. Biol. Med.* 2, 640–653. doi:10.1002/wsbm.86

Shabalin, B. Y. A. A., Weigman, V. J., Perou, C. M., and Nobel, A. B. (2016). Finding Large Average Submatrices in High Dimensional Data. *Ann. Appl. Stat.* 3 (3), 985–1012. Publishe 3, 985–1012. doi:10.1214/09-AOAS239

Shay, T., and Kang, J. (2013). Immunological Genome Project and Systems Immunology. *Trends Immunol.* 34, 602–609. doi:10.1016/j.it.2013.03.004

Shea, J. J. O., Paul, W. E., and Cells, C. D. T. (2010). Mechanisms Underlying Lineage Commitment and Plasticity of Helper CD4+ T Cells. *Science* 327, 1098–1102. doi:10.1126/science.1178334

Sohn, K.-A., Ho, J. W. K., Djordjevic, D., Jeong, H.-h., Park, P. J., and Kim, J. H. (2015). HiHMM: Bayesian Non-parametric Joint Inference of Chromatin State Maps. *Bioinformatics* 31, 2066–2074. doi:10.1093/bioinformatics/btv117

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B.-K., et al. (2011). Open Chromatin Defined by DNaseI and FAIRE Identifies Regulatory Elements that Shape Cell-type Identity. *Genome Res.* 21, 1757–1767. doi:10.1101/gr.121541.111

Sun, S., and Barreiro, L. B. (2020). The Epigenetically-Encoded Memory of the Innate Immune System. *Curr. Opin. Immunol.* 65, 7–13. doi:10.1016/j.coi.2020.02.002

Wagner, J. A., Wong, P., Schappe, T., Berrien-Elliott, M. M., Cubitt, C., Jaeger, N., et al. (2020). Stage-Specific Requirement for Eomes in Mature NK Cell Homeostasis and Cytotoxicity. *Cel Rep.* 31, 107720. doi:10.1016/j.celrep.2020.107720

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., et al. (2012). Sequence Features and Chromatin Structure Around the Genomic Regions Bound by 119 Human Transcription Factors. *Genome Res.* 22, 1798–1812. doi:10.1101/gr.139105.112

Xiang, G., Keller, C. A., Heuston, E., Giardine, B. M., An, L., Wixom, A. Q., et al. (2020). An Integrative View of the Regulatory and Transcriptional Landscapes in Mouse Hematopoiesis. *Genome Res.* 30, 472–484. doi:10.1101/gr.255760.119

Yoshida, H., Lareau, C. A., Ramirez, R. N., Rose, S. A., Maier, B., Wroblewska, A., et al. (2019). The Cis-Regulatory Atlas of the Mouse Immune System. *Cell* 176, 897–912.e20. doi:10.1016/j.cell.2018.12.036

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137. doi:10.1186/gb-2008-9-9-r137

Zhang, Y., An, L., Yue, F., and Hardison, R. C. (2016). Jointly Characterizing Epigenetic Dynamics across Multiple Human Cell Types. *Nucleic Acids Res.* 44, 6721–6731. doi:10.1093/nar/gkw278

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.