NON PARAMETRIC UNSUPERVISED CLUSTERING OF CHIP ENRICHMENT REGIONS PROVIDES ISOLATION VECTORS FOR DIFFERENTIAL FUNCTIONAL ANALYSIS

By
Alexander Griffith Candidate, B.Eng
August 2016

A Thesis
Submitted to the School of Graduate Studies and Research
in partial fulfillment of the requirements
for the degree of
Master of Applied Science in Biomedical Enginnering [1]

[1]The M.ASc. Program is a joint program with Carleton University, administered by the Ottawa Department of Electrical Engineering

Abstract

Gene transcription rates are influenced by proteins, known as Transcription Factors (TFs), that interact with DNA. The locations of TFs on the genome directly influence gene expression and the functional characteristics of a cell. TF binding locations can be estimated for entire genomes using high throughput chromatin immunoprecipitation sequencing (ChIP-Seq). While the analysis of ChIP-Seq binding locations is standardized for a single experiment, complications arise when data sets, taken from different labs and experimental conditions, are combined. In this thesis, I present my method for the simultaneous comparison of multiple ChIP-Seq data sets. My method of comparing multiple ChIP-Seq data sets extends the analysis of a single data set through the addition of two stages, a combination stage, and an extraction stage. Typically, one of two approaches are used to combine information from multiple datasets. Either estimated binding sites are extracted from each dataset and then combined (e.g. by various intersections or unions) or the "raw" genomic signals are analyzed by clustering or dimensionality reduction methods. Both approaches have strengths, but also substantial drawbacks. The method presented here relies both on estimating the binding sites and comparing the raw genomic signals between data sets. Once the binding locations have been found, the first step in the combination stage is to define an alternate feature space (AFS). The AFS is the union of all binding locations determined for all data sets. The AFS represents a subset of the genome that is likely to have TF binding in any condition where the protein is active. Once the AFS is defined, the read density is determined from the raw genomic signal of each of the data sets. The density is determined for all locations in the AFS resulting

in a unified density matrix (UDM). The UDM is the final product of the combination stage of the analysis. After the data sets are homogenized into the UDM, the extraction stage is applied to the matrix. The extraction stage consists of applying machine learning techniques and other methods used to analyze the raw genomic signal, to help elucidate underlying similarities and differences between the data sets. I applied this method to the binding locations of the TF TAL1 across 22 ChIP-Seq data sets from the hematopoietic and endothelial lineages. Once the UDM had been generated and normalized, using quantile normalization, hierarchical clustering and principle component analysis (PCA) were applied. Clusters, formed by hematopoietic stem cells (HSCs), Erythroid, and T-cell acute lymphoblastic leukemia (T-ALL), were found using hierarchical clustering. The principle components (PCs) of the UDM provided weights for each peak. Using those weights I could separate groups of cellular conditions including T-ALL, Erythroid, HSC, and Endothelial Colony Forming Cells (ECFCs.) The weights also provided a quantitative measure of importance for each peak in the AFS based on how much weight they provided towards the group of interest. Functional analysis techniques, including de novo motif search and Gene Ontology, were applied to the peak partitions defined using the PCs. Motifs that were enriched in the T-ALL TAL1 partition, and not the Erythroid, were annotated and found to be similar to those that had previously been published, including Runx1 motif and a preference for the CC Ebox (CACCTG). In addition to finding the CC Ebox in T-ALL, I also show that it does not form a composite motif with GATA, indicating an alternative mechanism for the binding of TAL1 in T-ALL. This thesis establishes that heterogeneous collections of ChIP-Seq datasets, from multiple labs and experimental conditions, can be meaningfully combined, and provides an algorithmic template for doing so.

Acknowledgements

I would like to thank both of my supervisors, Dr. Marjorie Brand, and Dr. Theodore Perkins, for their support. In addition I would like to thank the University of Ottawa, the QEII-GSST awards program, and the Ottawa Hospital Research Institute which provided funding for my position over the 2013-2015 period.

Contents

Chapter 1

Overview

1.1 Background

Transcription Factors (TFs) are proteins that have the ability to bind to specific locations on the DNA based on nucleotide sequence and chromatin conformation. The spatial binding potential of TFs, across the genome, influences gene expression levels, and controls cell fate. The binding potential of a TF will be influenced by changes in cellular state, such as changes in the expression levels of co-factors, or changes in chromatin accessibility [1, 2]. The interdependencies between TFs and cellular conditions permit for a high variability of cell functions while maintaining stable cell fates.

In select cellular conditions, specific TFs have critical roles, either in driving changes in cellular conditions, or maintaining a stable cell fate. The basic helix loop helix (bHLH) protein TAL1 is an example of a TF with a critical role in several lineages, including the hematopoietic and the endothelial. For instance, TAL1 is essential for the migration and adherence of endothelial progenitor cells [3]. Within the hematopoietic lineage, TAL1 is essential for the specification, survival and competence of hematopoietic stem cells (HSCs) and promotes cell differentiation towards the erythroid and megakaryocytic fates [4]. In contrast, TAL1 is not expressed in T-cells beyond very early progenitors, and the ectopic expression of TAL1 in T-cell

progenitors inhibits T-cell differentiation, which in turn leads to leukemic transformation and T-cell Acute Lymphoblastic Leukemia (T-ALL) [5]. As such, not only does TAL1 have a critical role in regulating different cell types within the hematopoietic and endothelial lineages, its functions also vary between cellular conditions.

1.2 Research questions

While it is clear that the distinct functions of TAL1 in different cell environments are mediated through differential genomic binding, the mechanistic basis underlying this differential binding as well as its consequences on TAL1-mediated functions in cell fate regulation are currently unclear. Important unresolved questions include the degree to which TAL1 binding locations are shared between conditions and the effect of unique binding locations on TAL1 function in distinct cellular conditions. Such questions are critically important to understand TAL1s role in controlling cell fate and differentiation in health and disease. Studying the roles of TAL1 in different cell environments requires a systematic comparison of TAL1 binding between multiple cell contexts.

The differential binding of TAL1 can be explained in part by variations in complex members. Structural and molecular studies have shown that the binding of TAL1 to DNA is highly dependent on its interaction with other proteins [2]. Furthermore, previous studies comparing two cellular environments have identified distinct types of E-Boxes and different composite DNA motifs under TAL1 bound sites in different cellular environments, suggesting that TAL1 DNA binding specificity may change between cell contexts [3]. This variation in binding applies to both the TAL1 protein itself and the TAL1 complex, based on the presence of members such as GATA1 and RUNX1 [5]. Despite these observed variations in binding preference, a systematic study comparing DNA motifs between all cell types has not been performed and as such, TAL1 DNA binding specificity between cell types remains unclear.

Chromatin immunoprecipitation followed by high throughput sequencing (ChIP-Seq) has been used extensively to analyze TAL1 genomic binding in multiple cellular environments, providing a novel opportunity for the systematic comparison of TAL1

binding in a wide variety of conditions. However, different labs have generated the datasets over several years using different protocols and sequencing platforms. This has resulted in variations in both the depth of sequencing and the non-specific signal contribution, i.e. background signal. While the differences in sequencing depth can be normalized between cellular conditions using standard approaches, differences in background signal cannot [6, 7]. This limitation makes it difficult to directly compare the available datasets.

Other issues that have so far prevented a systematic comparative analysis of TAL1 binding between different cellular environments include differences in the efficacy and specificity of the antibodies used by the different labs, variability in chromatin accessibility between each cellular environment, and the use of different controls (e.g. input vs. mock ChIP). In this thesis I investigate two current methods of comparing ChIP-Seq binding signal, overlap analysis and global analysis. I show their limitations when comparing datasets from disparate cellular conditions, and propose an alternative approach.

1.3 Goal, Objectives and Hypothesis

Our long-term goal is to better understand the common versus specific functions of TFs in multiple cell environments.

The objectives of the thesis were to design a ChIP-Seq data analysis method that allows for the systematic comparison of TF binding between multiple cell environments, and to use it for extracting biological information from TAL1 binding in different cell environments.

Our hypothesis was that a systematic comparative analysis of TAL1 binding in multiple cellular contexts would allow a better understanding of TAL1 DNA binding specificity, and a better understanding of TAL1s common and specific functions in each cellular environment.

1.4 Aims and Contributions

This thesis consisted of two principle aims. The first was to identify the limitations of current methods used to compare ChIP-Seq data sets between cellular conditions and to develop a method that overcomes these limitations. The second aim was to apply this method in the context of elucidating the role of TAL1 in the hematopoietic and endothelial lineages.

1.4.1 Contribution 1: Determine the limitations of the current approaches to comparing ChIP-Seq data and develop a method that overcomes these limitations

Rationale

Current approaches of comparing data sets are usually comprised of 2 steps: calling ChIP-Seq peaks separately using a predefined background (i.e. input or mock ChIP), followed by determining the overlap of peaks between conditions, to be used as a measure of similarity. However, this approach is limited in several respects.

1.) Stringency thresholds systematically underrepresent overlapping regions (Bias)

Due to the number of possible binding locations, and the variability of the background signal, peak detection often relies on the application of stringent thresholds, in order to limit the false discovery rate [8]. A stringent threshold helps to ensure that only the regions with the greatest binding potential are assessed. When determining the overlap of two sets of peaks, each determined independently, those that overlap must pass the statistical test twice. This double testing increases the stringency required for the peak to be found in the overlapping category, compared to the unique [6]. As the number of data sets being overlapped increases, the stringency of those peaks that are found to be present in all sets increases. This systemic underrepresentation of the number of overlapping peaks limits the utility of using overlap as a measure of cell condition similarity.

2.) Overlapping large numbers of data sets provides little insight into

the overall relations between several cellular conditions (Scale)

As the number of data sets being compared increases, so does the number of possible overlapping combinations. Currently, to overcome this limitation data sets are grouped based on a previously determined similarity and subsequently overlapped. This poses two issues, the first being that researchers apply bias to the investigation by presuming which data sets will have differential binding, and the second being that this method determines the distance between groups but not between data sets. While there may be inherent differences in binding between different data sets when several data sets are amalgamated into a group, based on previous evidence, the differences may be masked. As the number of data sets increases, the aggregate comparison becomes less specific, limiting the ability to scale to arbitrarily large numbers of data sets.

3.) Peak sets lose their quantifiable measure when overlapped (Quantifiability)

When an individual peak set is analyzed, a single measure, such as p-value or pile up count, may be assigned as an interval measure of peak "strength", "quality", or "importance". An interval measure is quantifiable, it allows for a distance between peak values to be determined, for the peaks to be ranked, and for the peaks to be categorized. In contrast, the overlap of two or more peak sets results in the categorization of peaks based on which data set they come from. This categorization provides a qualitative-only nominal measure of peak value. As such, the new value cannot be used to rank peaks or determine the distance between their values. Thus, if we simply overlap the peak sets, the quantifiable information is lost.

Method Description

To address these limitations, I designed a new comparative method with two distinct steps. First, the peak sets were identified for each data set and combined into a single unified set. Then, natural underlying clusters were extracted from the unified set for downstream functional analysis. Further details on the method are provided in Chapter 3.

Each of the ChIP-Seq data sets were individually aligned and their enriched regions

(peaks) were identified. The peaks from the individual data sets were combined into a single unified set. This set of unified peaks comprised the subset of the genome that the TF was most likely to bind to. This subset is referred to hereafter as the Alternate Feature Space (AFS).

Once the AFS was defined, the read pile up count for each data set was then determined under each peak within the set. The matrix of read pile up counts was normalized by peak width to form the unified density matrix (UDM). The UDM, containing information from each raw read data set for each peak in the unified peak set, represented the inclusion of quantifiable information from each data set. Quantile normalization and Principle Component analysis (PCA) were then applied to the UDM, in order to identify naturally occurring subsets in the unified data set. PCA was applied to the matrix such that it resulted in as many principle components (PC) as their were data sets being contributed, and each peak in the AFS was represented by a weight in each PC. By taking the dot product of the quantile normalized UDM and the PCs of interest, the weighted positions of the data sets could be visualized.

The PCA of the UDM concentrated ChIP-Seq signal variation into a handful of dimensions for easier visualization. Visual inspection of the lower dimensions could be used to identify PCs that separated data sets of interest. Alternatively, k-means clustering could be applied to a subset of the dimensions to determine if the data sets could be separated along that subsets.

The PC vectors, since they provided a weight for each peak along each dimension, could be used to identify which peaks contributed most to the separation of data sets. Since the PC vectors fell on a range, they were used determine the distance between peak values, rank peaks, and partition the peaks, i.e the PCs were a quantifiable interval measure. I used the PC vectors that separated data sets of interest, and defined the peaks that had weights that fell at the extremes of the PC vectors as those most important to the conditions they separated.

Using the PC vectors as a measure of cellular condition, coupled the identification of subsets and the value assigned to the peaks in each condition. In other words, the analysis no longer relied on predefined relations between data sets. This allowed the data to be naturally partitioned, and for the importance of each peak in each

partition to be retained in a quantifiable fashion. Additionally, the partitions that were found were used as an alternative to predefined relations, to help remove bias from the comparison of a large number of data sets.

By creating a unified peak set, the dependency of stringency on the number of data sets being overlapped was overcome. Rather than having the number of times a peak is tested related to the number of sets being intersected, each peak included in the unified set had to pass only a single test. The stringency of the single test may be altered based on the needs of the downstream analysis. Thus, the comparison of two or more data sets was no longer biased based on the degree of overlap.

The coupling of partitioning and peak ranking, along with the mitigation of the impact of cut off bias, permitted for the systematic investigation of the role of specific TFs across several cellular conditions. As such, the method permitted for the novel large-scale comparison of TAL1 binding across the hematopoietic and endothelial lineages.

1.4.2 Contribution 2: Compare TAL1 binding between conditions, and assess the functional differences between conditions, based on binding locations.

Using the method designed in first contribution, we compared TAL1 binding in 12 cellular conditions (22 datasets) with the aim of elucidating the role of TAL1 in 4 different contexts, T-ALL, erythroid, endothelial colony forming cells (ECFCs), and hematopoietic stem cells (HSCs).

First, I found that clustering based on TAL1 binding alone is sufficient to recreate the known hierarchical differentiation relationships between the different cell types within the hematopoietic and endothelial lineages. This demonstrates that the binding of TAL1 is specific to each cellular environment. Additionally, I found that ectopic TAL1 binding in leukemic cells contains a completely unique set of binding locations. This set differs from TAL1 binding in hematopoietic stem cells (HSCs) and erythroid cells. This indicates that the role of TAL1 in leukemic cells is unique, not solely a

remnant of its function in HSCs or healthy differentiated hematopoietic cells expressing TAL1. Another interesting finding was that megakaryocytes cluster closely to the HSCs, which is consistent with previous comparisons of the two hematopoietic cells [9]. Taken together, these results indicate that our new comparative approach provides biologically relevant separation of TF binding locations. Our approach provides a useful means to identify the set of TAL1 binding sites that are the most important for defining the unique function(s) of TAL1 in each specific cell environment (i.e. cell type-defining TAL1 binding sites).

The PCA of the UDM separated the 12 cellular conditions into 4 contexts along 3 eigenvectors. The four contexts were T-ALL, Erythroid, ECFC, and HSC. For the purpose of comparing the role of TAL1, the Megakaryocyte and HSCs were defined as a single context (HSC). This grouping was decided since both conditions grouped in the PCA and the hierarchical clustering.

I used each of the eigenvectors as a weighting vector to partition the peaks that contribute most to each condition. In addition the peaks that fell within 0.25 standard deviation of the mean of each eigenvector were included in a Central data set and used as background for the downstream functional analysis. With peaks that best separate TAL1 binding in each specific cell environment isolated, I then identified the DNA binding specificities that provide TAL1 with its unique functions in each cell environment. The specificities include:

1. Motifs specific to each condition

2. The E-protein motif binding preference shift between conditions

3. The change in importance of composite motif preferred distances between conditions

4. The genes associated with each condition and the resulting GO analysis

These specificities helped to characterize the role of TAL1 in each condition. They also showed that using the eigenvectors of the quantile normalized UDM provided subsets of the peaks that had biological relevance.

1.5 Content

I have broken my thesis into six sections. The background material, seen in Chapter 2, includes information on TAL1, the hematopoietic and endothelial lineages (Section 2.2), ChIP-Seq (Section 2.6) and current available methods used to compare ChIP-Seq data sets (Section 2.7). I outline previous methods described by Stark and Hardision, which this thesis builds on, in Section 2.7.4 [6, 10]. The methodology is presented in Chapter 3. The methodology was split into two key sections. Section 3.4 describes the methods for combining the data into a unified form, and Section 3.5 describes my approach for extracting biologically relevant information. In addition to the methodology, this chapter presents variations on the approach used for validation seen in Section 3.6 and the functional analysis I applied in Section 3.7. The results are presented in two chapters. The technical results can be seen in Chapter 4 and the biological in Chapter 5. I present my discussion in Chapter 6 and the conclusion in Chapter 7.

Chapter 2

Background

2.1 Overview

In this chapter I present the biologically relevant background in Sections 2.2 to 2.5. The technical background, including a description of the currently available methods can be found in Sections 2.6 and 2.7. This section presents an overview of the material that will be provided in the body of this Chapter.

The cells that make up the blood differentiate from the hemogenic endothelium, a common precursor for endothelial progenitor cells and hematopoietic stem cells (HSCs) [11]. Despite sharing a common precursor, the functionality of the terminally differentiated conditions varies significantly. These variations are in large part based of TF mediated regulation of cell fate. The differentiation of the endothelial and hematopoietic lineages is mediated by TFs including PU.1, GATA, RUNX1 and TAL1 [5]. More details on the hematopoietic and endothelial lineages and the TFs that influence cell fate determination can be found in Sections 2.2 and 2.3.

TAL1 is of particular interest for investigation between cellular conditions in the hematopoietic and endothelial lineages. Once dimerized with an E-protein, TAL1 has the ability to interact with the DNA directly, through its bHLH DNA binding domain [12]. Alternatively, TAL1 may interact indirectly through recruitment by other TFs as part of a complex such as GATA [13]. Apart from being recruited directly by these other TFs, the spatial binding probability of TAL1, i.e. the likelihood that TAL1

will bind to a location on the genome, is influenced by other co-factors present in the complex [2]. It is this variation of spatial binding probability based on the presence of other TFs, which makes TAL1 an interesting protein to investigate differentially between cellular conditions.

On top of changes in binding potential between conditions, TAL1 has also been shown to have different functional roles in different cellular conditions. TAL1 is critical in the formation of HSCs and in driving differentiation to erythroid and megakaryocyte cell fates [13]. In contrast, when ectopically expressed in T-Cells, TAL1 stalls differentiation, which leads to T-Cell Acute Lymphoblastic Leukemia (T-ALL) [5]. Observing the changes in binding probability between conditions could elucidate changes in the role of TAL1 between conditions. More details on TAL1 and its role in healthy and diseased states can be found in Section 2.4.

ChIP-Seq experiments provide insight into how TAL1 binds across the genome in different cellular conditions. In a single data set there are statistical methods of determining which regions are enriched compared to a control [8]. These enriched regions are referred to as peaks and are often used as a representation of TF binding locations. Current comparison of ChIP-Seq data between conditions relies on the overlap of two sets of peaks. The size of the overlapping proportion of peaks is used to determine the similarity of the TF's role in the two conditions. This overlap method has several limitations, including, statistical bias, limited ability to scale, and loss of quantification. These limitations have prevented the systematic investigation of large numbers of ChIP-Seq data set from several cellular conditions [6]. A summary of the ChIP method, accompanied by the current analysis methods available for the comparison of multiple data sets is presented in Section 2.6 and 2.7 respectively. More details on the currently available methodologies can be found in Section 2.7.4.

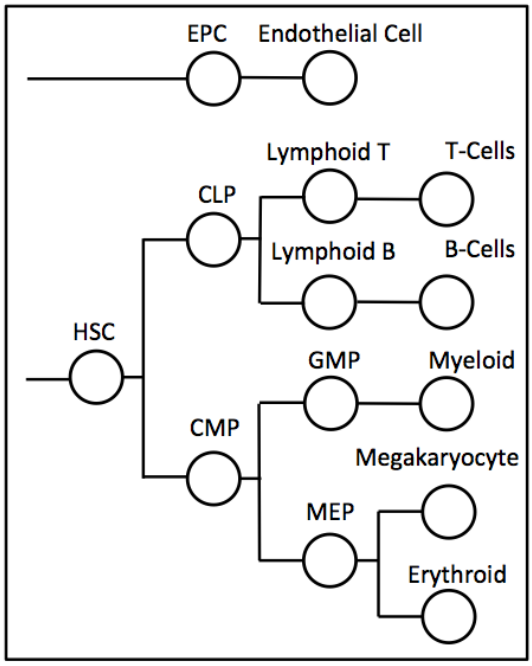
The first aim of this thesis was to establish a method to compare an arbitrarily large number of ChIP-Seq data sets, and to extract subsets of the data that could be analyzed using traditional techniques. The second aim of the thesis was to apply functional analysis techniques to the underlying clusters of peak regions, to elucidate the role of TAL1 in both the hematopoietic and the endothelial lineages.

2.2 Generation of All Blood Cell Types from Endothelial Progenitors and Hematopoietic Stem Cells

Blood and endothelial cells originate from a common precursor (i.e. the hemogenic endothelium) that gives rise to both HSCs and EPCs. There are several levels of cell fate regulation within the endothelial and hematopoietic lineages. The regulation of cell fate ensures that the appropriate proportion of each cell type remains in the body [14].

Figure 1 depicts a common representation of the hematopoietic and endothelial lineages. This tree is not an absolute measure of how similar the cellular conditions are, with respect to TAl1. The following section provides a contextual background on the hematopoietic and endothelial lineages, including the relations between progenitors and differentiated cells and some genes involved in the different members of each lineage.

Figure 1: Visualization of the differentiation pathway in the Hematopoietic and Endothelial lineages. The Endothelial Progenitor Cell (EPC), Common Lymphoblast Progenitor (CLP), Common Myeloid Progenitor (CMP) Megakaryocyte Erythroid Progenitor (MEP).



2.2.1 The Hematopoietic Lineage

The HSC is the source of all cells in the hematopoietic lineage. Red blood cells, white blood cells, and platelets are all final products of HSC differentiation. There are several stages of hematopoietic differentiation, each with different options for further cell fate commitment. Each of these stages serves a purpose in maintaining the population of blood cells within the body [15].

Downstream of the HSC, the hematopoietic lineage is split into two primary branches, the lymphoid and the myeloid. Within each of these branches there are progenitor members, which may be multipotent or bipotent, precursors, which are cell fate committed, and the final differentiated cells [16, 17]. The common lymphoblast progenitor (CLP) is the primary progenitor of the lymphoid branch. Lymphoblasts

have the capability to differentiate into immune response cells such as T and B Cells [18]. The myeloid portion of the lineage has several levels of progenitor control and gives rise to the other members of the blood, which include megakaryocytes and erythroid cells [19].

2.2.2 The Endothelial Lineage

Whereas the hematopoietic lineage gives rise to a variety of cells with a large range of functionality, the endothelial lineage principally provides cells for the development and repair of functional vasculature [20]. The differentiated endothelial cells form the inner most layer of the vasculature and facilitate nutrient transfer through endocytosis. The precursors to the fully differentiated endothelial cells have the capability of migrating from their source of origin to locations where new vasculature is forming, or where there has been damage done to existing blood vessels [21].

Endothelial Colony Forming Cells (ECFCs) are derived in cell culture from endothelial progenitors present in cord blood [22]. ECFCs have the capacity to migrate and differentiate to form capillary-like networks of endothelial cells (or tubes) in vitro upon plating on Matrigel. It has been shown in vivo using mice models that ECFCs have the capability to move to locations where blood vessels have been damaged and initiate the formation of new vasculature [22].

Members of the endothelial and hematopoietic lineages are regulated by a similar subset of genes including CD34, c-kit, GATA2 LMO2 RUNX1 and TAL1 [23, 24]. For both of these lineages, these shared genes drive cell fate commitment. Small changes in the expression of these genes can alter the functionality of the cell. The interaction of genes and the effect of the changes on one another are referred to as gene regulatory networks.

2.3 Transcription Factors Control Cell Fate Decisions During Hematopoiesis

Eukaryotes employ several regulatory mechanisms that allow for fine control of gene expression and protein formation. The rate of functional protein formation can be controlled at any point from the DNA to post translational modifications of proteins. Transcription controls regulate the probability of a transcription initiation factor, such as RNA polymerase II, being able to access the transcription start site(TSS) of the gene of interest [25, 26, 27]. Some factors that influence these probabilities include the physical accessibility of the DNA through the positioning of nucleosomes, and attraction or repulsion factors, such as the presence of complementary proteins (either co-factors or antagonistically binding proteins) [28]. Pretranscriptional control mechanisms, principally TFs, play a fundamental role in cell fate determination of hematopoietic cells [29]. The focus of the following sections is to investigate the functional action of TFs in genetic regulation, specifically how they influence the endothelial and hematopoietic lineages.

2.3.1 Transcription Factors

Genes are influenced by multiple TFs that are both proximal and distal to their TSS. As such, TFs can form complex interconnected regulatory networks that can either reinforce the stability of gene expression levels or drive rapid changes in the gene expression throughout the genome [30].

Transcription factors bind to specific locations throughout the genome and the possibility of them binding those sites is conserved across cell types [6]. Despite the possibility of binding to these sites the probability changes dramatically between cellular conditions. These changes in probability are caused by changes in the accessibility of the DNA, changes in the number of proteins available to bind, and changes in the co-factors that are present, which can interact with the TF.

Several conditions must be in place for a genomic location to become a preferential TF binding location. The binding region must match the DNA binding domain on

the protein (DNA binding motifs) [28]. The DNA must also be accessible in order for a protein to bind. This accessibility implies there must either be an opening between nucleosomes for larger proteins, or thermal fluctuations that allow smaller proteins (pioneering proteins) to attach and initiate changes in accessibility [31]. Finally the protein must be conditioned to attach. The conditioning may include homodimerization, heterodimerization with a partner or an alteration of the protein through a process such as phosphorylation or acetylation [32, 33].

2.3.2 Regulation of Transcription factors

The regulation of transcription factors is critical for maintaining steady gene expression [30]. There are several methods through which the rate of transcription factor binding can be regulated across the genome. Including the rate of the protein formation, the permeability of the protein to the nuclear membrane and interactions with other proteins.

Transcription factors, as with any other proteins, are generated through the process of transcription and translation. As such TFs can be controlled by the same mechanism as any other gene. TFs can regulate one another and may form feedback networks with their own genes [34].

As with other proteins, TFs are generated by the ribosomes outside of the nucleus. In order to interact with the chromatin, TFs must be within the nucleus. Not all TFs are inherently permeable to the nuclear membrane. Some TFs, such as nuclear receptors, require interaction with a ligand in order to become permeable to the nuclear membrane [35]. Through this action extra-cellular signals can indirectly influence the rate of TF binding.

Transcription factors do not interact with chromatin in isolation. There are other proteins present that can either share the direct DNA binding domain as the TF or interact directly with the TF [36]. As such many of these proteins can compete for binding locations or alter the binding potential of other TFs. Alternatively some proteins must interact in order to bind to DNA [32]. TAL1 is an example of one such protein. It must dimerize with an E-protein in order to directly bind to DNA.

2.3.3 Transcription Factors are critical for regulating the Hematopoietic and Endothelial lineages

Cell fate determination, within the hematopoietic lineage, is driven by several genes including RUNX1, GATA1, GATA2, and TAL1 [37]. These genes transcribe TFs that act as critical controls of differentiation and cell fate determination.

The combinations of these genes, and others that transcribe critical TFs, can be activated and repressed during differentiation. For instance, GATA2 is expressed in HSC but is turned off during erythropoiesis. GATA1 behaves in reverse. During erythropoiesis it becomes dominant, however in HSCs it is not highly expressed [38]. RUNX1 is another critical TF. It interacts with TAL1 and is critical for the formation of healthy HSCs. RUNX1 is also expressed in healthy T-Cells in the absence of TAL1 [39, 15].

As mentioned above, TFs do not function in isolation. The interaction between TFs can lead to the formation of complexes. Based on the members present, each complex may have a different role and be able to bind to a different region of the genome. The transition, from the expression of one complex member to another, can alter the binding preference of the whole complex, changing the genes that it regulates. In the hematopoietic lineage, TAL1 is a critical regulator and interacts with proteins including, LMO1, GATA1, GATA2, RUNX1 and E2A [5, 40].

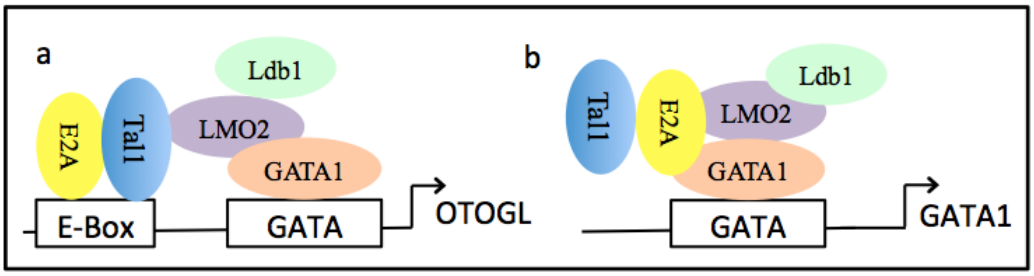
2.4 The TF TAL1 has Multiple Functions in the Different Endothelial and Hematopoietic Cell Environments

2.4.1 TAL1

The relation between TAL1 and genetic regulation is complex. TAL1 has a dynamic capability to influence regulation through different mechanisms in many conditions. TAL1 possesses a basic helix loop helix (bHLH) DNA binding domain, however in order to bind to DNA it must first heterodimerize with an E-protein such as HEB

or E2A [32]. Alternatively to binding directly to DNA, TAL1 may be recruited by other transcription factors such as GATA1 or RUNX1 see Figure 2 [13]. Due to the dynamic nature of TAL1-interacting proteins, it can form a wide variety of complexes that let it act both as an activator and as a repressor. While TAL1 primarily binds to enhancer regions it also binds to distal and proximal promoters [39, 15].

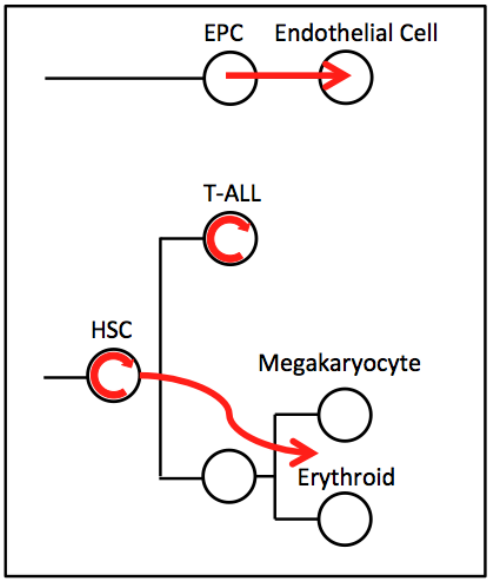
Figure 2: TAL1 can either heterodimerize with an E-Protein and bind directly to the DNA or be recruited by a co-factor



2.4.2 Healthy Hematopoietic Cells

TAL1 is expressed in HSCs and much of the myeloid branch of the hematopoietic lineage [41]. The role of TAL1 in these different cellular conditions varies significantly [10]. These role differences make TAL1 a critical regulator of cell fate determination within the hematopoietic lineage. Figure 3 depicts the subset of the hematopoietic and endothelial lineages that have been investigated in this thesis [5]. In each of the branches, the role of TAL1 changes.

Figure 3: The roles of TAL1 in differentiation of cells within the hematopoietic and endothelial lineages. The directional arrows indicate the direction of cell condition progression as TAL1 is expressed. The internal circular arrow indicates the ability of a cellular condition to maintain its population and proliferate. While TAL1 is critical in the differentiation of the erythroid and megakaryocytes cells, its presence in T-Cells inhibits differentiation and results in increased similarity with HSCs due to the reactivation of repressed genes via TAL1.



TAL1 forms distinct complexes with GATA LMO2 and RUNX1 that are critical for the formation of HSCs, but not for their maintenance [42, 14]. In HSCs, low expression of TAL1 drives quiescence while high expression leads to differentiation to terminal members of the myeloid lineage, principally erythroid cells and megakaryocytes [16]. Based on the pressence of complex members, such as GATA, TAL1 binding patterns shift. This shift drives the reprogaming of gene expression [10].

TAL1 is critical for both erythropoiesis and megakaryopoiesis. Despite the fact that the expression of TAL1 initiates both of these processes, the differentiated products are distinct [10]. These functional differences are associated with variations of the spatial binding preference of TAL1 between the two conditions [9] and possibly

a difference in TAL1 isoform preference [43]. In addition to changes in binding locations, the partners of TAL1 also change during differentiation. For instance, the dominance of binding between TAL1 and Gata2 in HSCs is superseded by binding with Gata1 in erythroid cells [38].

2.4.3 T-Cell Acute Lymphoblastic Leukemia

TAL1 is not highly expressed in primitive lymphoblasts and is not present in any subsequent T or B cells. While in most of the hematopoietic lineage TAL1 drives differentiation, it has been shown that causing aberrant expression of TAL1 within T-Cell lymphoblasts blocks healthy T-Cell differentiation. This block leads to T-Cell acute lymphoblastic leukemia (T-ALL) [44, 45]. The difference in binding preference of TAL1 between erythroid cells and the T-ALL cell line Jurkat has been investigated previously [3]. It was shown that TAL1 displays differential Ebox preferences in these two cellular environments (i.e. CA**GA**TG in erythroid cells and CA**CC**TG in T-ALL). Furthermore, RUNX and ETS binding sites were identified under TAL1 peaks preferentially in a leukemic T-ALL environment [3].

2.4.4 ECFCs

ECFCs share expression with hematopoietic stem cells. Accordingly many TFs that influence HSCs also influence ECFCs, including GATA2, LMO2, RUNX1 and TAL1 [23, 24]. Within ECFCs, the complexes TAL1 forms drive its specific functional capabilities, which include cell migration and differentiation (i.e. capillary-network formation) for the purpose of forming new blood vessels [22].

To better understand the different role(s) of TAL1 between cellular conditions, the binding preference of TAL1 must be determined. Chromatin immunoprecipitation (ChIP) is the principal method for assessing TF binding preference. Performing ChIP using a TAL1 antibody in different cellular conditions will provide the binding information necessary to compare TAL1 binding preference.

2.5 Chromatin Immunoprecipitation

ChIP is a method of isolating the regions of DNA that interact with a protein of interest [46]. ChIP relies on breaking chromatin (with TFs attached) into short regions and isolating those regions based on the protein of interest, by IP. Once isolated, the DNA represents the subset of the genome that interacts with the TF of interest.

ChIP follows a general process that includes cross linking, sonication, immunoprecipitatoin, elution, reverse crosslinking, and purification. In the following few paragraphs I have summarized the process described by Nelson et al [46]. Note that each of the stages can be varied based on the number of cells and the cell types being assessed. Also the sonication stage may be replaced with the addition of MNase, which cleaves the DNA around the nucleosomes.

The DNA and protein are first cross linked. Cross linking firmly attaches the protein to the DNA and can be achieved using formaldehyde. Once the protein and DNA have been cross linked it is sonicated to break the DNA into smaller strands. The size of the strands is dictated by the intensity and period of the sonication. For high throughput sequencing the preferred strand length falls between 200 and 500 bp [7].

Following sonication the chromatin is immunoprecipitated. Magnetic beads that have been linked to antibodies, are used to isolate strands of DNA. The antibody binds to the protein of interest. The strands of DNA that the protein are bound to can then be isolated using magnetic force.

Once the chromatin has been immunoprecipitated the DNA is eluted using a specific elution buffer. NaCl is added to the eluted DNA to reverse the cross linking and RNase A and proteinase k are added to degrade the RNA and protein in the eluted product. The final step is to purify the DNA, this can be accomplished using chloroform.

2.6 ChIP-Seq: A Method to Study TF Binding Genome Wide

2.6.1 Chromatin Immunoprecipitation and High Throughput Sequencing

Once the DNA has been isolated there are several approaches to determine the genomic locations that the protein of interest has bound.

Polymerase Chain Reaction (PCR) is a common approach used to determine if a specific gene is expressed and can be applied to determine if a specific region has a protein bound to it. Primers that bind to specific nucleotide sequences close to the genomic location of interest are added to the extracted DNA and the region following the probes are amplified [47]. The resulting amplified DNA can be run on a gel to determine whether or not the region was present in the initial product, i.e. whether the gene was expressed, or the protein present.

qPCR can be used to investigate the differential binding strength quantitatively between trials. This allows for comparison between proteins in the same cellular condition or between the strength of binding of a single protein between different cellular conditions. qPCR, while fundamental in determining binding in known regions, has limited utility in identifying novel binding sites, and does not reveal binding across the genome.

ChIP high throughput sequencing (ChIP Seq), in contrast to ChIP qPCR, has the ability to determine the TF binding signal across the entire genome. ChIP Seq applies all of the immunoprecipitated DNA, rather than only analyzing predefined regions of interest. These sequenced regions can then be mapped to the genome [7].

ChIP Seq allows for the genome wide analysis of TF binding, without requiring the definition of expected binding sites before hand. This analysis allows for the large scale discovery of preferred binding locations and is especially useful when trying to locate distal enhancer regions. There are several techniques available to estimate the regions of greatest binding likelihood from a single experiment [8]. This allows for the systematic analysis of single ChIP Seq data sets. The static role of the TF in a

single condition may be determined through the analysis of an individual data set, however, in order to observe how the role of the TF changes, multiple ChIP data sets must be compared.

2.7 Current ChIP Seq Comparison Methods

The number of ChIP Seq data sets has increased dramatically over the past few years. The comparison of multiple data sets is key to understanding the relationships between proteins and to elucidate changes in function between cell types. In order to observe changes in binding preference and TF function, there must first be a systematic method of comparing ChIP Seq peak sets. The methodology for the comparison of ChIP Seq data between cellular conditions is non trivial and current approaches face several limitations [6].

2.7.1 Overlap Analysis

Overlap analysis of two ChIP Seq data sets is a direct extension of peak calling. Each data set has enriched protein binding regions identified independently. The ratio of co-localization (overlapping binding locations) can then be used as a distance measure between two or more data sets [48]. Overlap analysis can be used to select unique and shared binding locations and to cluster the data sets.

This approach can be applied to different proteins within the same cellular condition, or to the same proteins shared between different cellular conditions.

2.7.2 Global Analysis

Rather than comparing enriched regions, global analysis compares the full genomic signal. The genome is split into equal sized bins, and the read pile up count is determined for each bin, for each data set being compared. The correlation between the pile up counts is then used as a distance between data sets. Unlike the overlap analysis method, global analysis does not require statistical thresholds. This allows for fair comparison of marginal and moderate peaks when cellular conditions are the

same. When comparing binding between cellular conditions, however, the background signal is not shared and can dominate the correlation between data sets.

2.7.3 Limitations

There are a wide variety of factors that limit the efficacy of true quantitative comparison between ChIP-Seq data sets. Even when comparing data from the same set of experiments, within the same cellular conditions, sequenced on the same platform, the output signal may be composed of different levels of bias and variations in binding signal due to the efficacy of the antibody and the solubility of the chromatin [6]. When comparing the same protein different antibodies can be used with different efficiency. Also the development of more efficient high throughput sequencing methods has lead to a significant difference in read depth between datasets. While a dataset published in 2014 may contain 200 million reads a data set published in 2008 may only have 2 million [49, 50]. When comparing experiments performed by different labs, working with different antibodies, in different cellular conditions, and having the data sequenced over a period of several years on different platforms, the number of possible complications increase dramatically.

These variations contribute to the highly variable levels of both systematic bias and random variation within the genome-wide signal. Due to this bias, when determining enriched locations, the cut offs used must be stringent. Performing a stringent peak detection analysis for an individual data set helps to decrease the impact of background on the results. Coupling the peak detection of treatment and control datasets further mitigates the prevalence of background regions in the final set of enriched regions. Extending this analysis to determining the overlap of peaks between data sets is a common approach.

When biological replicates are compared, peaks found in one condition will not necessarily be found in the other. Only the most enriched peaks pass the detection threshold in both data sets. The marginal peaks in each replicate, despite being statistically enriched in at least one, are much less likely to be determined as overlapping. This double stringency leads to an chronic underestimation of similarity between data

sets [6].

As an alternative to overlap, the global similarity of cellular conditions may be estimated using the height of the genome wide signal. The genome wide signal may either be the read height at each location or the number of reads within a series of subsets of the genome. The correlation of the global signal provides an indicator of cellular condition similarity. The contribution of non specific binding (background) to the signal is the principle limitation of global analysis. When working with data sets from the same cellular condition, the chromatin accessibility, which accounts for a large portion of the variation of background signal magnitude, is conserved between sets. When the cellular condition is different however, the background binding signal is not shared between sets [6] .

2.7.4 Alternative Approaches

My alternative approach for the comparison of ChIP-Seq binding extends methods described by Stark and Hardision [6, 10]. Stark in his 2011 paper proposed an extension to naive global analysis [6]. His alternative was to determine a subset of the genome that is statistically enriched in at least one cellular condition. This region will be referred to as the AFS.

The AFS was defined as the union of peak regions that had been identified independently for both data sets. A fixed peak length was used to avoid basing the analysis towards longer fragments. The union was taken such that it expanded the peak length to contain overlapping peaks (refer to the merge bed utility for details [51]). By including all peaks regions identified the issue of over stringent analysis preventing the comparison of marginal peaks was mitigated. The effect of changing background, which limits global analysis, was also reduced by the inclusion of peak calling. Peak calling can also be used to include the control data sets in the analysis [8]. Since the binding locations are conserved across the genome, more information on marginal binding, which may be sub-threshold in one replicate, will be included in the analysis, reducing the stringency bias.

To identify differentially enriched regions, Stark compared the log2 fold change

of the peak heights between two conditions. The differential peaks were selected by taking those that had the greatest increase for each condition. Since it relied on using the binary fold change between two conditions, Stark's method was limited to binary comparison.

Hardison extended the method to compare six TAL1 mouse data sets. Similar to Stark the peaks were called independently and a unified set was defined, however, rather than using a log2 fold change to separate peaks in a binary fashion he used k-means clustering. K-means clustering was applied to the AFS such that each peak was represented in 6 dimensional space, 6 being the number of data sets being used. 16 subset of peaks were declared to be grouped using this approach. Each subset related to a combination of data sets that it was enriched in. Note that the role of TAL1 in some data sets was to similar to have differential peaks separated using this approach.

In addition to using k-means clustering to group the peak heights, hierarchical clustering was applied to group data sets. The hierarchical clustering separated the stem like hematopoietic cells (megakaryocytes and HPC-7) from the erythroid progenitors and erythroblasts [10].

In Appendix B I demonstrate the reproduction of a subset of this approach on human TAL1 binding in the hematopoietic lineage. While k-means clustering of the read pile up counts was sufficient to separate peaks between a few data sets its efficacy degraded as the number of data sets increased.

2.7.5 Principle Component Analysis

This thesis focuses on generalizing the differential comparison of ChIP Seq signal between an arbitrarily large number of data sets, with and without the presence of controls and replicates. The comparison of two or three data sets can be visualized easily using traditional scatter plots and are amenable to the application of clustering methods like those applied by Hardison. However, as the number of datasets increases it becomes harder for human identification of clusters, and clustering methods face the curse of high dimensionality [52]. PCA can be applied to high dimensional data

to extract a subset of orthogonal vectors which contain the majority of the signal variation

The PCs of the PCA provide two functions. They can be used to facilitate clustering of the data sets, and they also provide weights representing the importance of each peak. These weights can be used to identify which peaks are most representative of either direction along the principle component [53].

If the TF signal variance between conditions is greater than that of the non-specific signal, then biologically relevant clusters will be apparent in a lower dimensionality. This should tell us in which conditions TAL1 binding is shared and in which conditions the binding varies significantly.

PCA facilitates the clustering of the data sets by focusing the variations in the data into fewer dimensions. Each PC was a set of weightings relating each peak to its importance in the observed separation. To compare the distance between data sets in a select set of PCs, the dot product of the select PCs and the normalized UDM was taken. If there are specific dimensions that separate clusters, the PCs are identified which separate the peaks of these conditions. Since each PC is a set of weightings, they can also be used to partition the peaks into sets that contribute most to each condition. In addition to their use in partitioning the data, the PCs also retains its quantitative measure of importance for all peaks in each partition.

2.8 Rationale

2.8.1 Goal

Our goal is to develop a stable method that will scale for comparative analysis of transcription factor binding between multiple cellular conditions. Using this method we will investigate the molecular determinants of TAL1 differential binding, with a particular emphasis in understanding what differentiates TAL1 binding between T-Cell acute Lymphoblastic leukemia and "healthy" environments.

2.8.2 Hypothesis

Our hypothesis is that the differential binding patterns, which are observed between different cellular environments, will yield molecular determinants that will help to elucidate the differential functions of TAL1. Determining these differential functions is key to understanding the role of TAL1 in TAL1-mediated leukemogensisis.

2.8.3 Aims

In order to investigate the validity of our hypothesis we have designed the following specific aims.

Aim1 Establish a method of combining ChIP-Seq data sets and extract biologically relevant genomic regions

The AFS method will be assessed in combination with hierarchical clustering and PCA to cluster cell types and to partition peaks. We wish to determine if TAL1 centric differentiation tree can be generated using only TAL1 binding data. Additionally, we will use each PC as a vector of weights, representing a quantitative measure of importance for each peak in each.

We expect that the biological replicates will group closely as a positive control for data set similarity and that the ECFC that belong to the endothelial lineage will fall outside the other groupings that all belong to the hematopoietic lineage. From previous publications it is clear that the Erythroid and Leukemia cellular conditions should form individual clusters however the position of the HSCs is unknown.

Aim2 Investigate the regions of TAL1 differential biological relevance to identify the molecular properties of importance in TAL1 mediated leukemia

Through the analysis of the AFS we hope to be able to first recreate the relations between T-ALL and Erythroid cell types that have been found previously in the literature. Once we have shown that the AFS method can at the minimum recreate expected results, we aim to identify the molecular determinants that distinguish TAL1 genomic binding in the different cellular environments.

Chapter 3

Methods and Experimental Design

3.1 Overview

In the following chapter I describe my methodology for the comparison of multiple ChIP Seq data sets. I applied this methodology to 22 human TAL1 ChIP Seq data sets. These data sets come from several labs and from cell lines and primary cells that are within either the hematopoietic or the endothelial lineages. In summary, each data set was acquired in short read array (SRA) format from the gene expression omnibus (GEO), and processed separately resulting in raw aligned reads and approximated binding peaks. Once assessed independently, the peaks and raw reads were compared. The comparison of multiple ChIP-Seq data sets was split into two distinct stages, a combination stage, and an extraction stage. For the combination stage the information representing each of the data sets was combined into a single matrix. Once combined underlying subsets were extracted through clustering, and those peaks that contributed most to each cluster were isolated.

The AFS was defined as the combination of the peaks (found independent of one another for each data set) into a single unified set. The formation of the AFS was described by Stark in his 2011 paper [6] and my implementation of it is discussed in Section 3.4. Refer to Section 2.7.4 for details on Stark's implementation.

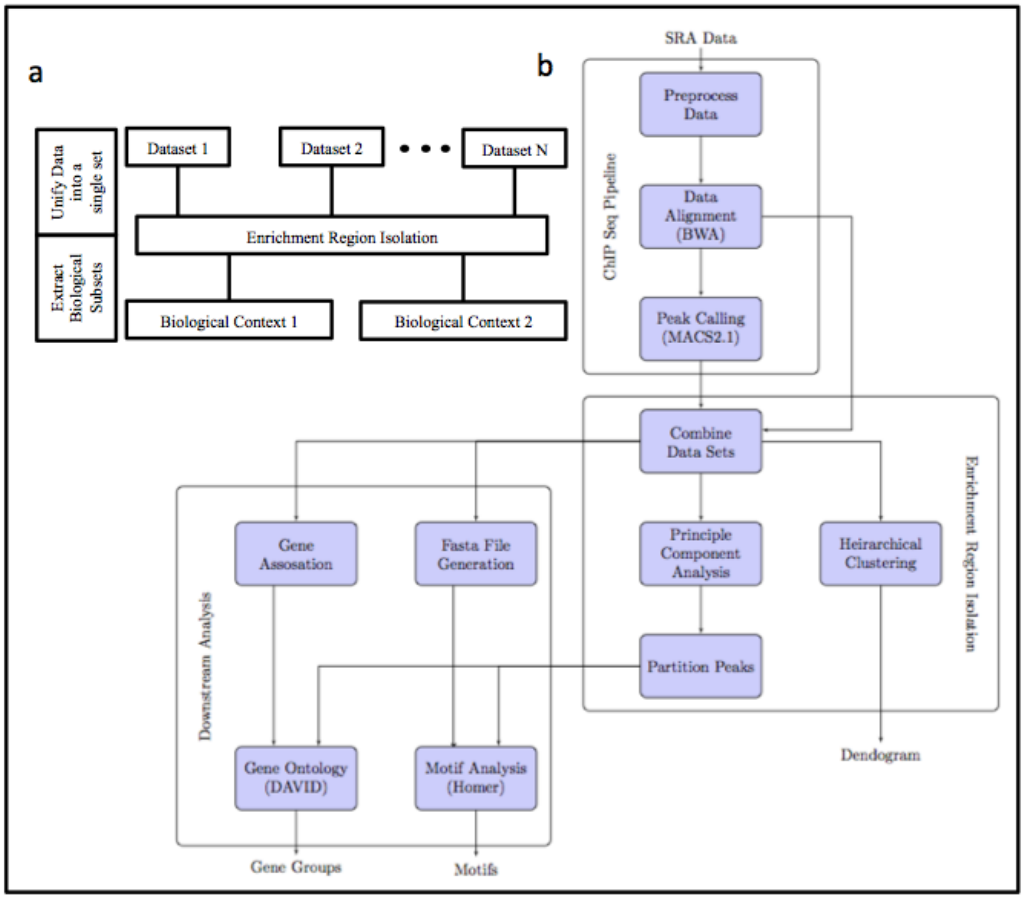
The UDM was a matrix of read count pile ups, one column for each dataset, one row for each peak in the AFS. Each raw read count was normalized by the width

of the peak to provide a read count density. Once the UDM had been generated, hierarchical clustering, k-means clustering, and PCA were applied to the matrix. This was done in order to group the data into clusters based on biological similarity, and to isolate the features that define that similarity. The resulting clusters were assessed for stability through the variation of features such as cut off stringency and background variation, in order to ensure that they were resilient to changes in the preprocessing stage. The application of these machine learning algorithms, coupled with the application of PCA, generalizes Starks analysis beyond his definition of a binary comparison and allows for quantifiable comparison of conditions.

The generation and analysis of the AFS was applied in two experiments. A pilot was done using 4 data sets. The pilot was compared to previously published methods described by Hardison et al. [10] of applying clustering to ChIP-Seq peaks. Hardision's approach was discussed in Section 2.7.4 and further details are provided in Appendix B. The pilot included data sets from the HSC, Erythroid, and T-ALL conditions. Only a portion of the pilot analysis is presented in the body of this thesis. For the complete pilot analysis refer to Appendix B. Once it was shown that applying PCA to the AFS could split these three conditions, the analysis was extended to all 22 data sets.

Once the 22 data sets had been assessed and their peaks isolated, the underlying function of TAL1 in these clusters was investigated. Functional analysis techniques, *de novo* motif search and gene ontology (GO), were applied to the regions that were isolated using the PC vectors of the AFS. The functional analysis was first applied to the well-studied comparison of the Leukemic and Erythroid conditions and subsequently to the novel comparison of the Leukemic and HSC conditions. The overview of the full process can be seen in Figure 4b. Full details on the functional analysis can be found in Section 3.7.

Figure 4: Summary of the process. a.) The abstraction that allows for the combination of multiple data sets relies on two steps. The first is generating a unified data set, against which analysis can be performed. The second is extracting subsets of peaks, which were important in defining clusters that can be used for functional analysis. b.) A flow chart representation of the final process of ChIP-Seq data set comparison.



3.2 Data Sets Analyzed

The cell types of interest included members from both the hematopoietic and endothelial lineages. From the hematopoietic lineage HSCs [50, 54, 55], Erythroid [56, 57, 58], Megakaryocyte [59], and T-ALL cell lines and primary cells [39, 56] were contributed. ECFCs [22] from the endothelial domain were used as a non-hematopoietic base point.

In total 22 TAL1 ChIP-Seq data sets were used in the analysis. From the HSC group there were two CD34 samples provided, as well as one CD133 sample. For the endothelial cells, one ECFC set was used. There were three primary erythroid cells, one fetal and two adult, and two K562 cell line replicates. There was a single megakaryocyte sample.

Additionally there were 5 T-ALL sample types three of which were cell lines and two of which were primary cells. The cell lines included 3 Jurkat data sets, 2 RPMI replicates and 2 CEM replicates. There were samples from two patients included, henceforth called patient 1 and patient 2 (prima1 and prima2). There were two replicate data sets available for prima2 and two labs analyzed prima1 independently. The data sets are summarized in Table 1. More infomation regarding these data sets is available in Appendix A.

Table 1: The summary of data sets analyzed in this thesis. Note that the Prima1(brand) has not, as of yet, been published. The control data set labels included Mock, WCE (Whole Cell Extract) and Input.

Cell Type	No. of Reads	Treatment Code	Control Code	Mock Label
Prima1(brand)	64,400,090	NA	NA	Mock
Prima1	400,006	SRR372690	SRR372682	WCE
Prima1	481,923	SRR372691	SRR372683	WCE
Prima1(rep)	509,859	SRR372693	SRR372695	WCE
Prima2(rep)	446,551	SRR372692	SRR372694	WCE
Jurkat(rep)	5,063,810	SRR443847	SRR443856	Input
Jurkat	5,529,172	SRR443848	SRR443856	Input
Jurkat(brand)	5,831,704	SRR070589	SRR070593	Mock
RPMI	4,842,091	SRR519118	SRR519119	WCE
RPMI(rep)	1,690,681	SRR519120	SRR519119	WCE
CEM	1,183,173	SRR372688	SRR372696	WCE
CEM(rep)	3,181,126	SRR372689	SRR372696	WCE
ECFC	32,407,916	SRR1051799	SRR1051798	Mock
Megakaryocite	10,742,953	SRR070379	SRR070380	Mock
CD133	6,088,943	SRR094806	SRR094805	WCE
CD34	3,791,775	SRR189205	SRR091681	Input
CD34(rep)	17,405,262	SRR1522112	NA	NA
Erythoid	6,479,544	SRR070589	SRR070591	Mock
Erythroid(fetal)	7,630,154	SRR452947	SRR452965	Input
Erythroid(adult)	2,274,458	SRR452948	SRR452966	Input
K562	22,936,696	SRR502380	SRR502109	Input
K562(rep)	15,343,051	SRR502381	SRR502110	Input

3.3 Data Preprocessing

Each of the ChIP-Seq data sets were first processed using a standard ChIP-Seq analysis pipeline. Each data set was assessed independently of the others during this initial stage. The inputs to the analysis were the unprocessed short read archive (SRA) files of TAL1 ChIP-Seq data in the hematopoietic and endothelial lineages, generated from several labs. The outputs of the pipeline were the uniquely mapped non-duplicated reads, mapped using BWA, and the peak locations, determined using MACS.

3.3.1 Acquiring and Formatting Raw Data

Before analysis could take place the ChIP-Seq data had to be acquired and mapped. The unprocessed SRA files were downloaded from the GEO ftp server. The SRA files that were acquired were quality tested using fastqc (v 0.10.0). The SRA files were expanded into fastq files using fastq-dump from the sratoolkit version 2.3.4 [60]. From the fastq format, the data was aligned to hg19 using BWA 0.6.2 [61] on default settings. Once aligned, the reads that were uniquely located on the genome had their duplicates removed. The mapable unique non-duplicated reads were used as the primary data for analysis. The alignment process was performed independently for each of the data sets and their accompanying control set.

3.3.2 Determining Enriched Regions

Peak calling is the identification of regions of the genome found to have statistically enriched read density in the treatment compared with the control. The peaks were called using MACS (version 2.0.1) [8]. MACS is a model based analysis of ChIP-Seq data, which looks to adjust reads found on opposing strands based on a binding model and subsequently identify regions that are significantly enriched, i.e. peaks. After the reads have been adjusted, MACS identifies the regions that are globally enriched by looking at the Poisson expectation of reads within overlapping bins of a uniform size covering the genome. The regions that are globally enriched are further investigated locally by constructing a new dynamic lambda. Lambda represents the

average number of peaks in a given genomic interval, for a Poisson process. If a control is provided the local lambda is built using 1,000bp around the peak summit. If no control is provided the local lambda is generated from the treatment data using 10,000bp around the peak summit[8]. Peak calling was applied to each of the data sets independent of one another. The process of alignment and peak calling for each data set independent of one another was referred to as the independent portion of the analysis.

The output of the independent analysis was modified such that each peak was defined as 701bp (+/- 350) around the maximum read density (summit) of each peak. A single sized peak width was chosen to mitigate the impact of read density signal amplification of outlier peaks that have narrow widths. The width of 701 was chosen between 1.3 and 1.7 times the average peak width of the data sets in order to ensure that the majority of the peaks were contained within the range chosen. Additionally, in order to determine the effects of stringency cut off, a series of p-value cut offs was set from 10^{-5} to 10^{-70} in intervals of $10^{-2.5}$.

3.3.3 Controls Used When Calling Peaks

Three different control definitions were used to determine the peaks. These were use of no control, of a combined control, and of a data set specific control. The no control case used the MACS algorithm without passing it a control data set. The combined control represented a portion of each of the available control data sets. Reads were taken from each control set, such that each set contributed half the number of reads available to be taken from the smallest set.

3.4 Unify the Data Sets

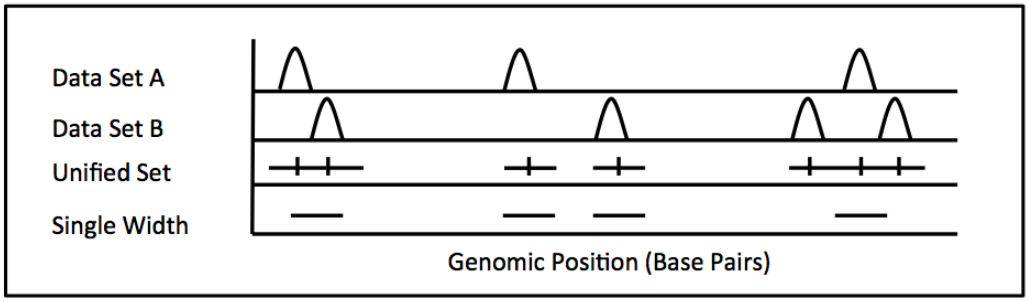
Once the data sets had been preprocessed independently they then needed to be compared. The first stage of the multiple data set comparison was to generate a single data structure representing all data sets. Several methods were investigated, including overlap analysis, and global analysis. Overlap analysis used the number

of peaks that overlapped between two datasets as a measure of condition similarity. Global analysis broke the genome into equal sized bins and used the read pile up in each bin to cluster the data sets. These results of these approaches can be found in Appendix C and D. Going forward the principle method that was used to unify the data sets was the AFS. The following sections outline the methodology of combining data sets first into an AFS, and subsequently into UDM.

3.4.1 Defining an Alternate Feature Space

The AFS represented a subset of the genome that was enriched for TF binding in at least one cellular condition. Since TF binding sites are conserved, a peak location found in one data set may have marginal binding in several other data sets. This marginal binding may not appear in a peak comparison by overlap due to the double statistical stringency required when identifying shared regions.

Figure 5: The unification of several peak regions occurs in two steps. The first step involves overlapping the data. The second step takes the new average of the peak, based on the means used to generate the peak, and limits the distance from the new summit.



The AFS was determined through data set unification. The unification involves two steps, which are outlined in Figure 5. The full analysis, from aligned data to unified data set required four steps, those steps are listed below.

1. **Peak Calling and Peak Summit Selection :**
   Enriched regions were determined independently for each of the cellular data

sets. The peak summits were selected as the identifier as the peak genomic location.

2. **Determine Overlap:**
   The regions of enrichment were overlapped between different cellular conditions. Overlap between two peaks was determined by the number of bp that separate their summits. If two summits were within a set distance of one another they were grouped into one peak. For this analysis the set distance was 350 bp.

3. **Reduce Overlap to New Peaks:**
   The result of the overlap was groupings of peaks, each group represented by a list of summit locations. The average of each grouping was used as the new summit location for each of the combined peaks.

4. **Make Genomic Ranges From Summits:**
   The unified summits were extended upstream and downstream by a predefined distance to form a genomic range. The predefined distance was 350bp. Note that the overlap distance between summits and the peak extension distance need not be the same.

The AFS represents a set of peaks that have come from a variety of data sets and have been modified such that overlapping peaks have been combined. Despite being modified the peaks were associated to the one or more data sets that contributed them. This was accomplished by noting the contribution information for each peak in the AFS. The ability to map the peaks of the AFS to their source is referred to in this thesis as tagging, and was important for determining the efficacy of data set separation later.

3.4.2 Analysis of the Alternate Feature Space

Once the AFS was defined, the pileup counts were determined for each of the data sets. The pileup counts represented the number of unadjusted reads that overlapped each of the peak regions of the AFS. The results are represented as the UDM, which had a width of the number of data sets combined and a length of the number of

peaks within the AFS. The UDM was then normalized using quantile normalization. Quantile normalization normalizes the peaks between conditions based on their rank within each condition. The goal is to make each of the data sets statistically identical. The normalized UDM was analyzed using global analysis techniques, which included determining the correlation between conditions, clustering hierarchically based on those correlations, and using PCA to observe the PCs with the greatest variability. The hierarchical clustering was useful for splitting the data sets into groups. PCA was used to generate a separation vector, through which the importance of a specific region to the definition of a cellular condition was measured.

3.5 Extract Biological Subsets

My method of extracting subsets that make up biological contexts can be summarized as follows. Once the unified matrix of ChIP-Seq binding data had been created, the data was analyzed using clustering methods. The purpose of clustering the peaks and datasets was to determine if there were underlying similarities between data sets. When there are similarities between data sets that share similar cellular conditions, the clusters are referred to as biologically relevant. These clusters can be used to quantify the differences between cellular conditions and ultimately isolate subsets of peaks that have the greatest influence in differentiating the cellular conditions. These peaks could then be assessed for functional purpose. The first step in comparing cellular conditions was to determine the similarity between data sets. The following sections present my method for extraction in greater detail.

3.5.1 Data Set Similarity Using the AFS

My alternative to global and overlap analysis was to assess the pairwise similarity between data sets using the Pearsons correlation of the read density of a subset of a genome, the UDM. By using the correlation of UDM to determine similarity between data sets, the bias from the overlap approach was mitigated, and the proportion of non-specific signal was attenuated.

The correlation on the UDM provides a suitable means to compare two data sets. The direct comparison however cannot be cleanly extended beyond this. In order to see how a large number of data sets relate to one another clustering methods must be applied. The two principle methods that were applied to help determine underlying differences were hierarchical clustering and principle component analysis (PCA).

3.5.2 Hierarchical Clustering

I used Hierarchical clustering to group the 22 data sets into their cellular conditions. Hierarchical clustering groups the data sets based on a defined distance measured between data sets and a linkage used to combine distances. Several distance metrics and linkages were investigated for this analysis. Three linkages were assessed, single, complete, and average. Single linkage compares the shortest distance between two nodes, complete linkage compares the greatest distance between two nodes, and average linkage compares the average distance between two nodes. The distance metrics investigated included dissimilarity, euclidean, manhattan and a pnorm distance where p was 5.The dissimilarity of the conditions is represents $1 - correlation$.

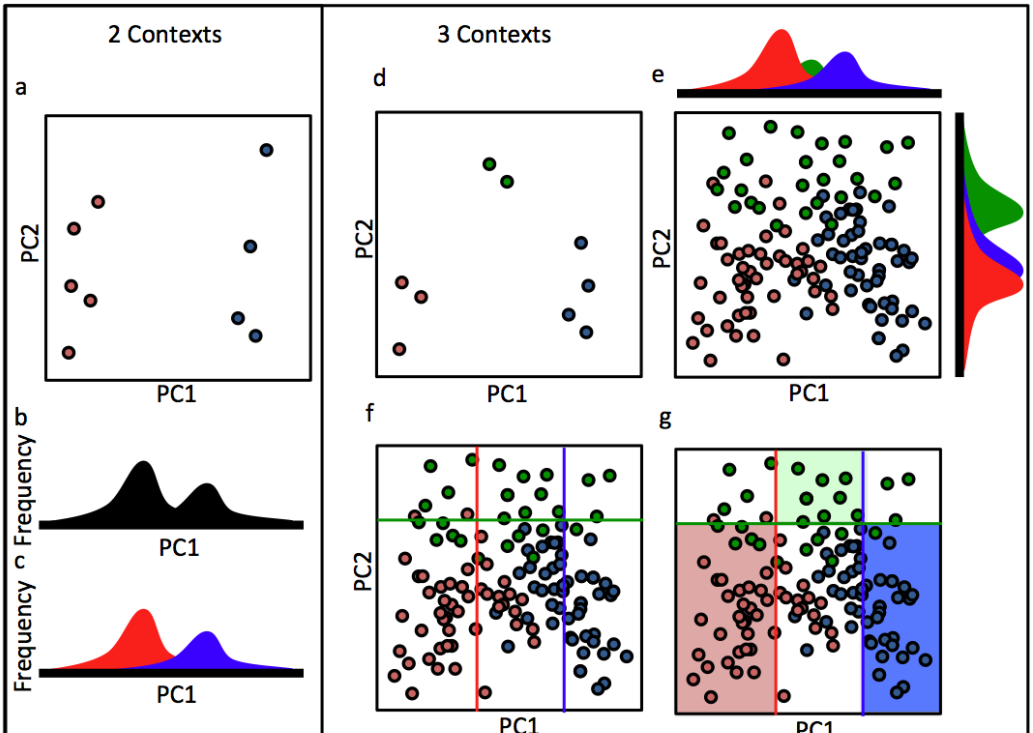
The hierarchical clustering of the 22 datasets was represented as a dendrogram. Biological replicates were expected to have the highest correlation and the cellular conditions of the most disparate cell types should have the lowest correlation (ECFCs should form an outlier in the data). Hierarchical clustering was useful for grouping data sets and k-means clustering could have been used to group peaks, however neither method provided an interval value for each peak, dictating its importance to the definition of each cell type (i.e. a quantifiable value relating the importance of a peak to a specific cluster). If a dimension could be found that separated the peaks into sets that are shown to have biological relevance, than the ability to group data sets and peaks would be combined. This combination would provide a quantifiable measure of how important each peak was to each condition.

3.5.3 Isolation of Subsets with Biologically Similar Members

Q-Mode PCA was applied to the quantile normalized UDM. The UDM matrix had dimensions of nxp, where n represented the number of data sets, and p represented the number of peaks in the AFS. PCA was applied to the transposed UDM matrix, which resulted in one PC vector of length p for each data set being compared i.e. the result of the PCA was a nxp matrix.

For the purpose of this report, the PC Vectors of the PCA were chosen as the method of isolating subsets of biological relevance. The PC Vectors were selected based on the data sets they separated and the proportion of total signal variability they contained, i.e. the proportion of total signal variance present in each PC. The data sets that formed the clusters determined the biological relevance of that cluster. For instance, a cluster that contained only T-ALL data sets was presumed to contain peaks important in the T-ALL condition.

Figure 6: An example application of using principle component analysis to select regions based on their significance to biological definition. The figure demonstrates the application to two separate experiments, one that has two cellular conditions and one that has three. a.) The weighted position of 9 data sets from two contexts along the first and second principle component. The results of the PCA clearly separate the type a data sets (red) from the type b data sets (blue). It indicates that the context is dependent on the first principle component and independent of the second. b.) The projection of peak weights onto the first principle component. c.) The projection of the weights and the color their based on the data set they came from, either red or blue. This example shows that the majority of the regions that contribute to the extrema of the first axis come from the respective cell types. Differential peaks will fall at their respective extremes, i.e. red more negative, blue more positive. d.) The weighted locations of 8 data sets from three cellular conditions or contexts. The weighted positions are defined by the dot product of the normalized UDM and the principle component weights along the first and second principle axis. e.) The PC Vectors for the first and second principle component. The dot product of these vectors with the normalized UDM yields the results seen in subplot d. Each point represents the weight of a single peak in each dimension. Note that it is the projection of these points along the first PC that forms the peak distributions like those seen in subplot b and c. f.) The thresholds used to separate the differential peaks. g.) The resulting groups of peaks.



Comparing the weighted location of the data sets was useful for determining if the data set locations were dependent on specific PCs. This process is shown in Figure 6a-c. From a we can see that the separation of the two cellular conditions (blue and red) is dependent on PC1 but not PC2. Thus PC1 can be used to separate the blue data sets from the red.

Once the PCs were identified that separated data set of interest the peaks which had the greatest influence on the weighted locations were identified. Figure 6e-g represents the separation of three data sets. The proportion of peaks from each data set along each PC was visualized by ploting the peaks based on their PC Vectors. The peaks were selected by taking those peaks that fell a set multiple of the standard deviation from the mean.

The full method can be described by the Equation 1. $\mathbf{M}$ represents the quantile normalized and column mean adjusted UDM. It has n columns representing the number of data sets, and p rows representing the number of peaks in the AFS. Equation 1b represents the singular value decomposition (SVD) of $\mathbf{M}$ into its left and right components, $\mathbf{U}$ and $\mathbf{V}$ respectively. Since we are doing Q-Mode PCA, i.e. applying PCA to reduce the complexity in the samples rather than the features (or rows rather than columns), I took the left factor ($\mathbf{U}$) as the set of PC Vectors. The first PC Vector is represented by the first column of $\mathbf{U}$. If data set weighted position is dependent on $\mathbf{U}_1$ then those peaks with a weight greater and less than N times the standard deviation of $\mathbf{U}_1$ σ are isolated as the peaks with the greatest influence on the separation of data sets.

$$\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} & \dots & m_{1n} \\ m_{21} & m_{22} & \dots & m_{2n} \\ \dots & \dots & & \dots \\ m_{p1} & m_{p2} & \dots & m_{pn} \end{pmatrix} \tag{1a}$$

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^t \tag{1b}$$

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & ... & u_{1r} \\ u_{21} & u_{22} & ... & u_{2r} \\ ... & ... & & ... \\ u_{p1} & u_{p2} & ... & u_{pr} \end{pmatrix} \tag{1c}$$

$$\mathbf{PCV} = \begin{pmatrix} u_{11} & u_{21} & ... & u_{p1} \end{pmatrix} \tag{1d}$$

$$\mathbf{Peaks}_{TypeA} \in \mathbf{PCV}_i \quad \forall\, i\ s.t.\ \mathbf{PCV}_i \geq \overline{\mathbf{PCV}} + \sigma \cdot N \tag{1e}$$

$$\mathbf{Peaks}_{TypeB} \in \mathbf{PCV}_i \quad \forall\, i\ s.t.\ \mathbf{PCV}_i \leq \overline{\mathbf{PCV}} - \sigma \cdot N \tag{1f}$$

As an example, if the first hypothetical PC clearly separates cell type A from cell type B, with the remaining cell types floating around the center, then each value in the PC represents a weight separating A from B. The peaks that have the greatest magnitude in each direction have the most influence in defining the separation. Thus, isolating the members at the vector extrema provides cell condition defining subsets of the genome.

3.6 Determining the Stability of the Comparison

With the analysis defined, the focus shifted to determining the stability of the comparison (i.e. how much do non biological factors, such as noise and data set size, impact the clustering). There were several technical factors that could alter the consistency of the analysis. These factors were assessed to ensure that the clusters formed did not change dramatically with small technical alterations. The technical factors that were investigated included sensitivity to cut off stringency, contribution of background, and the possible dominance of large data sets.

3.6.1 Sensitivity to Cut Off Stringency

The first stability concern that was investigated was the issue of sensitivity to cut off stringency. The genomic regions that made up the AFS were identified as those found to be statistically enriched. Ensuring that the regions were enriched in at least one dataset helped to limit the contribution of non-specific or transient binding signal. The stringency of the statistical threshold was varied using the p-value.

In order to determine the impact of statistical cut off on the analysis the p-value was shifted and the impact on the inputs to the clustering were observed. The influence of cut off stringency was measured for both the overlap and correlation of the data sets. It was accomplished by comparing the correlation and overlap of similar and different data sets as the statistical stringency was increased from 10^{-5} to 10^{-70} by steps of 2.5. The sensitivity to cut off is a measure of how much variation occurs as the stringency changes.

3.6.2 Contribution of Background

As with the stringency cut off, background signal also impacts the comparison of multiple data sets. The background signal of TF data is mainly composed of transient and non-specific binding, and is related to how accessible the DNA is. Whereas naive global analysis would take into account large regions of the genome where the TF was unlikely to be bound, the AFS reduced this region to one where the TF was most likely be bound. Despite its reduction, the background signal will still be present in the analysis. We want to be sure that the clusters that are found on the vectors generated by the PCA are not simply products of the background signal.

To do this we kept the AFS constant and swapped the treatment raw read sets with those of their control. The analysis was repeated and the PCs were investigated to see if the clusters remained. It was expected that replicates would still appear close due to the similarities in their chromatin structure, but the separations between cell types would not be as clear.

3.6.3 Dominance of Large Data Sets

Even if statistical cut off and background signal have no influence on the clusters found, the groupings could still be dominated based on the size of the data set. Each data set varies in size in two ways. First, the number of peaks it provides to the AFS, second, the number of raw reads available to be counted under those peaks. Due to the large variance in the sizes of the data sets being analyzed, we were concerned that the biological variance may be dwarfed relatively.

To determine the effect of data set size on the clustering, the PCs were assessed based on both their correlation with measures of data set size, and the impact of removing members from their formation. The numbers of peaks and raw reads were determined and used as measures of data set size. Additionally for each cluster the largest member was removed and the analysis was repeated. This was done to ensure that the largest members were not explicitly required to form that cluster.

3.7 Functional analysis

Once the 22 data sets had been assessed and their peaks isolated, the underlying function of TAL1 in these clusters was investigated. In overview, functional analysis techniques such as Gene Ontology (GO) and de novo motif search provided insight into the relation between TAL1 binding locations and its functional role in the hematopoietic and endothelial lineages. Functional analysis techniques were applied to the regions that were determined to have the greatest genomic influence. The following sections dictate how the data sets were separated and how Motif analysis and GO were applied to the partitions defined using the PCA of the UDM.

3.7.1 Cellular Context Separation

The final parameters I used for the separation of peaks were a MACS2 score of 5 when calling peaks, a MACS2 score of 20 ($pvalue = 10^{-20}$) when selecting peaks for the AFS, and a two standard deviation cut off when selecting important peaks using the PCs. The two standard deviation cut off was applied along each PC that separated

the context. Additionally, the peaks selected for each context were required to be unique to their specific context, i.e. Erythroid peaks could not also be defined as Luekemic. Refer to Figure 6d-g.

There were two backgrounds that were used when comparing the Ebox preference and preforming the motif analysis. The two backgrounds were all peaks included in the AFS (referred to as ALL) and all peaks that fell within 0.25 standard deviations of the mean for all PCs used to separate peaks(referred to as NONE).

3.7.2 Motif Analysis

The primary functional analysis, applied to the TAL1 binding locations, was motif analysis. DNA binding motifs represent a sequence of nucleic acids, on which the TF has a strong potential to bind. For analysis, DNA motifs representing the four nucleic acids Cytosine, Guanine, Adenine and Thymine were declaimed as C, G, A and T respectively. Additionally IUPAC code names were used to represent each of the possible nucleic acid combinations. These names can be found in the functional appendix.

Ebox Frequency Analysis

The frequency of Ebox (CANNTG) occurrence was compared between the defined conditions. The 10 Ebox combinations in addition to the NN Ebox were assessed for all contexts separated using my PCA method. Only the Eboxes that fell closest to the summit of each peak were used in the analysis. This limited the number of possible Eboxes per read to 1. The frequency was represented as the ratio of Eboxes to the total number of peaks in each condition. In addition to the frequency of Ebox occurrence, the proportion of peaks that contained no Eboxes was also determined.

De novo DNA Motif search

DNA motif de novo techniques determine statistically enhanced motifs in the genomic subset of interest compared to a control. The genomic subsets of interest were defined using the PCs of the AFS. Two controls were used, ALL (representing all peaks in the

AFS) and none, which represents those peaks that fall within 0.25 standard deviations of the separating PCs.

De novo DNA Motif search was performed using the HOMER alignment tool (version 4.7) [62]. Motifs lengths from 6-9 were investigated.

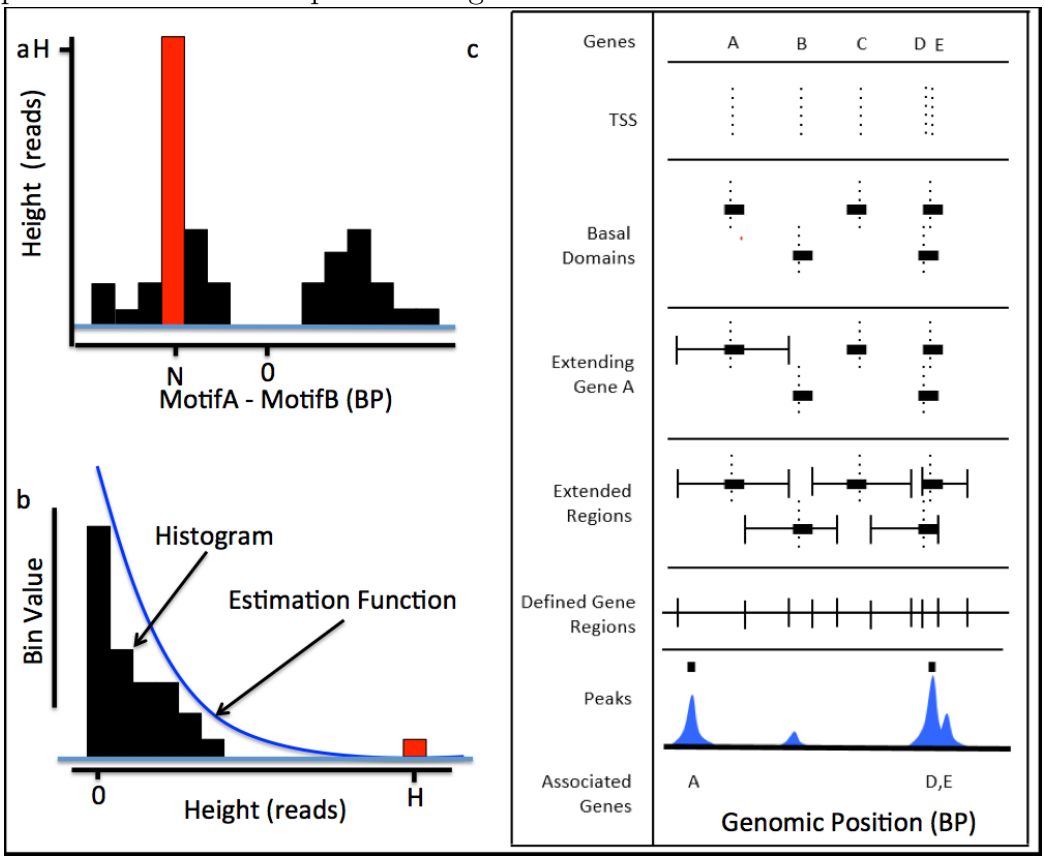
Annotation of Motifs

Once the dominant nucleotide sequences had been found, the next step was to determine if they were similar to previously found motifs. There are several databases containing known motif preferences for different TFs [63]. STAMP was used to determine if the motifs found using DENovo were similar to those found in the Jaspar and Transfac databases. STAMP is a motif annotation web application, the settings used for annotation were the correlation coefficient as the comparison metric and the ungapped Smith Waterman alignment method [64].

Preferred Distances

Once the most likely motifs were found and annotated, their relation to one another was then assessed. If proteins bind to each other and to the DNA, the geometric relation between the binding locations on the genome is limited by the organization of the chromatin and the characteristics of the proteins involved. In one dimension this limitation can be observed as a preferred distance between the two motifs. This effect can be observed by determining the distance between two motifs in a relevant region of the genome and observing if at a specific distance the motif distance distribution is enriched. Additionally by changing the region of interest the significance of these relations may change. The method of isolating preferred distances that were enriched can be seen in figure 7

Figure 7: Functional analysis. The frequency that two motifs are a specific distance from one another under shared peaks is shown in subplot a. The red region indicates a distance which is statistically greater than the rest of the signal, having a height of H. The significance estimation can be seen in subplot b, where an estimation function, based on the entire set of differences, is used to predict how significant the outlier H is. N is the preferred distance between motifA and motifB c.) McLean et al's 2010 GREAT algorithm. The algorithm associates genes with ChIP-Seq peaks in two stages. The first stage is to define a basal domain around each of the gene TSS. Once the basal domain has been determined the regions are extended up to a point where they either intersect another basal region or reach a predefined maximum range. Once the extended defined regions are found the locations are crossed with the summits of the peaks to associate the peaks with genes.



To determine if there were any underlying motifs that had differential preferred

distances between cellular conditions, all motifs discovered in the DENovo and all the motifs found in the Jaspar database were compared. The importance of these preferred distances was compared between conditions to determine if there was a dependence on cell type.

3.7.3 Gene Ontology

The role of the motif analysis was to estimate which proteins interact to when bound to the DNA. This gave no insights into changes in gene regulation between conditions. Gene Ontology (GO) maps a set of genes to categories that have known roles. These categories are then given a statistical weight based on how many genes are included in the input set and how many genes are part of the category. When coupled with gene association, GO can be used to estimate the differential functional roles of TFs between conditions.

Gene Association

The data regions being analyzed are not genes, as such the location of these regions must be used to approximate which genes are being influenced. The Stanford GREAT gene analysis method was used to generate a list of genes from the relevant peaks [65]. Due to interaction limitations with the GREAT API, the algorithm for association was repackaged in R. The method can be visualized in Figure 7 c.

The association consists of two steps. The first is to break the genome down into a series of mutually exclusive regions relating to any number of genes. Once these regions are determined, the list of peaks is crossed with the regions to generate a list of associated genes. The distance of association for each gene by default in the GREAT analysis is the minimum of either 1GBp or the maximum extension up to a basal region of another gene.

DAVID Functional Analysis

Once the peak regions of interest have been associated with genes the gene list can then be analyzed. For the GO analysis DAVID was used. DAVID provides the

relevant GO terms in addition to possible disease relations and pathways.

RNA Seq Experiment

RNA Seq data was analyzed for the knock down (KD) of TAL1 in the Jurkat cell line (Leukemia). The reads were aligned using Top Hat and DESeq was used to determine the differential expression between the KD and wild type (WT) [66, 67]. The log2 fold change (LFC) in expression was used to compare the genes between conditions. Those genes that had a LFC greater than 1 were presumed to have different expression between conditions. An increase in fold change (KD relative to WT) indicated down regulation of the gene and a decrease an up regulation.

I first investigated the regions that changed between WT and TAL1 KD, regardless of whether they were either up or down regulated. I compared these regions for the genes associated with peaks that were in the Leukemic context. The Leukemic peaks were defined as those that were less than a SD from the mean of the principle component that isolated the leukemic context. The genes were associated using the GREAT algorithm, using their default parameters. Subsequently, the effect of the presence of an Ebox under each peak was assessed. The fold change between those peaks that had an Ebox and those that did not were compared using Welchs T-Test. Finally, since the TAL1 complex can act as a repressor and as an activator, the impact of the important Ebox variants were compared to genes that were strictly up regulated and down regulated individually. The gene fold changes were compared with the fold change of all up regulated and all down regulated genes respectively.

Chapter 4

Evaluation of Comparison Framework

4.1 Overview

In this section I present my results for the comparison of ChIP-Seq data between disparate data sets. I present my results in four stages: the comparison of correlation and overlap (Section 4.3), the definition of parameters used to cluster the data sets hierarchically (Section 4.4), the separation of contexts using PCA (Section 4.5), and the stability of the PCA to technical alterations in the analysis (Section 4.6). For the analysis I defined two distinct AFSs and UDMs, a pilot scale, and the full scale. The pilot scale contained four data sets. I used the pilot subset to investigate overlap bias, hierarchical clustering parameters, and PCA on a limited scale. The limited scope was useful for narrowing the parameters that should be used in hierarchical clustering and for showing the efficacy of PCA in separating peaks based on their biological context. The full AFS, containing all 22 TAL1 ChIP-Seq data sets, was used to finalize the parameters in hierarchical clustering and to find principle components that separated biological contexts. The principle components of the quantile normalized UDM were used to partition peaks for the downstream functional analysis.

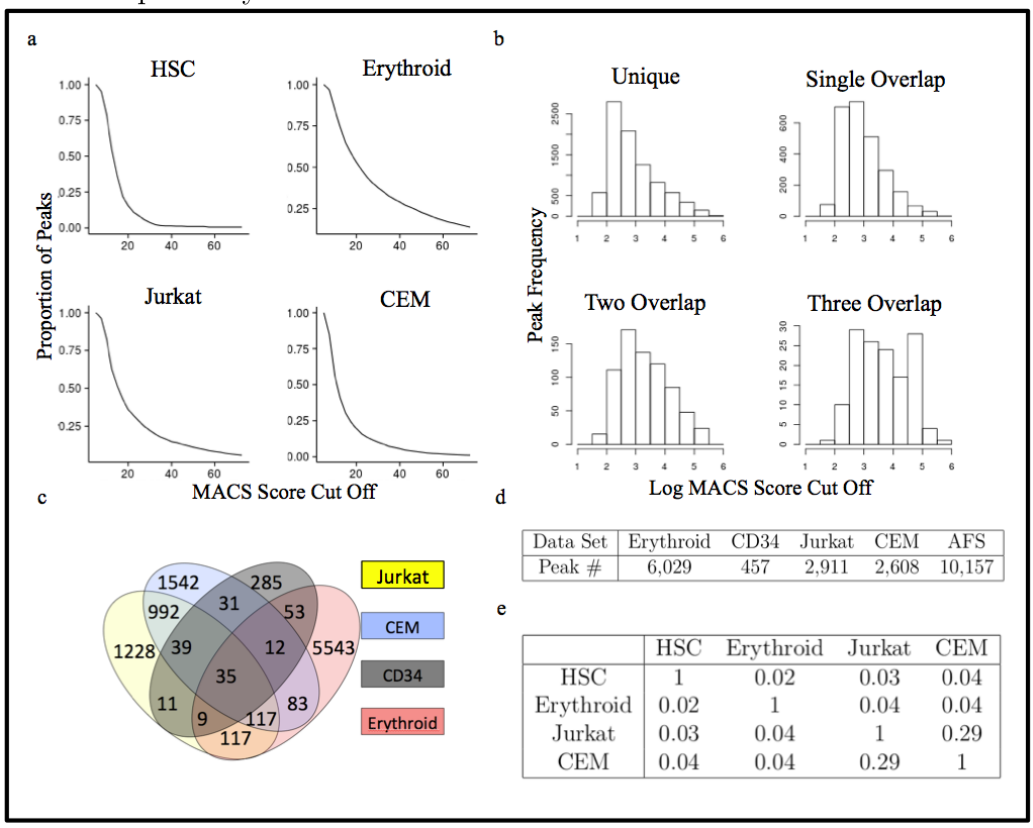
4.2 The Data Sets

4.2.1 Pilot Scale Peak Calling

For the pilot scale AFS, I compared four data sets: two leukemic, Jurkat and CCRF-CEM/C1 (CEM), an Erythroid and a CD34. These data sets had a 5 fold range in the number of mapped unique non-duplicated reads, and had been sequenced between 2010 and 2014 by three different labs. Additionally, I used two K562 (an erythroid cell line) data sets in the analysis of the hierarchical clustering parameters. The aims of the pilot scale analysis were to compare and contrast the use of overlap and correlation as a measure of data set similarity, to help define the parameters to be used in the hierarchical clustering, and to investigate the efficacy of using the PCA to partition data sets.

To identify peaks in the four data sets, I ran MACS2.1 with a p-value. cut off of 10^{-5} (the default defined by MACS) and a combined control The outputs generated from calling peaks, using the standard p-value cut off provided by MACS (10^{-5}), can be seen in Figure 8 d. The most peaks were seen in the erythroid condition and the least were seen in the HSC 8d. Referring to Table 1 in Appendix A, the erythroid and HSC data sets have comparable number of peaks. The difference in TAL binding my be due to the relatively open chromatin in stem cells allowing binding to many positions that may be sub threshold [68]. For the default p-value using the combined controls there were 10,157 peaks in the pilot AFS.

Figure 8: The number of peaks called and overlap between the four pilot data sets and the impact of increasing the statistical stringency. a.) Proportion of peaks as a function of MACS score cut off represented as 1 - the empirical CDF. Note the non-increasing nature between the number of peaks identified and the MACS score, where MACS score is $-log10(pvalue)$. b.) The Histogram of log MACS scores. The Histogram indicates that the degree of overlap increase is accompanied by increasing MACS scores, i.e. MACS score is dependent on the number of data sets being overlapped. c.) is a Venn diagram depicting the number of peaks that fall into each overlap bin. Subsections d and e represent the number of peaks called and the overlap proportion respectively



Data Set	Erythroid	CD34	Jurkat	CEM	AFS
Peak #	6,029	457	2,911	2,608	10,157

	HSC	Erythroid	Jurkat	CEM
HSC	1	0.02	0.03	0.04
Erythroid	0.02	1	0.04	0.04
Jurkat	0.03	0.04	1	0.29
CEM	0.04	0.04	0.29	1

I also filtered the peaks from the initial peak calling results, defining increasing stringency thresholds ranging from a p-value of 10^{-5} to 10^{-70}, at intervals of $10^{-2.5}$. As the p-value cut off increased, the numbers of peaks found in each of the data sets

were seen to be non-increasing. The result of increasing the p-value on the number of peaks called can be seen in Figure 8a. Just as with the individual data sets, the number of peaks in the AFS decreased as a function of increasing p-value.

4.2.2 Complete Set of Peaks Called

The full analysis included 22 TAL1 ChIP-Seq data sets representing 12 unique cellular conditions and five contexts of interest. The contexts were T-ALL, Erythroid, HSC, ECFC, and Megakaryocyte. Information regarding the peaks that I used to generate the full AFS can be found in Appendix A. As with the peaks used in the pilot AFS, the peaks for the full AFS were identified using MACS 2.1. The final AFS, created using the standard cut off, contained 130,305 peaks. The numbers of peaks in each data set varied by more than two orders of magnitude. The smallest data set (CD134) contained only 109 peaks and the largest data set (ECFC) had 68,254.

4.3 Comparing Overlap and Correlation Between Data Sets

Before I applied hierarchical clustering to either the pilot or full scale UDM, I compared correlation and overlap in respect to their utility in measuring distance between data sets. In this section, I show how, using the pilot scale subset, I confirmed that there was statistical bias present in complete overlap based on the number of overlaps. Additionally using the full scale analysis, I show that when looking at the effects of changing p-value cut off, correlation is more resilient than overlap when measuring distance between conditions.

4.3.1 Pilot Scale Assessment of Correlation and Overlap

The complete overlaps between the four conditions are presented in the Venn diagram seen in Figure 8c. The bin pertaining to the four-way overlap had 9 times fewer peaks than the bin that was unique to the smallest cellular condition. Since three of the

four conditions come from different contexts, the four way overlap was not anticipated to be large. However, I hypothesized that this category would be artificially shrunk, since with each increasing overlap the peaks would have to pass an additional test, biasing the peaks to fall in bins with fewer overlaps. Thus the innermost category (the overlap of all data sets) would have to pass 4 tests while the peaks that were found to be unique to a data set had to pass only one.

As I anticipated, there was an increase in average p-value as the overlap level increased. The histograms of peak p-values for the four levels of overlap can be seen in Figure 8b. The geometric mean p-value for unique regions was $10^{-27.5}$. The mean for all regions with at least one overlap was $10^{-30.7}$, for two overlaps it was $10^{-42.8}$ and for three overlaps it was more than double that of the unique regions at $10^{-56.9}$. In order to avoid this bias I investigated the use of correlation as a measure of distance between cellular conditions, and limited overlap to the pairwise comparison of data sets. Since pairwise overlap had the same degree of overlap for all conditions being compared the relative distance would not be impacted by the degree of overlap.

The correlation between the 4 conditions can be seen in Figure 9a. I determined this correlation using the default MACS p-value (10^{-5}). Figure 9 also depicts the effect of changing p-values on overlap and correlation. As was expected, CEM and Jurkat had the highest correlation while the erythroid data set was seen to correlate marginally with both. CEM and Jurkat are both cellular conditions from the T-ALL context while the Erythroid data set is from another, self-titled, context.

In order to determine which measure was more resilient to statistical cut off when identifying peaks, I observed how changing the stringency of the peak detection would effect both the correlation and the overlap between data sets. As the p-value increased both the overlap and the correlation between the two disparate data sets (Jurkat and Erythroid) decreased. The correlation of the two T-ALL datasets increased with increasing stringency. As such, if dissimilarity (1 - correlation) was used as a distance measure, as the p-value increased the distance between the similar data sets would decrease and the distance between the different contexts would increase. Unlike the correlation, the overlap between the two T-ALL data sets decreased with p-value. The decrease of the similar data set overlap was roughly the same ratio as the decrease

in the overlap of the disparate data sets. If the ratio of decrease in overlap is shared between all data sets than it may be possible that the relative overlap between data sets is independent of the cut off used. The analysis of this subset, however, is not sufficient to strongly demonstrate this point. From this scale of comparison it was not clear whether binary overlap or correlation should be used as distance metric for relative comparisons, although the evidence favored the use of correlation.

Figure 9: Comparing the correlation and overlap of similar and disparate data sets as a function of p-value cut off. Jurkat and CEM are both T-ALL cell lines. Subsection and b are respectively tables indicating the correlation between each of the data sets for an AFS defined using a MACS score cut off of 5 and 70. c.) The relation between overlap and correlation depending on MACS cut

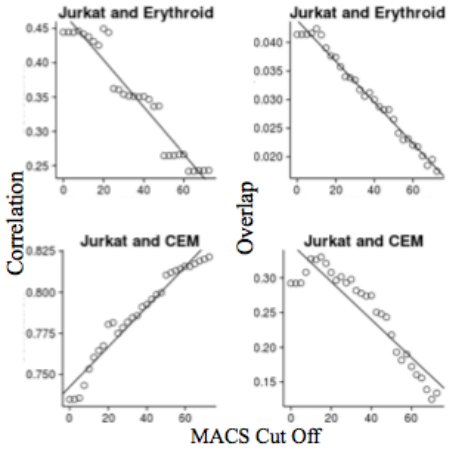
a

p-value : 10^{-5}	HSC	Erythroid	Jurkat	CEM
HSC	1	0.39	0.59	0.52
Erythroid	0.39	1	0.44	0.37
Jurkat	0.59	0.44	1	0.73
CEM	0.52	0.37	0.73	1

b

p-value : 10^{-70}	HSC	Erythroid	Jurkat	CEM
HSC	1	0.32	0.53	0.67
Erythroid	0.32	1	0.24	0.27
Jurkat	0.53	0.24	1	0.82
CEM	0.67	0.27	0.82	1

c

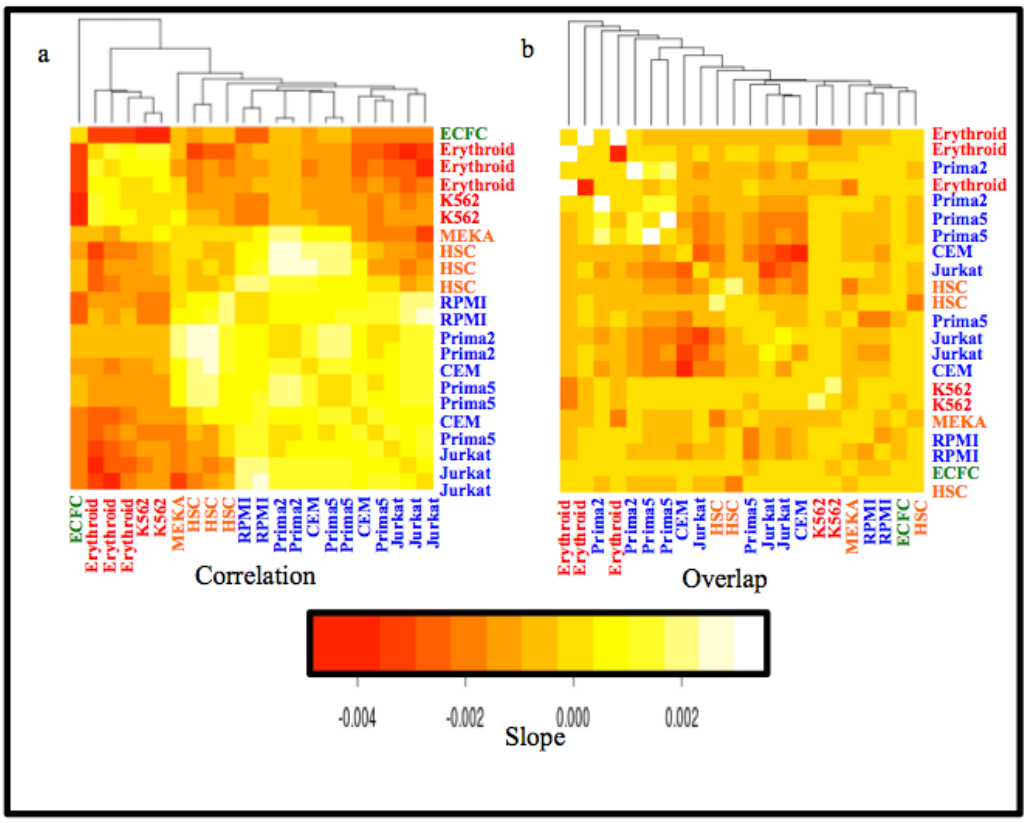


4.3.2 Complete Assessment of Overlap and Correlation

Using the pilot set of data, relative overlap was shown to be independent of p-value and correlation was shown to diverge disparate data sets with an increase in p-value. I applied both measures to the full UDM to determine if these observations could be generalized to all TAL1 ChIP-Seq data sets in the analysis. To confirm the relations, I looked at the slope of the overlap and correlation as a function of p-value for the

relationships between all 22 data sets.

Figure 10: Heatmaps representing the slope of the overlap and correlation as a function of p-value. The data sets were clustered using single linkage. a.) Representation of the rate of change of the correlation, as a function of p-value, between data sets. b.) represents the rate of change of overlap, as a function of p-value, between data sets.



The results, seen in Figure 10, rebuke the idea that p-value has no impact on relative overlap. While the Jurkat CEM overlap and Jurkat Erythroid overlap slopes are similar this is not true for other conditions. Additionally, while some replicate data sets, such as K562, have slightly increasing overlap as a function of p-value, there is no strong relation between context and overlap slope. Conversely, the results for the correlation, Figure 10 a, show that the change in correlation is directly linked

to the cellular context. It can be seen that using the rate of change in correlation is sufficient to separate the ECFC, Erythroid and T-ALL contexts. This direct relation indicates that, when using dissimilarity as a measure of distance, adjusting p-values may change the distance between clusters but should not change the composition of clusters. This relation makes correlation of the UDM a more robust measure relating the distance of cell types from one another since, unlike pairwise overlap, the changes in slope are tied to biological contexts.

4.4 Hierarchical Clustering

Following the confirmation that correlation would be a resilient measure of distance, I investigated the parameters that would be best suited for hierarchically clustering the 22 TAL1 ChIP-Seq data sets. The parameters that I investigated were distance metrics and linkages. For distance metrics, I compared the dissimilarity, i.e. 1 - correlation, Euclidean, Manhattan, and P-norm (where the p was 10). I also compared three linkages: single, complete and average linkage.

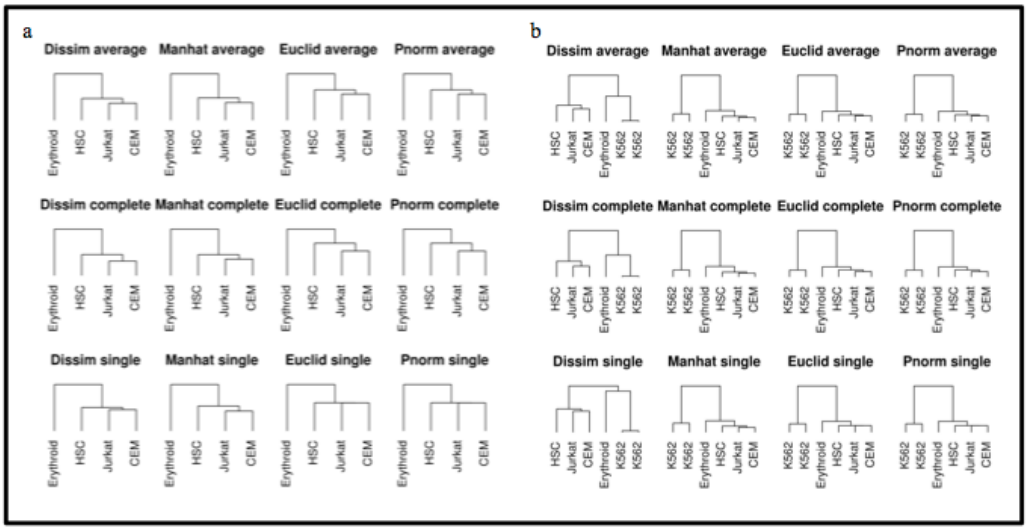
4.4.1 Clustering the Pilot Scale Subset of Data

First, I applied hierarchical clustering with a series of parameters to the pilot UDM, to determine which parameters formed distinct groups that were biologically related. This small scale comparison was done to limit the chances of over training the selection of parameters for all data sets. The result of the 12 parameter combinations applied to the pilot data can be seen in Figure 11 a. Using the four data sets, it was not sufficient to determine which linkage and distance metric would be best suited to differentiate cellular conditions. The two T-ALL data sets grouped as expected for each of the 10 of the 12 combinations. Only the results using the Euclidean and Pnorm with a single linkage were ambiguous.

To help elucidate which parameters would work best, without applying all of the data sets, I added two additional data sets to the analysis. The two data sets were replicates of the erythroid cancer cell line K562. These data sets were expected to

cluster closely with the erythroid data set. The results of the hierarchical clustering with the 6 data sets can be seen in Figure 11 b. For each of the distance metrics, dissimilarity provided the clearest distinction between the T-ALL context and the other contexts.

Figure 11: The result of applying hierarchical clustering to the AFS combined data was assessed using a dendrogram. The Figure displays the 12 combinations of interest. a.) Representations the clusters formed using the 4 key members of the pilot scale analysis. b.) The impact of including two additional data sets (K562 replicates) in the clustering



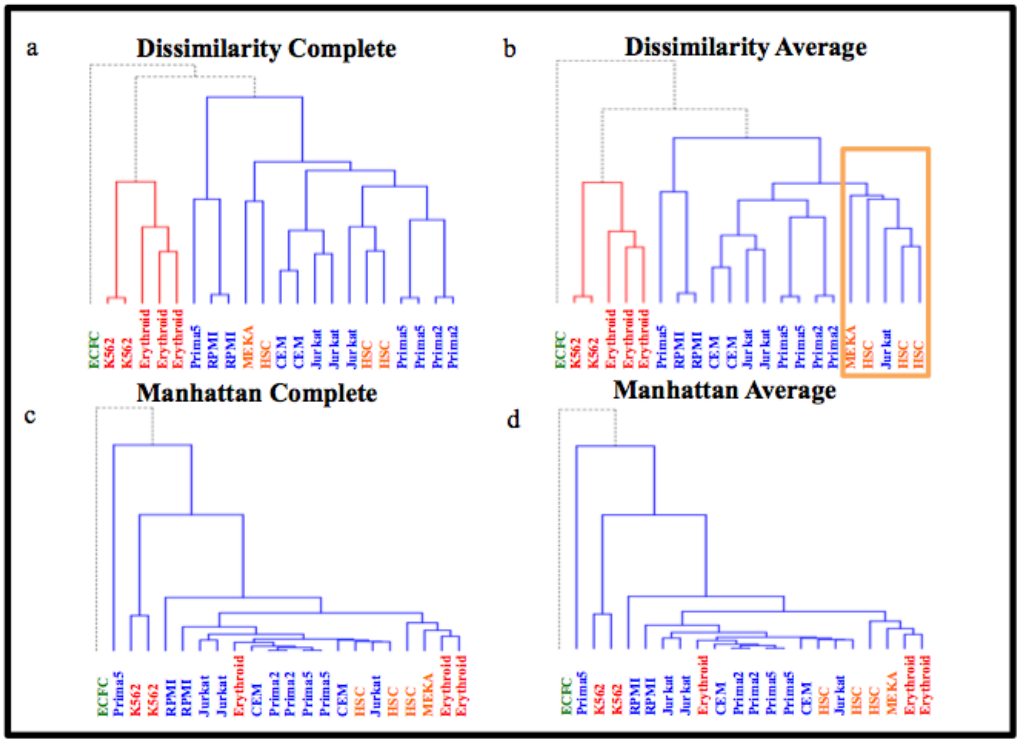
4.4.2 The Full Clustering Analysis

The pilot sized analysis was sufficient to identify dissimilarity as the most likely measure to separate the contexts, however, its scope was too small to indicate whether the complete or average linkage should be used in the full analysis. In order to confirm that Dissimilarity was the best approach I compared it with the Manhattan distance, using both average and complete linkage. The results can be seen in Figure 12.

It is clear that the Manhattan distance, as seen with the addition of K562 in the pilot case, failed to satisfactorily group a more diverse range of data sets. The

use of dissimilarity separated the ECFC, Erythroid, and T-ALL contexts using the complete linkage. When the average linkage was used the HSC and Megakaryocyte cellular conditions formed a distinct sub cluster within the T-ALL. Thus, without normalization, Dissimilarity coupled with average linkage provided the most biologically relevant separation of the data sets.

Figure 12: The four parameters used in hierarchical clustering. Applied to the unified density matrix before quantile normalization. a.) complete linkage using Dissimilarity as a distance. b.) average linkage using dissimilarity as a distance. c.) complete linkage using Manhattan distance. d.) average linkage using Manhattan distance.

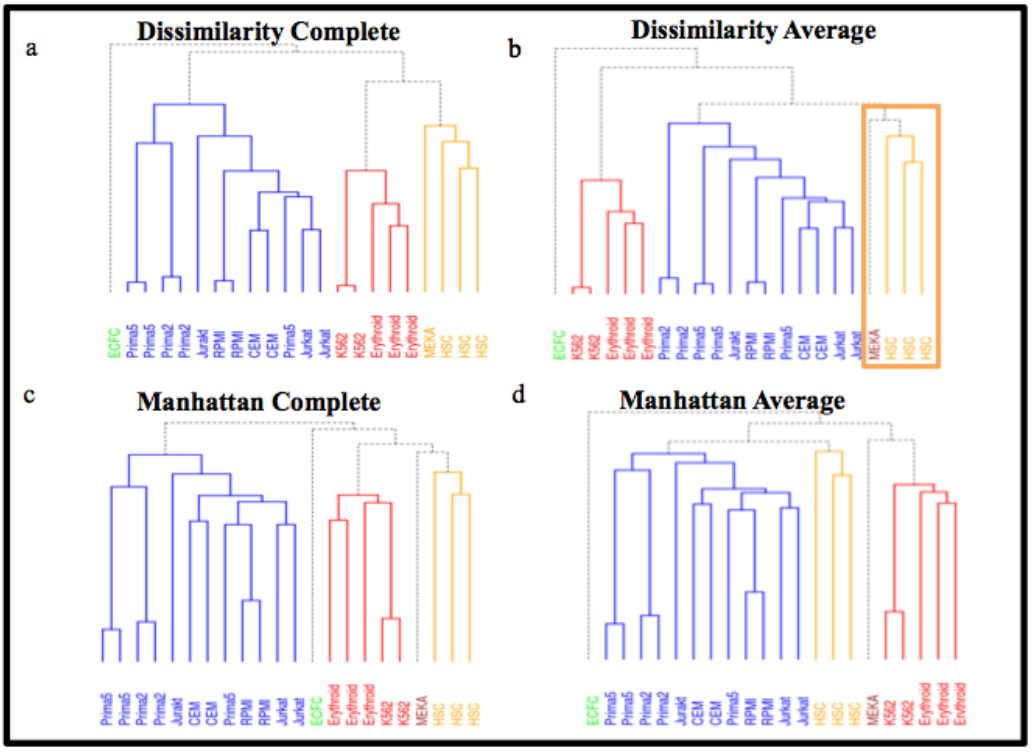


Correlation, unlike the Manhattan, normalizes the distance between two data sets by the variance of each data set. Since the data sets contributed different numbers of peaks and had highly varying total read counts, there was potential for bias based on these technical factors. To make a more even comparison between the two distances, I applied hierarchical clustering to the UDM after it had been quantile normalized.

The results can be seen in Figure 13. From this Figure it is clear that both Dissimilarity and the Manhattan form clusters for at least four of the five contexts of interest. While the structure of each of the clusters remains stable (with the exception of the inclusion of megakaryocytes with the HSC group), the relationships between clusters change as the parameters used in the clustering are altered. With the use of the complete linkage the HSC is grouped closer to the Erythroid, while with the average linkage it groups closer to the T-ALL. When using average and complete linkage with the dissimilarity distance and the complete linkage with the Manhattan distance the HSC and megakaryocyte cluster closely. Conversely, the clusters formed with average linkage with Manhattan distance indicate that the megakaryocyte may be closer to the Erythroid than to the HSC.

These results show that the use of hierarchical clustering coupled with quantile normalization was sufficient to form the clusters that were expected using a variety of parameters, however, due to the high variability of the positioning of those clusters chosen no conclusive relation could be made on the distance between all clusters. In other words, it is clear that the Erythroid, ECFC and T-ALL data contexts are different from one another, it is not clear whether the HSC and Megakaryocytes are closer to T-ALL or Erythroid, or whether the HSC and Megakaryocytes are closer to one another than any other context.

Figure 13: The four parameters used in hierarchical clustering. Applied to the unified density matrix following quantile normalization. a.) complete linkage using Dissimilarity as a distance. b.) average linkage using dissimilarity as a distance. c.) complete linkage using Manhattan distance. d, average linkage using Manhattan distance.



Without quantile normalization the use of dissimilarity with an average linkage formed the most clusters pertaining to cellular contexts. The contexts separated were the Erythroid T-ALL and ECFC. With quantile normalization both the Manhattan and Dissimilarity, using both the complete and average linkage, clustered the data into four to five distinct groups that aligned with each of the biological contexts of interest. Since quantile normalization decreased the variations between parameters it was used going forward with the PCA analysis.

4.5 Principle Component Analysis

While I have shown, using hierarchical clustering, TAL1 binding is sufficient to group data sets, I have not demonstrated the ability to partition peak regions that are most important to each cluster. Partitioning peaks is a prerequisite of applying downstream functional analysis, such as DNA Motif DENovo and Gene Ontology (GO). For the third stage, I applied PCA to the quantile normalized UDM. The goal of applying the PCA was to observe if the eigenvectors could be used to separate cellular contexts from one another and if so to partition the data sets that contributed to the separations.

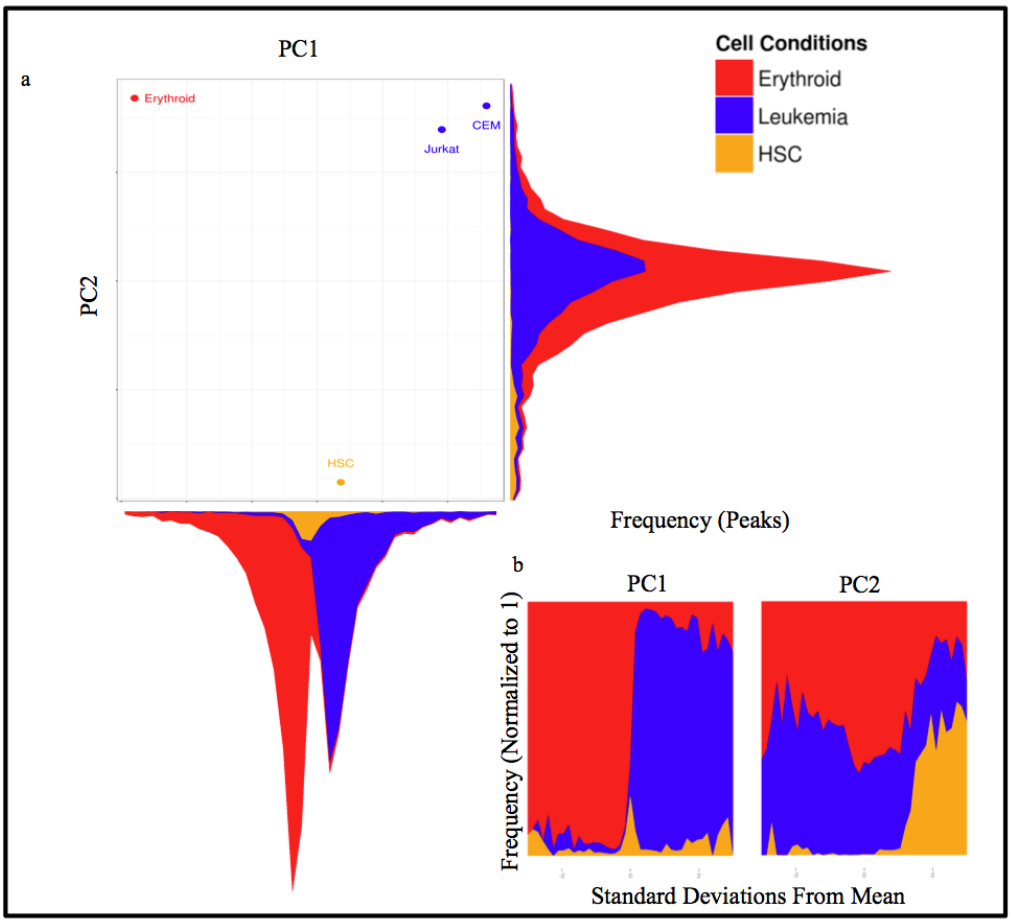
4.5.1 Application to the Pilot Scale AFS

Each value in the PC relates to the weighting of a single peak along that axis. The weighted data set position of the first PC, found by taking the dot product of the first principle component and the normalized heights of each peak, can be seen in Figure 14. The first principle component separates the Erythroid and T-ALL context, and the second principle component separates the differentiated hematopoietic cells from the HSCs. The result of the erythroid leukemic separation is clearly bi-modal. It is clear that along the first principle component, within 1 standard deviation from the mean, there is a switch from a higher proportion of Erythroid to Jurkat, going left to right. In addition to separation of the Leukemic and Erythroid along the first PC, the HSC can be separated from peaks that are not important to HSC along the second.

This second separation shows one of the key advantages of using the PCs of the UDM when identifying regions that are important to specific conditions. Whereas the TF TAL1 in the differentiated conditions is strongly bound to key locations, in HSC TAL1 binding is more ubiquitous and weakly bound. Because TAL1 is weakly bound in HSC, few regions are identified as enriched when peaks are called. However, by observing the raw read density of the HSC data set in a wider set of TAL1 peaks, we can identify a larger set of regions that are important in the HSC context. Additionally we can identify regions that are explicitly unimportant in the HSC context. Having a scale of importance, rather than a binary definition, allows for a differential

comparison of each context using a varying level of stringency. This is especially useful for data sets such as HSC, which are expected to have a large number of binding locations that are marginally bound but shared with other contexts.

Figure 14: The projection of the quantile normalized unified feature matrix onto the first two principle components. a.) the two principle components and their contribution histograms. n.) the normalized contribution histograms, where each bin is normalized to 1.



4.5.2 Application to the Full Scale AFS

Since it was clear that the contexts in the pilot data could be separated using PCs, I applied PCA to all 22 data sets. Five of the PCs of the PCA crossed with the normalized UDM can be seen in Figure 15. The first PC separates the Erythroid and T-ALL contexts and the third separates the endothelial context from the differentiated hematopoietic. The projection of the peaks and their sources onto the first and third PC can be seen in Figure 15, and more projections can be seen in Figure 16. It is apparent that a switch in the prevalence of peak contribution sources occurs around the mean in both dimensions.

Along the first eigenvector, the region that is less than the mean is dominated by peaks from the T-ALL data sets and the region above the mean is dominated by the Erythroid context. For the third principle component there are very few peaks contributed from ECFC on the left while half of the peaks on the right come from ECFC. The projection onto the fifth eigenvector shows that the HSC and megakaryocyte data sets are weighted higher in the region greater than the mean and along the seventh eigenvector the HSC and megakaryocyte conditions are split negative to positive.

While the peaks that contribute most to the T-ALL, Erythroid, and ECFC conditions may be separated using a single vector, those that contribute most to the HSC and Megakaryocyte require two dimensions. The combination of the third and fifth vector separates the HSC and Megakaryocyte, first from the differentiated hematopoietic cells, and subsequently from the ECFC. Using a combination of the third and seventh eigenvector the Megakaryocyte and HSC can be separated from one another. The results for separating the megakaryocyte using the third and seventh dimension may be lackluster since neither dimension clearly differentiates it from the ECFC.

Figure 15: The first and third principle components, found using the full AFS defined with a statistical cut off of $10^{-}5$. Each projection is accompanied by the contribution histogram of peaks along its PC. a.) the representation of the data weighted by the first and third principle components. b.) the projection of peaks and their sources along the third eigenvector. c.) the projection of peaks and their sources along the first eigenvector. Leukemia specifies T-ALL cell types, MEKA is the abbreviation of megakaryocytes

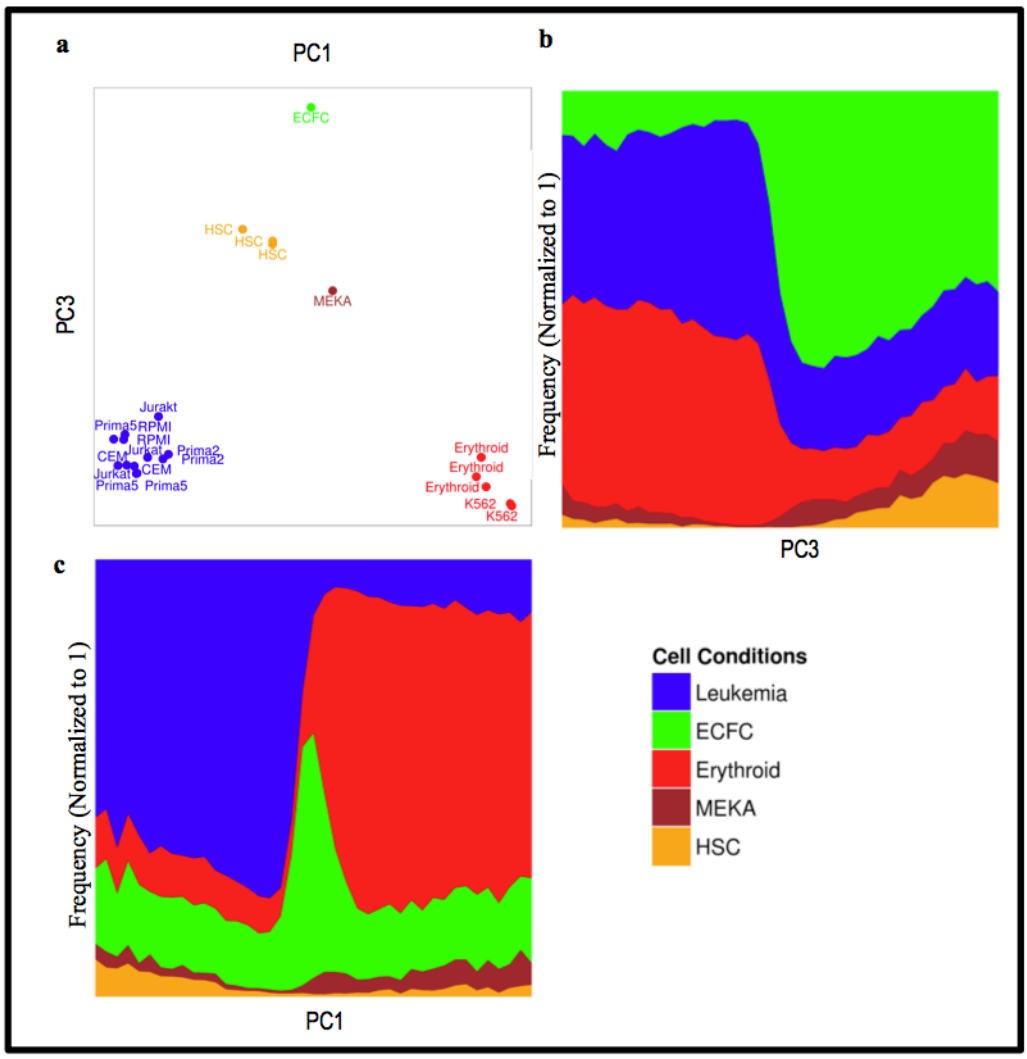
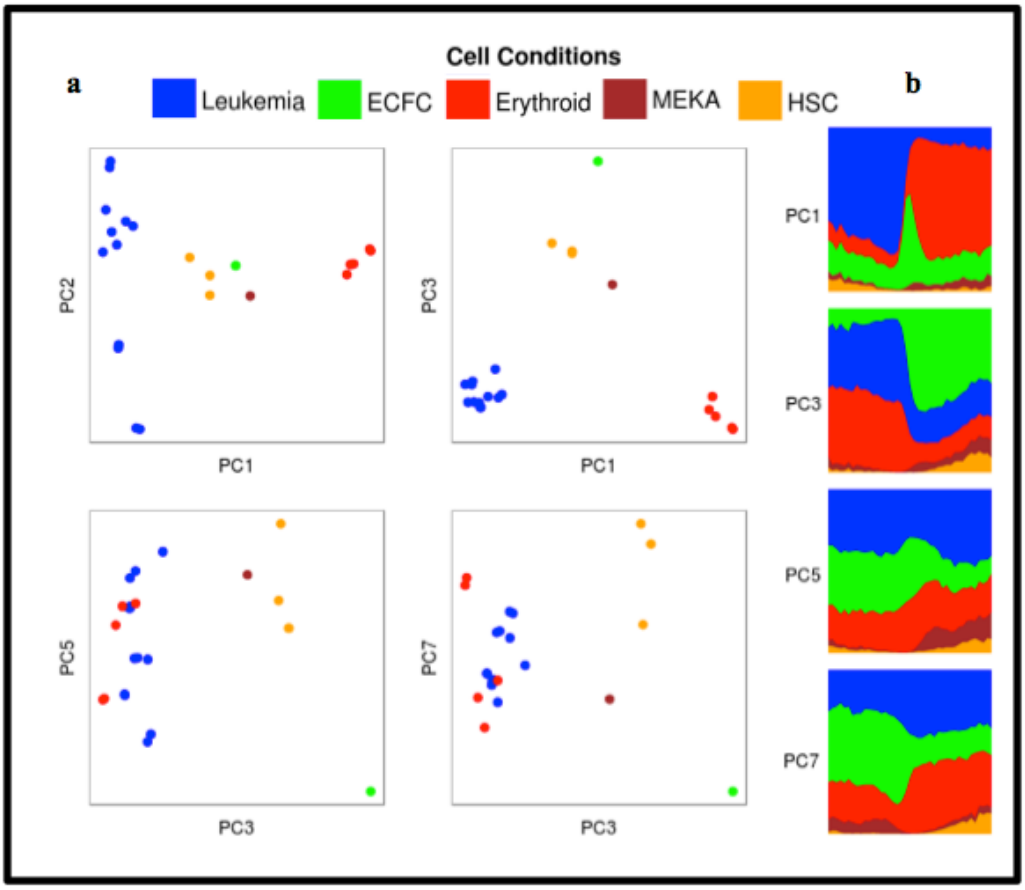


Figure 16: The positions of members of each context on the first, second, third, fifth and seventh PC. a.) the positions of each data set weighted by the eigenvectors. b.) the projection of peaks along each of the eigenvectors of interest.



Using the pilot subset of the data, I showed that the Leukemic peaks could be isolated from the erythroid. The contribution histogram was clearly bi-modal with one peak being produced by T-ALL peaks and the other from Erythroid. Once the efficacy of this comparison was shown using the four data sets, I extended it to all 22 datasets. The relation between erythroid and T-ALL peaks was retained along the first principle component and additional information, which could be used to isolate peaks important in ECFC, HSC and megakaryocyte, was identified. Now that I have shown that the data sets can be separated and their peaks given a quantifiable value that can be used to measure distance, rank, and partition, I had to confirm that the

results were not heavily influenced by factors apart from the binding of TAL1.

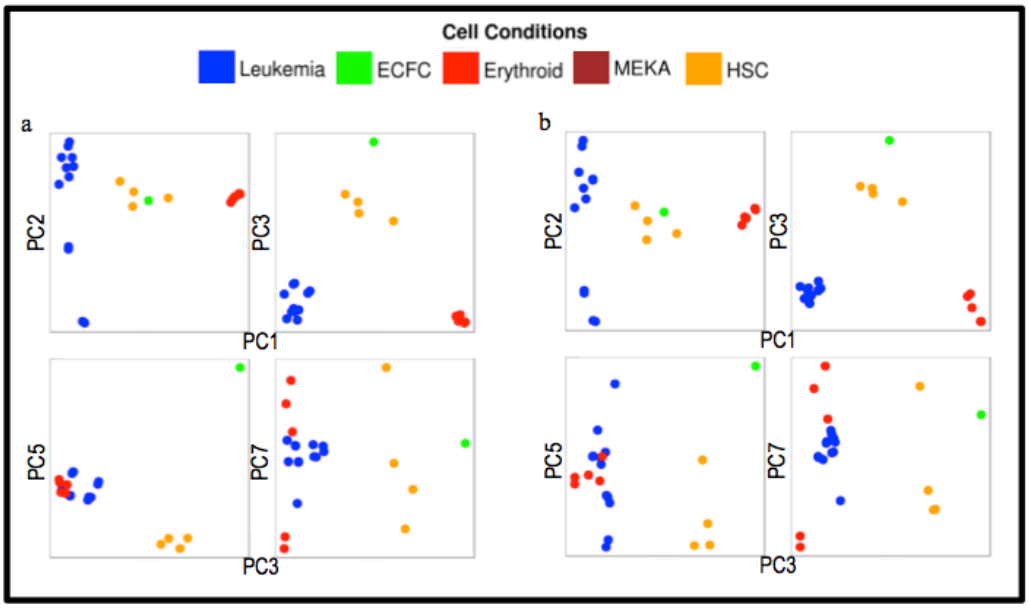
4.6 The Stability of the Principle Component Analysis

The final stage of the analysis was to determine how stable the separation of data sets using PCA was. Stability in this context refers to the resilience of the observation of underlying biological relations to changes in the technical approach, i.e. while some peaks may not be included in the T-ALL category if RPMI is removed from the analysis, as a whole the T-ALL context should still be differentiated from the Erythroid. In the following section, to determine if the analysis was stable, I investigated four sources of instability and bias.

4.6.1 The Effect of Changing the Backgrounds

In the previous section, to form the AFS I was using peaks that were generated using a combined control. In order to determine the impact of the background on the clusters found I repeated the analysis calling peaks using individual controls provided by each data set and no controls. The results can be seen in Figure 17. While there was some movement within each of the clusters, each unique cluster remained identifiable along the same principle component. This stability in the face of alteration in the analysis represents the underlying stability of the natural clusters in the ChIP-Seq data.

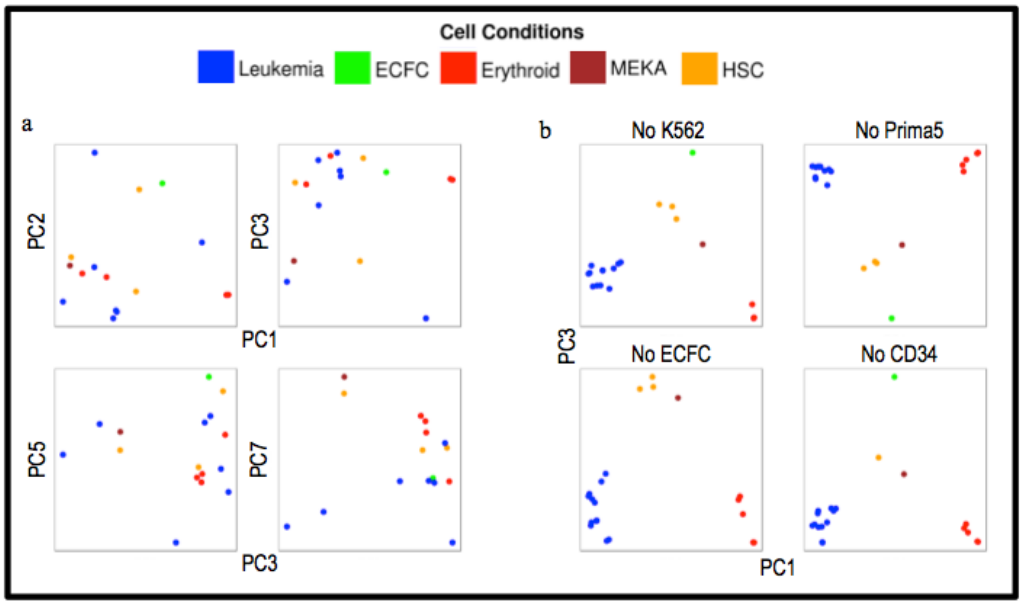
Figure 17: The impact of using different controls when defining the AFS. a.) using the individual control provided in each data set when calling peaks using MACS. b.) using no control when calling peaks



4.6.2 The results of using the control as the Treatment

To further assess the impact of the background on the analysis, and to confirm that it was actually TAL1 binding driving the results, I swapped the treatment data sets with the 16 individual control data sets during the unified density matrix generation phase, using the same AFS found using the 22 treatment data sets. From the results, shown in Figure 18 a, the clusters formed using the treatment were not reproduced using the control. While there was close proximity between replicate data sets there was not a relation between cellular contexts as was seen in the treatment.

Figure 18: Assessing the stability of the separation dimensions. a.) the effect of swapping the control with the treatment on the principle components of interest. b.) the impact of removing the largest data set from each cluster on the first and third principle component.



4.6.3 The Impact of Altering the Data Sets Contributing to the Analysis

To ensure that the clusters that I found using the treatment data were not dominated by their largest members, I removed the data set that contributed the most reads to each condition and repeated the analysis. While there was movement within each of the clusters the clusters themselves remained stable and in the same distance from others. The results can be seen in Figure 18b.

4.6.4 Principle Component Correlation With Technical Factors

To determine if any of the principle components were dominated by information concerning the number of reads in a data set or the number of peaks it contributes

to the AFS, I looked at the correlation between data sets and these two technical factors. None of the PCs were highly correlated with either factor. The results can be seen in Table 2. The highest correlation that was seen was 0.4 and it was not in any of the dimensions that were found to separate cellular contexts.

PCs	Reads	Peaks
PC1	-0.329	0.139
PC2	0.349	0.319
PC3	-0.119	-0.06
PC4	-0.167	0.323
PC5	0.135	0.2
PC6	-0.188	0.336
PC7	-0.061	-0.02
PC8	0.661	0.062
PC9	-0.105	0.012
PC10	0.158	0.092
PC11	-0.036	-0.154
PC12	-0.104	0.023
PC13	-0.324	-0.582
PC14	-0.052	-0.173
PC15	0.029	0.02
PC16	0.044	-0.05
PC17	-0.034	-0.02
PC18	-0.08	-0.185
PC19	-0.139	-0.241
PC20	0.18	0.231
PC21	0.126	0.254
PC22	0.122	0.487

Table 2: Summary of the correlation of non biological factors and the principle components found for the AFS analysis of the combind control condition

First I have shown that the use of controls when calling peaks has little impact on the final clusters found. I hypothesize that this is due to the combination of many data sets. For a peak to be ranked as highly important in the leukemic context, for example, there must be a high pile up in 12 data sets, and little pile up in the non leukmeic data sets.Then, using the control data set to generate the unified density matrix, I show that the clusters formed using the treatment were not apparent. Third, to ensure that no single data set dominated the analysis I have also shown that the removal of the larges data set from each cluster did not impact the clusters formed. Finally to ensure that the PCs chosen to separate the contexts were not dominated by signal pertaining to technical factors I looked at the correlation between each PC and the number of reads and peaks in each data set. The next step was to investigate the stability of the clusters as the stringency of peaks was altered.

4.7 The Impact of Changing P-Values

In the comparison of the correlation and overlap I have shown that the correlation between data sets is dependent on the p-value, and that the change in correlation is related to the context of the cellular condition, i.e. similar data sets correlate more with increasing p-value and disparate data sets less. To determine the effect of p-value on the separation along eigenvectors, I extended the PCA to investigate UDMs generated using a series of AFSs defined using MACS scores ranging from 5 to 72.5 at intervals of 2.5. The relation between p-value and the size of the AFS can be seen in Figure 19 a. As would be expected from the pilot analysis the size of the AFS was non-increasing. Figure 19 b shows the log mean square difference between the pairwise data set distances for the first four PCs, as the cut off increases. I did this to estimate how stable the relation between data sets were. If there was no change in the overall distance between datasets when the statistical threshold was increased, it can be assumed that the relation between data sets is independent of threshold cut off. From the figure it is apparent that the difference settled between 2.5 and 3, and that the time to settle increased with the increasing principle components. By a MACS score of 25 the first three PCs had settled while the fourth settled after 30

The impact of changing the cut off on the separation using the PCs can be seen in Figure 20. As the cut off increases the separation between T-ALL and Erythroid along the first PC remains consistent. For the remaining PCs however there are significant alterations. The separation between the differentiated hematopoietic cells and the endothelial cells seen in the third principle component becomes apparent in the second principle component for MACS scores greater than 10. The variation between HSC and the endothelial condition becomes apparent in the fourth PC as the cut off rises above 15.

Apart from the trends in separation, the overall contribution of some contexts are seen to diminish. At a cut off of 25 there is an observable diminishment of the overall representation of HSC in all data sets. This loss can be explained by referring to Figure 14 in the pilot data, which indicates low MACS scores for the peaks called in the HSC conditions. Thus as the cut off rises, the proportional contribution of HSCs to the AFS decreases. Despite this the clusters still stabilized, indicating that the read density of the HSC conditions in peaks contributed by data sets other than HSC remained stable.

Selecting a p-value cut off beyond the settling point for the PCs we wish to investigate can help ensure that the analysis is resilient to increases in stringency. This relation has to be weighted with the loss of information with increasing cut off. For this further analysis, I chose to compare the results using a MACS score cut off of 20, the results can be seen in Figure 21.

Figure 19: The relation between MACS cut off scores and the settling of key PCs and the size of the AFS. a.) represents the log mean squared difference between two cut off values separated by a MACS score of 2.5. b.) the proportion of the AFS that remains as the MACS cut off score increases with points indicting the remaining portion after the macs score has been increased to 10, 20, and 30.

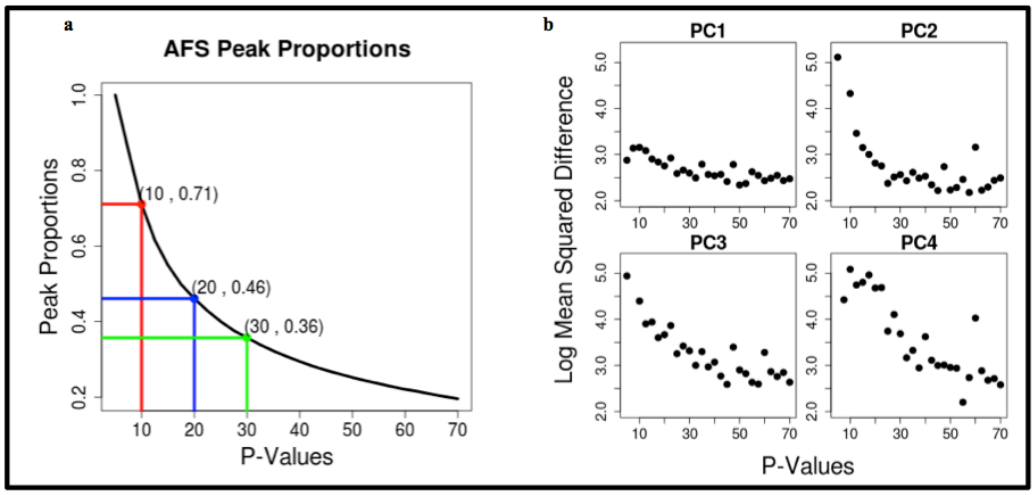


Figure 20: Comparing the PCs of interest as the MACS cut off is increased from 5 to 25. a.) is used to compare the PC weighted positioning of each data set at a macs score of 5 and 20. b.) shows the contribution histogram of along the first four PCs as the MACS cut off is raised from 5 to 25 at in intervals of T5

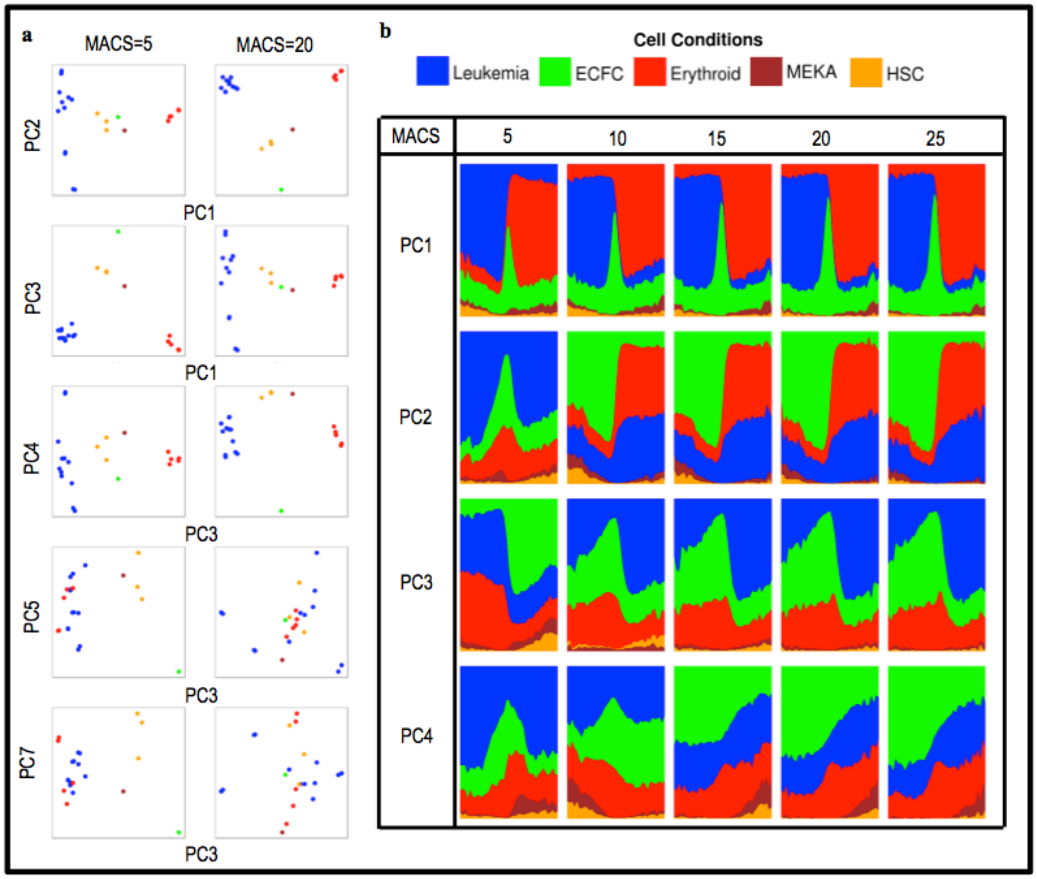
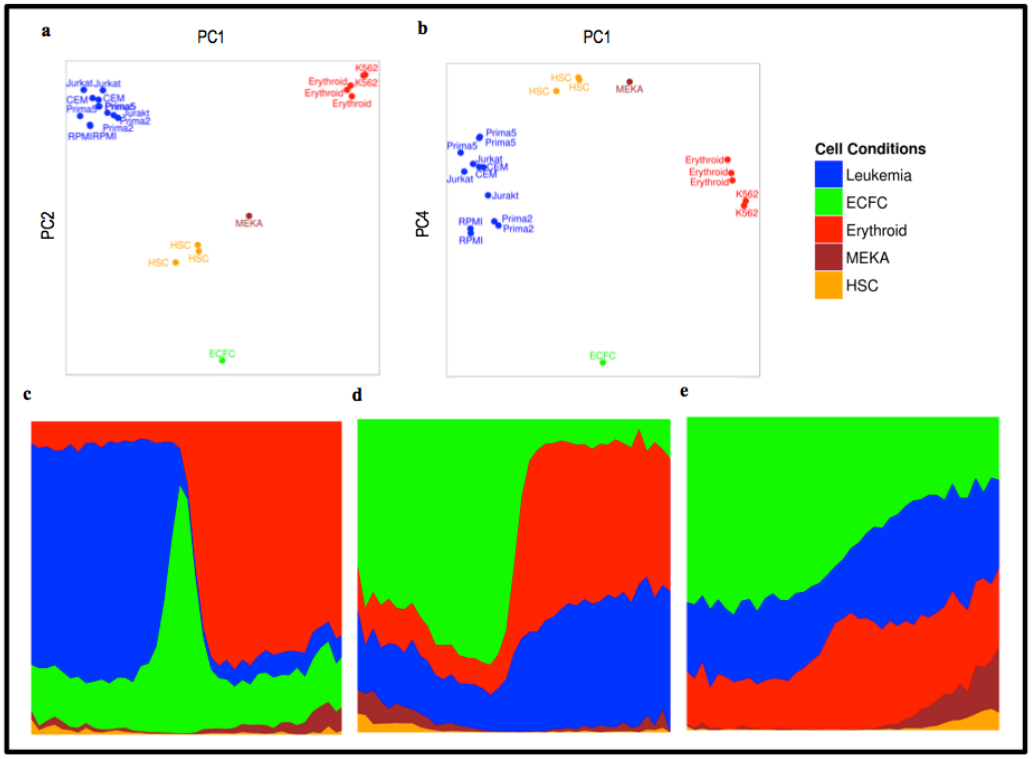


Figure 21: The three PCs that offer clear separation between four cellular contexts. a.) the weighted positing of the normalized unified density matrix along the first and second PC. b.) the weighted positioning along the first and fourth PC. Subplots c, d, and e respectively depict the unity contribution histograms along the first, second, and fourth PC.



4.8 Summary

Through this pilot scale analysis I have confirmed that using overlap as a measure of cell type similarity is biased based on the number of overlaps being used. Looking at the rate of change of overlap between all data sets as a function of statistical cut off, I have shown that there is no biological relation between the effect of changing stringency on overlap. This does not mean that there is no change in the overlap rate. In several cases the overlap of data sets from different conditions increase while the overlap between similar data sets decrease. This case is clearly shown with the relation

of the erythroid data sets. As the cut off increases, erythroid replicates overlap more with the jurkat data set that came from the same experiment and less with K562 (a condition where the function of TAL1 is shared with erythroid). Conversely with correlation I have shown that using the rate of change, I can cluster the data sets using hierarchical clustering and form the expected biological contexts.

For hierarchical clustering, it was clear, using an extended pilot set, that dissimilarity was the best distance measure. This was confirmed when the analysis was extended to all 22 data sets. After quantile normalization had been applied to the unified density matrix, however, both the Manhattan distance and Dissimilarity were shown to be effective. The results using different parameters showed that the ECFC, Erythroid, and T-ALL conditions were distinct from one another. The relationship between HSC and Megakaryocyte was less clear.

When I applied PCA to the unified density matrix I found that each of the contexts of interest could be isolated from one another using either one or two PCs. The stability of these separations, in respect to small alterations in the technical approach, indicated the strength of the underlying biological information.

Increasing the stringency of peak detection changed the results of the PCA slightly. While the first PC remained stable, information from the third principle component became apparent in the second. After the cut off had increased to a MACS score of 30 the analysis had stabilized. In the following sections I show how partitioning the data using the PCs provides biologically relevant information when functional analysis techniques are applied.

Chapter 5

Biological Results

5.1 Overview

The final result of the technical portion of my analysis was the separation of peaks into four cellular contexts. In this section, I have taken those TAL1 peaks that were separated and studied their characteristics to identify TAL1 DNA binding properties that are specific to each cellular environment. My aim was to confirm that these isolated peaks were biologically relevant, in each of the four contexts. The two key functional characteristics that I investigated were the DNA motifs (i.e. DNA binding sites), and TAL1 associated genes. These results are found in Section 5.4 and 5.5 respectively.

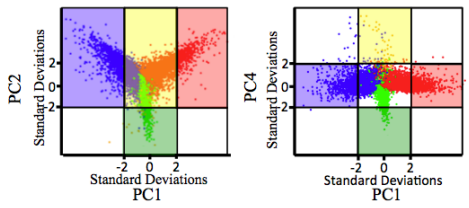
For motif analysis I first determined the frequency of Ebox occurrence in the four cellular contexts. Subsequently I performed a *de novo* search for motifs to uncover over represented sequences in each context. For the gene association, I used the GREAT algorithm and applied gene ontology to the associated peaks [65]. Additionally I analyzed the gene expression data upon TAL1 knockdown in the Jurkat (T-ALL) cell line to study the relationship between the specific motifs I have identified and genes that are functionally regulated by TAL1 in T-ALL.

5.2 Selecting Context Specific Peaks

The four cellular contexts that I identified, using the principle components (PCs) of the UDM, were Erythroid, T-ALL, Hematopoietic Stem Cell (HSC), and Endothelial Colony-Forming Cell (ECFC). The final peak detection p-value cut off was set to 10^{-20} (corresponding to a MACS score of 20), which was the point that the first four PCs began to stabilize. Figure 22 a depicts the distribution of peaks across the three PCs that separated the four contexts. As expected, the overall number of peaks that fell into each of the contexts decreased as the cut off became more stringent. This relation can be seen in Figure 22 b.

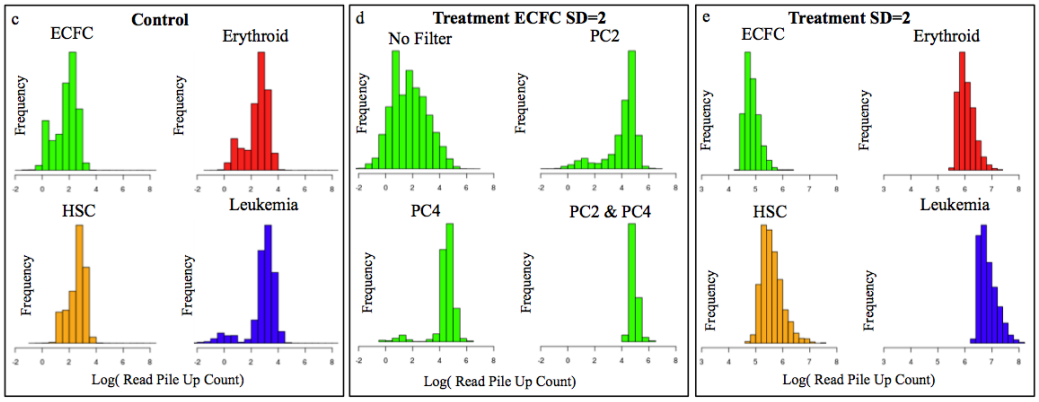
Figure 22: The definition of peak sets for each of the four contexts, Erythroid, T-ALL, HSC and ECFC. a.) Depiction of the three PCs that were used for the separation of these contexts (the first, second and fourth). The range in three dimensions for each of these data sets is colour coded. Erythroid in Red, T-ALL in Blue, ECFC in Green and HSC in Yellow. b.) A table showing the total number of peaks available in each peak set as the cut off (represented as colour coded ranges in subplot a increases from one to four standard deviations. c, d, and e.) Histograms of the max log normalized peak heights from the UDM. c pertains to the distribution of normalized heights in the control without isolation, d depicts the effect of combining multiple PCs for the isolation of ECFC peaks using a standard deviation cut off of 2. e.) represents the final histograms of each set peaks

a



b

Cut Off	PCA Seperated Peaks				Background	
	Erythroid	Leukemia	HSC	ECFC	Center	All
1SD	4,274	3,958	2,821	2,297	11,829	62,573
2SD	1,694	1,448	1,210	651	11,829	62,573
3SD	733	635	596	197	11,829	62,573
4SD	325	295	320	58	11,829	62,573



The first PC separated the Leukemic and Erythroid conditions. The second PC separated the differentiated hematopoietic data sets from the ECFC. The fourth PC separated the HSC and the ECFC. Peaks were isolated by selecting those that fell two SDs from the mean along the PC that separated them and those that were no greater than two standard deviations along any of the other separating axes. The

purpose of the second restriction was to ensure that all isolated sets of peaks were mutually exclusive. The background that I used in the *de novo* search for motifs in each cellular environment (see section 5.4) was defined as the central regions of the first, second and fourth principle components. Peaks that fell within 0.25 SD of the mean, along each of the separating axes, were included in this context.

To confirm that the peaks I isolated were enriched in the treatment data sets and not the control, the log read pile up counts were compared. The summed normalized pile counts from the control data sets can be seen in Figure 22 c. The normalized heights have the majority of their signal bound between 0 and 40. Following the assessment of the control data, I compared the distribution of heights for the treated ECFC peaks using a combination of filters. This comparison can be seen in Figure 22 d. In the case where the treatment was not filtered, it was apparent that the variance of normalized pile up counts was greater in the treatment than in the control. The goal of filtering was to take advantage of the increased variance and to select those peaks that fell outside the range of the control data. Without filtering, the majority of peaks in the AFS had a normalized read count less than 40, i.e. within the range of the control signal.

From Figure 22 a, the ECFC context can be isolated using both the second and fourth PCs. The separation of ECFC along either of these PC dramatically decreases the number of peaks that had a normalized height less than 40. However, it can be observed that there is still a subset of the peaks that have a normalized height within the range of the control. Combining the second and fourth PCs eliminated the majority of this sub population.

As can be seen in Figure 22 b, changing the standard deviation cut off effects the total number of peaks in each peak set. For example, increasing the cut off from one SD to two decreases the number of ECFC peaks four fold. Using this higher cut off however, guarantees that there will be no peaks in the ECFC context that have a normalized height less than 40. This is in contrast to the control set, which had no peaks with a normalized height greater than 40. The Erythroid peak set, along with the Leukemic and HSC were each separated using a single PC, refer to Figure 22 c. From comparing Figure 22 c and d, the resulting separation provided peaks that

were outside of the range of the normalized heights of the control in each respective context.

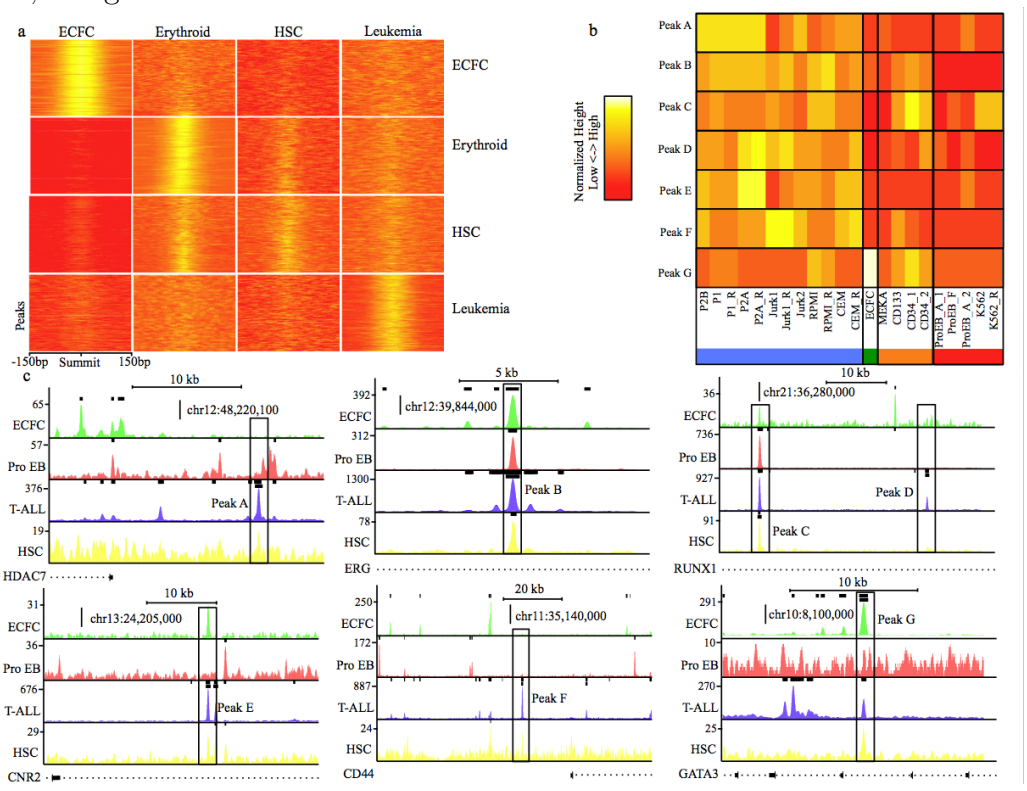
5.3 Quantifiable Results

An important advantage of the method I have developed for comparative analysis is that rather than just comparing the binary peak presence for qualitative comparison, it allows for a quantifiable analysis since it retains peak height information for the analysis. Figure 23 depicts the quantifiable nature of this analysis. The read heights for each of the peaks in the AFS can be seen for each separated context in Figure 23 a. The ECFC, Erythroid, and Leukemic peak sets each have a prevalence of read density around their distinct summits, for each of their corresponding raw data sets. This prevalence is an indication that the PCA method separates peaks successfully based on the combined heights of each of the contributing data sets. In contrast, when compared to the erythroid and leukemic contexts, the HSC a less distinct read distribution for the their raw data sets around their isolated peak summits. This limitation was anticipated, since 57% of the peaks that were indicative of HSC came from these two contexts (see Figure 21 e in the previous section). While only 22% of peaks in the HSC separated context come from a HSC data set, those 22% represented 65% of all HSC peaks. Contrast this with 1.6% of HSC peaks falling in the opposing context (PC4 ¡ -2), resulting in a 41-fold change in the proportion of peaks at +/- 2 SDs across the fourth PC. This difference indicates that the fourth principle component successfully isolates sub threshold HSC peaks from ECFC peaks.

Note that these read height distributions come before the quantile normalization. In order to compare the distributions across contexts the read pile-ups had to be summed, resulting in a single height for each peaks. To assess the normalized height across all members in each context, six locations impacted by TAL1 were chosen for investigation. The stacked heights (which have not been normalized) for each of the peaks can be seen in Figure 23 c. While HDAC7 and CD44 both had associated peaks that were unique to a single context (T-ALL) the other four peaks were identified by MACS, or were visually apparent, in two or more contexts. The most interesting case

is ERG (peak B). ERG had a peak identified by MACS for each context, however, it was determined using the PCA analysis to be most important in T-ALL. Referring to Figure 23 b, it can be seen that peak B, while having signal in each of the contexts, had a consistently higher signal in all 12 of the leukemic conditions whereas the signal for peak B is much more variable between the 4 HSC conditions. This indicates that the PCA analysis method does not just take the heights into account but also the variance between data sets within each of the contexts.

Figure 23: The quantifiability of the PCA analysis. a.) represents the read pile up +/- 350 bp from the peak summit for the four contexts. The regions depicted along the y axis are the sorted peak sets, the x axis represents the height distributions around the summit for each peak sets. b.) is a heat map representation the normalized UDM of the seven peaks. These values are the input to the PCA, for the seven peaks. The blue band represents peaks that are from the Leukemic context. The green band represents the Endothelial peaks. The yellow band represents the HSC peaks. The red band represents the Erythroid peaks. c.) is the raw read stacked pileup for each of the contexts (snapshot from the UCSC Genome Browser). For each context two peak sets (black bands above the pile up signal) are included. The top set of bands represents the union of peaks called using MACS for all data sets in the context. The lower band represents the peaks identified to be most significant for each specific context, using PCA.



5.4 Motif Analysis

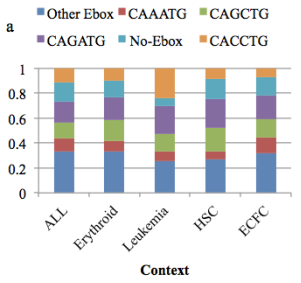
While the parameters used in the PCA analysis guaranteed a minimum normalized height for differentiating peaks and the results were observed to be quantitative, we have not yet shown any distinctive biological information regarding the discriminant contexts. Here I show how these discriminating peaks can be used to determine whether specific DNA binding sequences (i.e. motifs) may be differentially enriched between the different cellular environments. For the full set of results refer to Appendix E.

5.4.1 Ebox Preference Between Conditions

Figure 24 a shows the proportion of peaks with specific Eboxes in each of the four contexts. Overall, through the analysis of the discriminant peak sets, I found that there is very little difference in the proportion of the different Ebox variants between the cellular contexts with 2 exceptions. I found that the CC Ebox variant is disproportionately overrepresented in the leukemic environment. This increase appears to be at the expense of a lower representation of TAL1 binding sites that do not contain any Eboxes. These results suggest that there are fundamental differences in the DNA binding specificity of TAL1 when it is expressed ectopically in a leukemic environment compared to cell with normal expression. While the exact reason for these differential binding specificities remain to be established, it is possible to envisage that those would be related to the leukemic role of TAL1.

Figure 24: The analysis of motifs under TAL1 peaks in 4 contexts. a.) the heights proportion of E-boxes found under TAL1 peaks. b.) the top ranked motifs for the differential comparison of Jurkat and Erythroid cell types. c.) the top motifs for the general analysis. The motifs where identified using Homer and a central region as the background. The motifs were annotated using STAMP

a



b

Leukemia with Erythroid Background

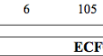
Motif	Rank	-log(pvalue)	Ann	E-Value
AACCAC	1	152	Runx1	2.12E-09
ACCACA	2	152	Runx1	1.27E-07
CACCTG	3	152	CC Ebox	2.39E-08
ACCGCT	4	152	Runx1	9.17E-05
TTTGAA	5	152	NR	1.81E-04
TKTGCG	7	138	Runx1	8.15E-06

Erythroid with Leukemia Background

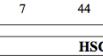
Motif	Rank	-log(pvalue)	Ann	E-Value
CTTATC	1	192	Gata	5.2E-08
TATCAG	2	192	Gata	1.1E-07
TATMTC	7	124	Gata	3.0E-06
GACASA	8	124	Smad	3.2E-06

c

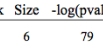
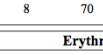
Leukemia with Background "Center"

Motif	Rank	Size	-log(pvalue)	Ann	E-Value	Logo
ACCACA	1	6	172	Runx1	1.265E-07	
CATCSG	2	6	172	Ets	1.75E-05	
CACCTG	12	6	105	CC Ebox	3.3127E-08	

ECFC with Background "Center"

Motif	Rank	Size	-log(pvalue)	Ann	E-Value	Logo
CGGAARC	1	7	44	Stat1	2.38E-06	
TGTTTTC	2	7	44	Fox	1.40E-05	
GATAACA	3	7	44	Gata	1.97E-10	

HSC with Background "Center"

Motif	Rank	Size	-log(pvalue)	Ann	E-Value	Logo
CCACAG	2	6	79	Runx1	1.72E-05	
GGGGGC	9	6	76	Erg	6.02E-07	
KGCGTG	12	6	71	CG Ebox	3.78E-06	
CTGGCTG	11	7	74	Smad	9.31E-06	
CSCYCCTC	21	8	70	PU.1	1.55E-05	

Erythroid with Background "Center"

Motif	Rank	Size	-log(pvalue)	Ann	E-Value	Logo
CTTATCT	1	7	240	Gata	4.13E-10	
CCAGCTG	6	7	163	GC Ebox	2.11E-11	
GCCTTGTC	3	8	221	P53	1.39E-04	
GGCTGTCA	10	8	171	Smad	6.67E-06	

5.4.2 Differential Leukemia and Erythroid De Novo search

The differential comparison of TAL1 binding between Jurkat (T-ALL) and Erythroid has been performed previously by palii et al [3]. To ensure that the PCA method had produced results that were reproducible, I performed a differential comparison between the erythroid and leukemic peak sets. The key motifs, which have previously been found to be over represented, were also over represented in the results of the PCA analysis. Principally, I observed the prevalence of RUNX1 variants and the preference of the CC Ebox in the leukemic condition, both of these motifs had been previously

found. I also found the over representation of GATA1 in the Erythroid condition. In addition to the motifs that were found previously, the differential analysis of the motifs also showed that the nuclear receptor (NR) motif was over represented in the Leukemic condition, and the SMAD motif was over represented in the Erythroid condition.

5.4.3 Generalized De Novo Search

While the differential analysis of the motif preference between the leukemic and erythroid conditions yielded expected results, the methodology was still limited to a binary analysis. The goal of the PCA analysis was to be able to perform general analysis of multiple data sets. To generalize the *de novo* search for DNA motifs that are overrepresented under TAL1 binding sites in each specific cell environments, I used the TAL1 ChIP-Seq peaks that showed no distinct difference between the cellular context as a background (i.e. Background Center, see Figure 22 a, left panel, the white square in the middle) to perform the search. Using the center as the background was useful in determining what binding locations were important in a context compared to the binding of TAL1 as a whole, rather than a single other context. As such, each motif could be shared between multiple contexts. The results can be seen in Figure 24 c. The peak set for the leukemic condition had over represented RUNX1 motifs, as well as ETS, which is consistent with the previously demonstrated importance of RUNX1 and ETS factors to target TAL1 to specific binding sites in leukemia [3]. Furthermore, the CC-Ebox was also identified, which is consistent with the results shown in Figure 24 a. For the endothelial cell type ECFC, the FOX and GATA motifs were enriched, suggesting that FOX and GATA families of transcription factors are involved in mediating the endothelial specific function of TAL1. In addition, Stat1, which is a TF important in cell signaling, was highly over represented.

The TAL1 peaks in the HSC condition contained a diversity of over represented motifs. The motifs that were of interest included RUNX1, ERG, PU.1, CG Ebox and Smad. RUNX1 is important for the formation of HSCs. However it has been shown that it is not required for their maintenance [69]. ERG was enriched in the HSC case

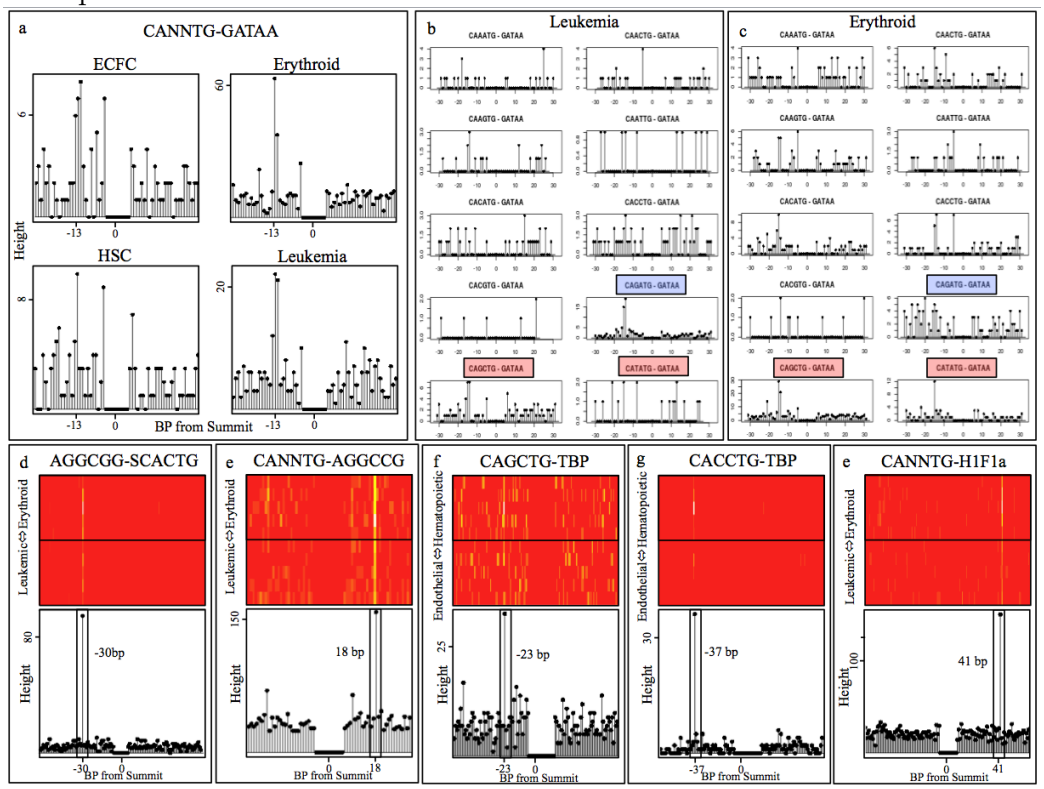
as well. It has been previously shown that ERG is required for self renewal in HSCs [70]. Interestingly, it has been shown that PU.1, as with ERG, may be required to maintain the HSC population within the bone marrow [71]. The CG Ebox was also over represented. This preference had been observed in the Ebox analysis, where HSC had a 68% increase in the proportion of peaks with CG Ebox compared to the other three contexts. However the absolute magnitude of this relation is still small, with only 2.2% of HSC peaks having a CG Ebox (in comparison to 23% with GA Eboxs). The Smad motif was also enriched in the HSC condition. Smad is an important TF in cell signaling and it has been shown that the role of Smad varies between cellular conditions and is tied to the closely to critical TFs [72] [73]. Smad was also enriched in the erythroid set of peaks, along with GATA, the GC Ebox, and p53.

5.4.4 Preferred Ebox GATA Motif Distance Under TAL1 Peaks

While the nucleotide combinations of motifs may be enriched in different conditions, so can the geometric relation between different proteins. These geometric relations manifest themselves in motif analysis in the form of preferred distances (i.e. a composite motif). As with the previously discovered over represented motifs there are also known preferred distances between motifs in the hematopoietic lineage. GATA and the CANNTG Ebox motifs (i.e. NN Ebox) have been shown, using ChIP-Seq data, to have a preferred distance under TAL1 peaks [3]. To ensure that this preferred distance was observable in peaks identified using the PCA method, I compared the distance between the GATA motif (GATAA) and each of the Ebox variant for each of the four contexts. GATAA was chosen, as it is the base for all three GATA variant motifs. The distance was measured from first base pair of the Ebox to the first base pair of the GATA motif. As expected, the preferred distance of 12 and 15 bp was found in each condition when comparing the NN Ebox. This result coincides with the previously published results specifying the distance between the motifs being 7 and 10 based on the number of base pairs between the end of the GATA motif and the start of the Ebox. Using the PCA method, it was possible to extend the analysis

of the preferred distance to determine the significance of this relation in each of the four contexts. The results, shown in Figure 25a, indicate that the preference, while apparent in each contexts, is greatest in the erythroid context.

Figure 25: Preferred distance between motifs. a.) shows the preferred distance between GATA and the NN Ebox for each of the four contexts. b.) shows the preferred distance between GATA and each of the Ebox variants for the leukemic and erythroid contexts. c-h.) show novel preferred distances that were identified under all TAL1 peaks in the AFS.



Following the confirmation that the Ebox GATA composite motif was shared between the Leukemic and Erythroid contexts (but was not distinctive in the other cell environments), the contribution of all Ebox variants to the composite motif were investigated for both conditions. The CC Ebox showed very low signal in both of the conditions. For the Erythroid context, the GC and TA Eboxes contributed over

75% of the signal that had been present in the general NN case. Conversely in the leukemic context, the contribution from these two combinations was marginal. In the leukemic context, I found that most of the signal was contributed by the GA Ebox (i.e. the in vitro preferred TAL1 binding Ebox). While contributing the majority to the leukemic NN signal, the GA Ebox showed no preferred distance signal to GATA in erythroid context.

5.4.5 GATA TAL1 Interactions in T-ALL and Erythroid

These differences in composite motif preference indicate possible binding mechanism differences between the Leukemic and Erythroid condition. TAL1 forms a complex with four other proteins. This pentameric protein complex includes LMO2, LDB1, a class II Eprotein (E2A, HEB), and GATA. TAL1 and the Eprotein first heterodimerize and are subsequently bridged to the GATA protein by LMO2 and LDB1, refer to Figure 26 d. Both the heterodimerized TAL1 and GATA possess a DNA binding domain. Peaks that have the composite Ebox GATA motif (see Figure 26 e), are likely locations where both proteins have bound to the DNA. To function however, both members with binding potential do not need to bind directly to the DNA. This has been shown in Erythroid cells. When the basic domain of TAL1 is altered such that it can no longer bind to the DNA, Erythroid cells can still function [13]. Thus, DNA binding can be driven by the overall potential of the complex, the preference of GATA, or the preference of TAL1.

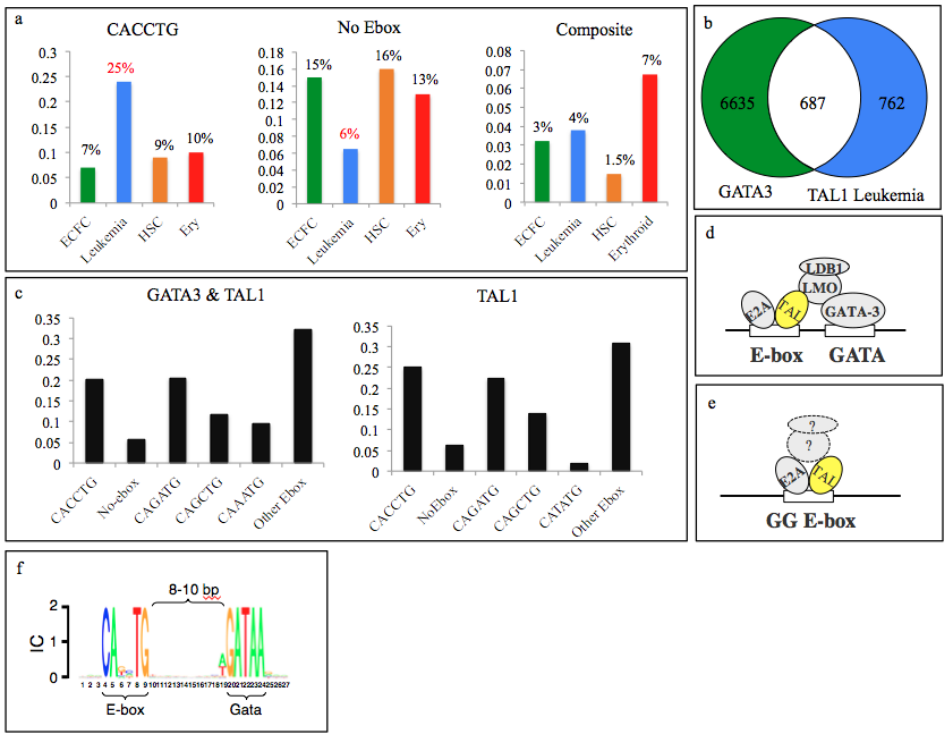
In Erythroid, the composite motif occurrence is greater than in the other three contexts (see Figure 26 a). Additionally, in the Erythroid context, TA and GC variants dominate the Ebox portion of the composite motif. Neither of these variants are the preferred in vitro binding target of the heterodimerized TAL1 Eprotein [5]. The prevalence of composite motifs with non-preferred Ebox variants suggests that GATA initiates the binding at these locations and subsequently recruits TAL1. Alternatively it could suggest that the pentamer has formed before binding to the DNA, and as a whole has a greater binding affinity.

In the Leukemic case the mechanism appears to be opposite. There is a smaller

proportion of composite motifs, and most are found with GA variants, the preferred in vitro binding target of heterodimerized TAL1. Conversely, the CC ebox variant has a very low rate of occurrence in the composite motif while consisting of almost 40% of the Eboxes found under Leukemic peaks. Both of these trends suggest that in T-ALL, TAL1 has the greater role of initiating DNA binding. It is not clear at this point whether the same pentameric protein complex was formed in T-ALL, following TAL1 binding at the CC Ebox.

By observing the colocalization of GATA3 and TAL1 ChIP-Seq peaks in T-ALL it was apparent that the colocalization ratio with the CC Ebox in overlapping peaks was no less than that what was observed in TAL1 alone, refer to Figure 26 c. This indicated that the same complex members group together in T-ALL peaks with a CC ebox as those with a composite motif. It was not clear whether the members formed the same complex or if a new complex binds to the CC Ebox in the leukemic context. In addition the transience of this relation is yet to be analyzed.

Figure 26: The TAL1 complex and its relation to GATA, a.) The proportion of peaks containing specific motifs for CC Ebox No Ebox and the composite TAL1 Gata motif (from left to right). b.) A Venn diagram showing the peaks unique to the GATA3 peak set, the TAL1 Leukemic set and the number shared between each . c.) the proportion of Ebox variants under Leukemic TAL1 peaks and the union of Leukemic TAL1 peaks and GATA3 peaks. d.) the pentameric complex e. The initiation of TAL1 binding to the CC Ebox. f.) The TAL1 GATA composite Motif.



5.4.6 Preferred Distances Between Other Motifs

Once I had confirmed that the preferred distance between GATA and the NN Ebox was present in the PCA analysis, I searched for additional preferred distances between other motifs. Figure 25 d comes from the analysis of all oligonucleotides where AGGCGG (a motif with no clear annotation) and SCCTG (a Pax5 motif) had a distinct distance of -30bp. The AGGCGG motif when compared with the NN Ebox had a distinct preferred distance of 18bp. Both of these preferred distances were found to

be slightly more prevalent in the erythroid than the leukemic contexts.

The TATA binding protein (TBP) tends to binds to gene promoters. Two Ebox variants had distinct preferred binding distance to the TBP. The CC Ebox is most likely to be 37 bp from the TBP, while the GC Ebox was found 23bp from the TBP. This suggests that there may be a context specific mechanism shared between the TAL1 complex and the TBP. Both of these relations are more prevalent in the differentiated hematopoietic condition than the endothelial.

HIF1a is a TF that is important for the response to hypoxia. Its gene has been shown to be over expressed in cancer [70]. Figure 25 e indicates that the NN Ebox and the H1F1a motif have a preferred distance of 41 bp. Furthermore, this distance is more preserved across both the erythroid and leukemic contexts. This suggests that in these cellular contexts, TAL1 may work with HIF1a to regulate some common target genes, a possibility that remains to be investigated.

5.5 Gene Ontology

In addition to motif analysis, gene ontology (GO) serves as a useful tool in elucidating functional characteristics of the genes that are associated to the specific TAL1 binding sites in each cellular context. Here I investigate the GO terms found for each of the four contexts. In addition to looking for set of discriminating peaks in each of the four contexts, I also investigated the genes associated to TAL1 peaks in the leukemic context that overlaps with CC Eboxes and NN Eboxes present at a preferred distance to GATA.

5.5.1 Gene Association

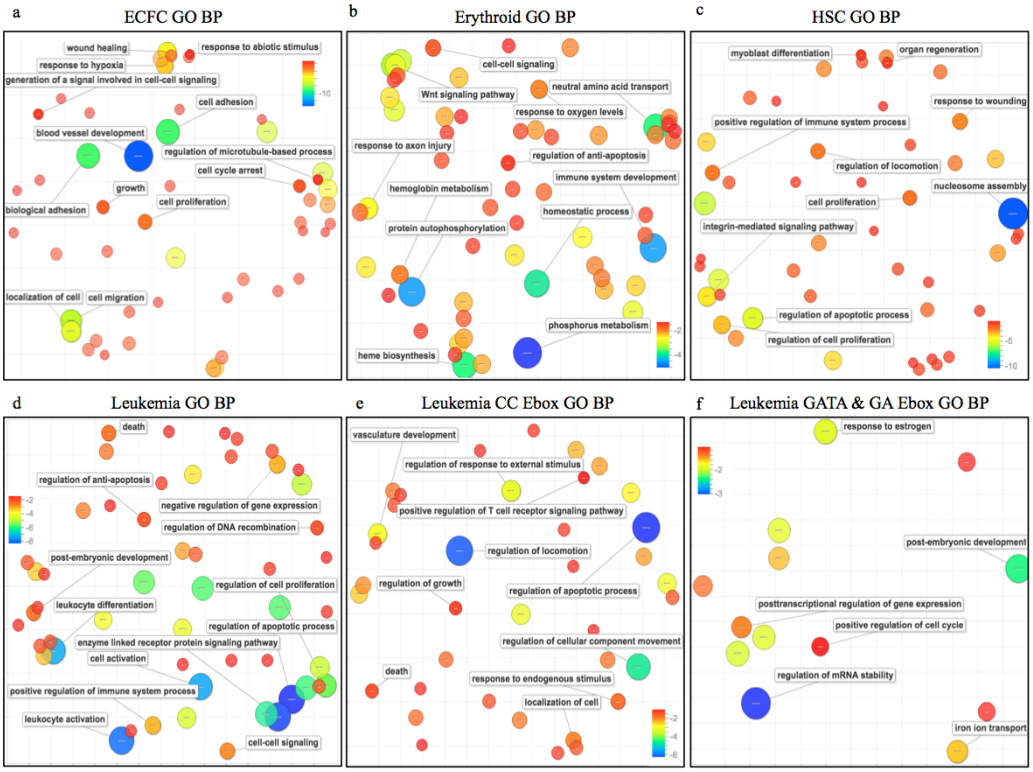
Before gene association took place I associated the peaks that were unique to each of the four contexts with genes in the RefSeq gene set. The GREAT algorithm with the standard configuration was used for the association [65]. The results of the association included 1687 genes associated with TAL1 in ECFC, 2635 in Erythroid, 1818 in HSC, and 2187 in T-ALL. While the peak sets were mutually exclusive the

gene sets were not. This was due to the possibility of multiple binding sites being associated with each gene. T-ALL and HSC had the greatest proportion of genes that were shared. This amounted to 9.28% of genes. On average only 8.17% of genes were shared pairwise between data sets.

5.5.2 Gene Categories Found for the Four Contexts

The over represented genes, in the gene set associated to ECFC peaks, can be seen in Figure 27a. The visuals were created using REVIGO [74]. The key fundamental biological processes in ECFC, including proliferation, migration, and adhesion were all present. In addition to the expected basic function of endothelial progenitors, genes important in blood vessel development and wound healing were also found. The erythroid biological process categories included several signaling pathways and genes relating the heme biosynthesis, a process specific to erythroid cells Figure 27b.

Figure 27: The results of gene ontology. The significance of each term is indicated by its size (largest to smallest) and colour. Blue are the most significant and red the least. Subplots a to d are the result for the four contexts, Subplots e and f specifically look at the role of the CC Ebox and the GATA Ebox composite motif in the leukemic context.



The GO terms found for the HSC include a diverse set of functionalities that would be expected from the precursor to the leukemic and myeloid condition. For example there is a category specific to the positive regulation of immune system processes. There is also a set of categories that are shared with ECFC. This includes terms such as wound healing.

The leukemic GO terms include those relating to leukocyte activation and differentiation. This is consistent with the previous experiments suggesting that TAL1 disrupts the process of T cell differentiation [3]. Additionally there is evidence of population misregulation through the prevalence of terms relating to apoptosis and

cell division. Again, this is consistent with the previous experiments showing that leukemic T-ALL cells undergo apoptosis upon TAL1 KD [3]. To determine if these terms related to ectopic TAL1 binding locations, normal TAL1 binding patterns and abnormal binding patterns were compared.

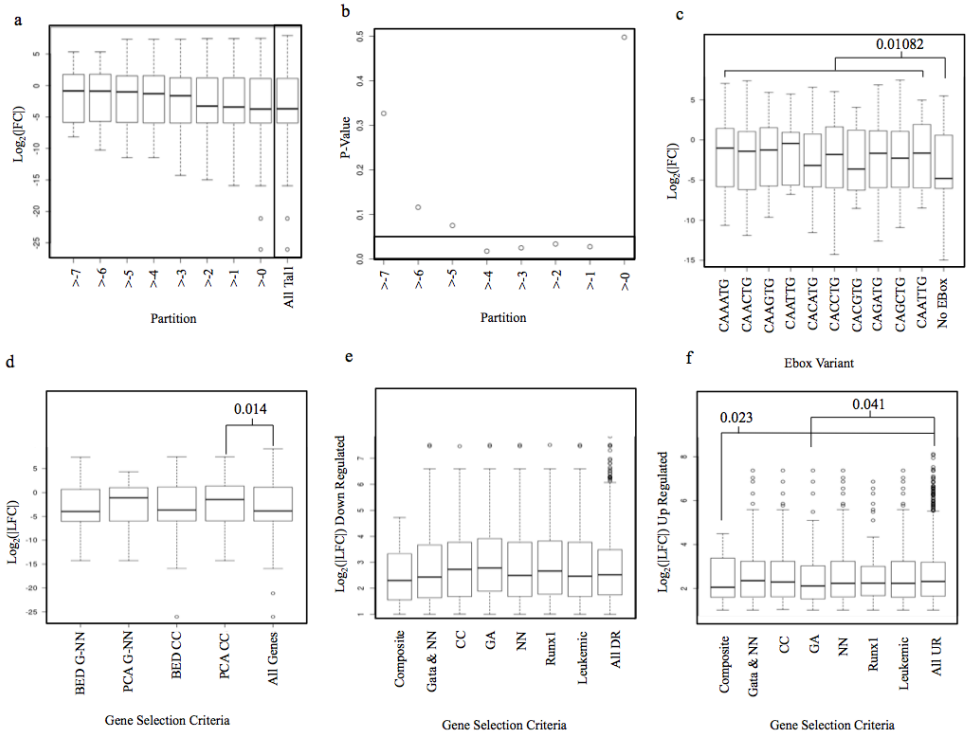
5.5.3 Leukemic Categories Contributed by Ebox Combinations

Figure 27 f and g depict the two Ebox combinations that were investigated for the leukemic context. The abnormal leukemic TAL1 peaks, i.e. those that had CC Eboxes, contained GO terms that related to apoptosis. The normally arranged TAL1 peaks, those with CAGATG and GATAA at a distance between 12 and 15bp , had no enriched terms relating to apoptosis. Both of the sets of peaks had genes representing embryonic processes however, suggesting that stem like properties may have been kept active through the ectopic binding locations of TAL1 present in T-ALL.

5.6 Differential Gene Expression

To elucidate the impact of TAL1 on the overall gene expression change in T-ALL Cells, I first assessed the magnitude of the gene expression LFC. The magnitude of the LFC was compared between the genes that were associated with different TAL1 peak sets. These sets were defined by their varying importance to T-ALL, i.e. how many standard deviations below the mean of the first PC the peaks fell. The magnitudes of the LFC at SDs ranging from -7 to -1 by 1 were compared with the expression of all genes associated to TAL1 peaks. The results can be seen in Figure 28 a and b. From figure 28, it is evident that between negative one and four SDs from the mean the Leukemic peaks have a statistically significant change in the magnitude of the LFC in gene expression, when compared with all genes associated to TAL1 peaks. In addition to peaks in the Leukemic context having a greater magnitude of LFC, the genes that were associated to peaks that had Eboxes were found to have a greater change than those that had no Ebox. This can be seen in Figure 28 c.

Figure 28: Leukemic gene expression with and without TAL1. a.) shows the log of the absolute log fold change of genes associated with Leukemic motifs isolated at varying levels of stringency, i.e. large positive values refer to large change positive or negative, and negative values refer to those genes whose values remain consistent between conditions. b shows the resulting p-value from the t test comparing all genes associated with Leukemic peaks to those associated to all TAL1 peaks. c shows the expression for genes associated to peaks which have a specific Ebox. d compares the absolute expression of the CC Ebox (CC) and the Gata Ebox composite (G-NN) genes. e and f compare CC and G-NN genes that are strictly down regulated and up regulated respectively



After confirming that there is a greater change in the magnitude of gene expression LFC in Leukemic peaks and that genes associated to peaks that had Eboxes had a greater change than those that did not, I compared the magnitude of the LFC between

TAL1 peaks that had motifs with relevance to T-ALL. I compared the genes associated to peaks that had a GATA Ebox composite motif (G-NN) and the genes associated to peaks that had a CC Ebox motif (CC) to the magnitude of the LFC of all genes. Figure 28 d shows the variation in expression level difference for the G-NN and CC gene sets. I did this using both the overlap method (BED) and the PCA approach. For the overlap method, the genes were associated with peaks that were found in the union of the twelve leukemic and not in the ten non leukemic data sets. The change of expression for down regulated G-NN genes was not found to be statistically greater than the overall change in expression of genes for either the PCA or overlap approach. The genes associated to peaks with CC Eboxes conversely, were found to have statistically greater gene expression difference than the body of genes, but only for those peaks identified using the PCA method. Both the genes associated to the CC and G-NN the PCA peaks had larger mean gene expression difference compared with those peaks identified using the traditional approach. Thus, the PCA method selects for peaks that have a greater impact on gene expression and the TAL1 peaks with a CC Ebox have a statistically greater influence on gene expression.

Since TAL1 can act both as an activator and a repressor these changes could either come from an increase or a decrease in regulation. In Figure 28 e and f I show the fold changes of the up and down regulated leukemic genes. For the genes that are down regulated upon TAL1 KD (i.e. genes that are activated by TAL1), those whose peak had a GA Ebox were found to have statistically higher difference in expression between conditions than the group of all down regulated genes. The G-NN category had statistically less impact on gene expression for down regulated genes. For the up regulated genes, the difference between the CC and all up regulated genes was negligible, however, the up regulation of G-NN was less than that of all genes. Less up regulation in the G-NN set of genes indicates that TAL1 has less impact on gene expression when binding to a GATA Ebox composite motif in the Leukemic context. When binding ectopically to a GA or CC Ebox in the Leukemic context the magnitude of the gene changes up and down are large. This indicates that in the case of binding to a CC Ebox TAL1 appears to behave as both an activator and repressor.

5.7 Conclusion

I have shown that the PCA method is suitable for isolating hematopoietic and endothelial TAL1 peaks into at least four cellular contexts with a minimum normalized height for each condition. Additionally I have shown that the isolated peaks represent a quantitative measure of importance for each condition, i.e. peaks may be shared between contexts but will be significant in only one. Using the isolated peaks, I confirmed that the CC Ebox preference was seen only in T-ALL and reproduced the comparison between Leukemic and Erythroid peaks showing that this approach can reproduce results that have been previously published. While these results were useful for validating my approach they did not provide the novel benefit of my approach, the ability to generalize the comparison between multiple cellular contexts.

Using the novel method of comparing the outlying peaks against the central peaks I have identified the motifs that are overrepresented in T-ALL. In addition I have identified motifs present in the HSC that had direct biological relevance, such as ERG, and PU.1. I also found some novel motifs, primarily that the P53 motif is overrepresented in the Erythroid context, and STAT1 is overrepresented in the ECFC.

Apart from the motif *de novo* search, I have also shown that the GATA Ebox composite motif is present in both the Leukemic and Erythroid conditions. However, there are important differences within this composite motif between the two contexts. In the erythroid context, the TA and GC Ebox variants contribute 80 % of the peaks that have a composite motif. In contrast, GA provides the majority of the peaks in the leukemic context. This suggests that the composition of the pentameric complex may be slightly different between contexts. I have also identified a handful of other novel preferred distances, such as the distance between a PAX5 motif and the NN ebox and the unidentified AGGCGG motif. A preferred distance was also identified between the Ebox and HIF1a and two distinct distances were found between the TATA binding protein and two Ebox variants.

GO analysis was used to complement the motif denovo search. GO revealed the expected results for each of the contexts. For the ECFC all of the characteristic

proliferation, migration and adhesion terms were found. For the Erythroid the JAK-SAT cascade was present along with heme biosynthesis. The HSC had mixed bag of terms found in the other contexts. The largest terms found in the Leukemic condition included leukemic activation and regulation of apoptotic processes. When the subset of Leukemic peaks that had CC Eboxes present was compared to those with the NN-G relationship it was determined that the regulation of apoptotic processes was impacted mostly by TAL1 binding the CC Ebox.

Using the expression data of the Jurkat cell line wild type and TAL1 KD I have confirmed that the peaks isolated using the PCA method have direct biological relevance and that the PCA method selects for peaks that have a greater impact on gene expression. Additionally, I have shown in Jurkat that the TAL1 peaks that have Eboxes present have a greater impact on the expression of local genes. I also suggest that the TAL1 peaks with CC Eboxes present have a greater impact on up and down regulation when compared to the TAL1 peaks that occur in a normal arrangement.

Chapter 6

Discussion

6.1 Advance in Comparison of ChIP-Seq Data

The methodology I have developed and implemented extends work that has been published previously by Stark and Hardison [6, 10]. The base of my analysis comes from a modification of global analysis that had been performed by Stark et. al. [6]. The modification reduces the impact of the background signal by limiting global analysis to a subset of peaks that are shared between the two data sets that are being compared. The fundamental step proposed was to create an AFS from the independently identified peaks. The AFS represents a subset of the genome that is most likely to have binding. From the AFS, Stark limited the analysis to the quantitative binary comparison of data sets.

The analysis of the AFS was extended by Hardison et al, when investigating the role of GATA in differentiated hematopoietic cells [10]. The comparison was extended to multiple data sets through the application of hierarchical clustering and k-means clustering. This was shown to work when generalizing a few contexts (6 data sets in 4 contexts) and identifying the peaks that were shared between contexts. However, through the application of k-means clustering, the quantitative nature of the analysis was lost.

My method is intended to not only allow the comparison of multiple ChIP-Seq data sets, but also to retain quantifiable information pertaining to each of the peaks.

By using the PCs of the UDM as separation weightings, I have retained a quantifiable interval value for each peak. The benefit of having quantifiable weightings that separate conditions is two fold. First, it allows the data sets to be clustered based on the information present in the ChIP-Seq data. Thus, predefined contexts, such as those outlined in a differentiation tree, are not needed. This decreases the inherent bias when comparing data sets since the difference must be present in the underlying data in order for a differential comparison to be made. Second, the PCs provide a spectrum of separation. Thus, each peak will have a value along each axis that separates contexts of interest. This separation may have an immediate biological implication or may cross the boundaries of expected contexts. Due to this property, a peak can be observed to be dominant in T-ALL but also have a minor presence in stem cells. Due to this flexibility, the separation threshold can be modified to ensure stability of the analysis or to alter the stringency of the comparison. By using the eigenvectors of the UDM (the PCs found using the PCA), I have retained the interval information present in the normalized heights of the peaks in each context, which allows for inter peak distance measures, peak ranking, and isolation of peak sets.

6.2 Comparison to Alternative Approaches

There have been approaches to comparing ChIP-Seq data sets that are independent of the work done by Stark and Hardison. While some of these approaches provide foundations for comparing data sets, their functionality decreases as the number of ChIP-Seq data sets increases.

MULTOVL is a method of selecting conserved peaks in replicates [48]. By requiring that peaks come from two or more replicates, the likelihood of those peaks representing sustained binding increases. This method can be extended to differential comparison between conditions. In this thesis, the approach used by differential MULTOVAL analysis was referred to as overlap analysis (as the more general term). The benefits of using this approach include the reduction of background signal influence on the analysis, and the flexibility of requiring varying degrees of overlap. By increasing the number of replicates that must overlap between conditions the method can

become more conservative, which helps to ensure the stability and reproducibility of the analysis. Unfortunately this narrows the spectrum of comparison to the locations with the strongest binding while disregarding the binding signal in the final analysis. Due to these limitations, MULT0VL can only provide a qualitative comparison of peak binding and disregards marginal binding signal.

DBCHIP is another method of data set comparison that was investigated. DBCHIP is a quantitative approach for the identification of differential peaks between two or more conditions [75]. Similar to Stark, they identify a unified set of consensus peaks from two or more data sets. Rather than comparing the normalized heights directly, they estimate the probability of the peak being differential between the two data sets. This approach, which unlike Starks method does not rely on quantile normalization, has several benefits. First, since the global binding signal is not equalized between data sets, global changes in binding preference can be observed. Second, since the analysis takes into account the control data set, data sets that have very low read counts and low signal to noise ratios will not end up with an inflated number of significant peaks. The principle limitation of this approach is that it only applies to differential binding between predefined contexts, i.e. what is being compared must be known before the comparison is made. As such, while it can be applied to a large number of contexts, it will only identify differences between those regions that are predefined as different. This provides no direct information about which data sets belong to different contexts, and does not provide a spectrum that separates conditions.

Quantile normalization was a critical step in my analysis, since the data sets being compared ranged two orders of magnitude in size. Unfortunately, quantile normalization has its limitations. Critically, it removes the overall difference in global binding strength between data sets. MANorm was investigated as a robust alternative to quantile normalization.

MANorm is a normalization method that allows for quantitative comparison between biological data sets [76]. It determines the mean and difference of peaks that are shared between data sets and determines the rate and intersect that normalizes the shared regions. The parameters used to normalize the shared regions are then

applied to the regions that are unique to each data set. For the purpose of general comparison between an unspecified number of data sets MANorm fails to scale. This is due to its reliance on defining shared peaks between data sets. As the number of data sets contributed from different conditions increases, the proportion of peaks that are shared between all conditions decreases. Additionally, as the number of disparate data sets increase, those that are shared between all conditions become more biased towards sampling for traits that are universally conserved. The proposed alternative to requiring all data sets to overlap was to select peaks that had a set number of overlaps between conditions, i.e. defining shared peaks as those regions where at least N data sets overlap. When applied as an alternative to quantile normalization in our pipeline, however, this approach was found to be unviable since the majority of the peaks that contained multiple overlaps were context specific, i.e. all of the overlapping peaks came from a single context such as T-ALL or erythroid. This heavily biases the normalization of the full set of data. Alternatively, if the peaks were selected based on those shared between two or more contexts, the analysis would no longer be a general approach to comparison since it would require the contexts to be defined before the analysis.

6.3 Limitations of the PCA Method

I designed the PCA method to achieve two goals. Grouping ChIP-Seq data sets based on underlying similarity and elucidating differences between those groups. The scope of questions that can be answered pertaining to elucidating differences are limited by the groupings that are inherent in the data. Thus, arbitrary comparisons cannot be defined before the analysis. There are several benefits inherent in imposing this limitation. When analyzing a large number of data sets, it provides a guide by indicating which data sets should be compared. This makes the analysis less reliant on previously published topics pertaining to the relationship between data, and simplifies the analysis of a large number of data sets. This reduced reliance on expected results also decreases the inherent bias brought to the analysis by the researchers presumptions. Additionally, by limiting the set of questions that can be

answered to those that truly have differential binding, my method makes it easier to elucidate differences that may not be anticipated from literature.

Apart from the limited scope of biological questions that can be answered using the approach, there are also issues that come from using quantile normalization. Since the analysis relies on quantile normalization, any real global difference may be hidden in the differences in data set sequencing depth. MANorm was investigated as an alternative to quantile normalization, however as mentioned previously, it was found to be unable to scale due to its reliance on the specification of a shared region. Applying PCA partially compensates for squashing global differences in binding strength. This stems from the relationship between the PC weightings and the conservation of high signal strength between data sets in a context. In other words, even if the absolute binding strength is greatest in a data set that falls in the second context, if the binding strength is consistently high in the first, it is a strong indicator that that peak is important for the first context. This relation helps to elucidate marginal binding in contexts where the signal may be weaker, due to differences in global binding or sequencing depth.

6.4 Summary

My method extends work done by Stark and Hardison by providing a general approach to the quantitative comparison of ChIP-Seq binding between disparate data sets. While other approaches are available for differential comparison of ChIP-Seq binding signal, they lack the quantitative, unbiased, general applicability of my approach to an unspecified number of data sets. Ultimately my analysis works best when applied with the aim of discovering underlying differences in a TF function between several cellular contexts.

Chapter 7

Conclusion and Recomendations

7.1 Technical Aim

Throughout this thesis I have presented my method for the analysis of multiple ChIP-Seq data sets. The technical aim of my analysis was to provide an unbiased general approach to analyzing multiple ChIP-Seq data sets that could be applied to any number of data sets and would provide a quantitative comparison between them. My biological aim was to determine the extent to which the regions found by my method are biologically relevant and to elucidate the differential role(s) of TAL1 within the hematopoietic and endothelial lineages.

Using a subset of the TAL1 hematopoietic ChIP-Seq data, I first showed that the PCs of the quantile normalized UDM can be used to distinguish between data sets that come from Erythroid and Leukemic cell lines using a single PC. This indicated that the PCs, as quantifiable measures, were sufficient for the isolation of conditions. From there, I proceeded to show that my analysis could be applied to many ChIP-Seq data sets from several labs and cellular conditions, and that the use of the PCA approach was unbiased.

To accomplish this I extended my analysis to all 22 TAL1 hematopoietic and endothelial data sets. Using hierarchical clustering, I showed that the differentiation tree can be recreated using only TAL1 binding, indicating that using the AFS provided an analysis mechanism that could scale as the number of disparate data sets increased.

In addition to using hierarchical clustering, the results of the PCA provided clearly separable contexts. I have shown that these separable vectors were not biased by technical factors, including data set size and the number of peaks identified in each data set. Additionally since these contexts were inherent in the data, the analysis was not biased based on the predefinition of the groups of interest by the researcher. The final results from the PCs of the quantile normalized UDM was the isolation of four contexts: Erythroid, Leukemic, HSC, and ECFC. The separation of these contexts was stable when the data sets were perturbed, and when different control schemes were used.

The combination of defining an AFS and applying PCA to the quantile normalized read pile up count under each region provided a general method for analyzing several ChIP-Seq data sets. The methodology provided a quantifiable weighting for each peak in the form of the eigenvector of the UDM. Finally, the approach reduced both the technical bias and the explicit experimental bias. While the use of quantile normalization and the PCA each had technical limitations individually, their combination provided a powerful tool for elucidating the peaks of greatest importance for each condition in a quantifiable manner.

In this thesis the method was applied to the TF TAL1 in the hematopoietic and endothelial lineages. However, this method is not limited to the comparison of TAL1. This method can used to compare the ChIP-Seq binding patterns of any TF that is shared between cellular conditions. This method can also be extended to compare different TFs within the same cellular condition.

7.2 Biological Aim

The role of TAL1 in the hematopoietic and endothelial lineages was the ideal test case for the application of my analysis. The fact that the role of TAL1 varies between conditions is known. However, the specific biological function(s) of TAL1 in each of these contexts, particularly in the leukemic context where TAL1 is acting as an oncogene, has yet to be elucidated. From the 22 hematopoietic and endothelial data sets that were being assessed, four contexts were isolated for downstream functional

analysis. The parameters used in the final separation of contexts included using a combined control, selecting peaks identified by MACS that had a p-value of at least 10^{-20}, and taking peaks that fell at least 2 SDs from the mean of the PC that separated them and within 2 SD of all other PCs that were used for separation.

Through motif analysis, GO, and analysis of RNA-Seq, I have indicated that the peaks separated using the PCA method could be used to reproduce previously published results and provided novel information pertaining to the role of TAL1 in different contexts. The comparison of motifs between the leukemic and erythroid condition was used to confirm that the peaks identified pertain to reproducible biologically relevant functionality. The key motifs found in the differential comparison, RUNX1 and CC Ebox for T-ALL and GATA1 for Erythroid had been identified in previous experiments comparing only two data sets. Additionally, I made the surprising finding that the CC Ebox is over represented in the leukemic context.

The novel aspect of the motif denovo search involved using the central region of the three separating PCs as the background. Using this background I was able to determine that the Smad motif is over represented in both Erythroid and HSC, and that P53, Stat1, and PU.1 are over represented in Erythroid, ECFC and HSC respectively. In addition to the over represented motifs, I determined that the Ebox and GATA composite motif, while conserved between T-ALL and Erythroid, prefers different ebox variants: CA**GC**TG and CA**TA**RG in Erythroid, and CA**GA**TG in T-ALL.

The assessment of the separated peaks from the genetic perspective, using GO and RNA-Seq data showed that the peaks separated using PCA contained biologically relevant information and that the PCA method selects for peaks that have a greater impact on gene expression. This change in expression was greatest in peaks that contained E-boxes. Applying GO to the leukemic peaks that had CC Eboxes and the GATA Ebox composite motif revealed that the genes related to apoptosis occur most often in the ectopic TAL1 peaks that bind to a CC Ebox. In Jurkat, the expression of genes associated to these peaks also had greater change (both activating and repressing) compared to the peaks with GATA Ebox composite motif, when associated to genes.

In conclusion, the approach of unifying the data set into a single set of consensus peaks and subsequently separating them based on the PCA of the quantile normalized read pile up count yielded biologically relevant subsets of the hematopoietic and endothelial lineages. This separation was robust in the context of changing contributing data sets and was present regardless of the control mechanism used. When compared differentially, the peaks that were isolated yielded similar functional characteristics to analyses that have been previously published. The general comparison of peaks yielded novel binding motifs present under the TAL1 peaks as well as relevant biological process following GO.

7.3 Future Directions

The analysis I have presented meets my aims of comparing ChIP-Seq data sets while retaining quantifiability, the ability to scale, and the reduction of bias. It was successful in isolating four stable contexts that were found to have biological relevance. There are still several possible advancements that could be made on my methodology, and the biological analysis of TAL1 in the future.

A key advancement would be to replace the quantile normalized heights, as the input to the PCA, with another measure of significance. For the application of my analysis, on the role of TAL1, there were control data sets available for each data set being compared. My methodology, however, did not take the read count pile up information from these controls into account. If a probability of binding, which related to the pile up around the summit in the treatment and control, were used in place of the normalized height then global changes in binding potential could be compared.

Another development would be to redefine the process of merging peaks while defining the AFS, to reduce the possibility of long chained peaks. If the peaks from individual data sets were to be clustered based on their distances from one another, rather than combining them and defining a new summit, a higher fidelity comparison may be possible.

Finally, the analysis so far has been applied only to TAL1 in the hematopoietic

and endothelial domains. To test it further, the inclusion of non TAL1 regions in the analysis could be done to quantify the reduction of background signal through using the AFS. Apart from this, including non TAL1 data sets in the analysis could help to elucidate which TAL1 regions are conserved across all contexts. This is a necessary step if the analysis is to be extended from differential comparison of data sets to identification of similarities between all data sets of interest.

In this thesis I compared the gene expression change in Jurkat for a TAL1 knock out. Gene expression changes could be compared between other cell types that were included in the analysis to validate the effectiveness of isolating key binding locations. In addition the expression level could be correlated with the key principle components, since it is another technical factor which may influence the clustering of cell types.

There are several experiments that could be performed to extend the analysis, and to confirm my biological findings. ChIP-Seq of GATA2 in T-ALL and GATA1 in Erythroid could be used to determine if there is a unique composite option for TAL1 in the leukemic condition when it binds to a CC Ebox. Additionally, in order to determine if the motifs found to be over represented in specific contexts merit further investigation it should be determined if they interact with TAL1 or the members of its complex.

In addition to elucidating the role of TAL1 in the hematopoietic and endothelial lineages my method can be extended to changes in the binding patterns of other proteins. As such the method could be applied to a variety of proteins for further validation and to elucidate novel relations between TF binding patterns and cell fate decisions.

Appendix A

Data Sources

A.1 Raw data sets

Tables 3, 5, and 4 summarize the data sets used in this analysis. The data sets have all been processed between 2010 and 2014. Data sets are included that use both a mock and an input. The data sets came from 7 published papers and 3 unpublished sources, and are all, save Brand Leukemia Patient 1, available on GEO.

A.1.1 Sanda 2012 Cancer Cell

The 2012 paper from the Young lab at the Whitehead Institute at MIT investigated the role of the TAL1 complex in the regulation of T-ALL. They investigated its colocalization with other key members of the pentameric complex including the e-protein, GATA3 a non pentameric transcription factor RUNX1. Sanda was focused on understanding the functionality of TAL1 within T-ALL and did not compare its binding outside of that context.

A.1.2 Palii 2011 EMBO

For the 2011 EMBO publication Palii et al investigated the differential role of TAL1. Specifically, the differences in the binding preference of TAL1 between the jurkat cell

line and erythroid. This investigation paved the way for my general analysis of TAL1 between members of the hematopoietic and endothelial lineages.

A.1.3 Tijssen 2011 Dev Cell

In their 2011 cell dev paper Tijssen et all compared the binding sites of 5 key transcription factors, GATA1, GATA2, RUNX1, FlI1 and TAL1 in megakaryocytes. They found that the co-localization of all five proteins was most enriched near known hematopoietic regulators and found several novel genes that may be regulated by the potential interaction of the 5 proteins.

A.1.4 Novershtern 2011 Cell

Novershtern et al were investigating cis-regulatory circuits and the role of critical hematopeoitic regulators.

A.1.5 Hu 2011 Genome Research

Hu et al were focused on identifying enhancers, specifically enhancers important in the differentiation of HSCs to erythrocytes.

A.1.6 Mansour 2014 Science

Mansour was focused on discovering super enhancer locations on the genome. To do this they investigated the binding locations of RUNX1, GATA3 and TAL1 conserved across HSC MOLT and Jurkat cell types.

A.1.7 Unpublished

In addition to the 7 publications there were 3 sources of unpublished data available. These data sets came from Brand, Young and Snyder. The data from Young and Snyder are available on GEO.

Table 3: Summary of the data sets and their published sources

Data Set	PMID	Lab	Author	Control	Journal	Year
Prima5(brand)	N/A	Brand	N/A	Mock	N/A	N/A
Prima2	3422504	Young	Sanda	Input	Cancer Cell	2012
Prima2(rep)	3422504	Young	Sanda	Input	Cancer Cell	2012
Prima5	3422504	Young	Sanda	Input	Cancer Cell	2012
Prima5(rep)	3422504	Young	Sanda	Input	Cancer Cell	2012
Jurkat(rep)	3422504	Young	Sanda	Input	Cancer Cell	2012
Jurkat	3422504	Young	Sanda	Input	Cancer Cell	2012
Jurkat(brand)	3034015	Brand	Palii	Mock	Embo	2011
RPMI	N/A	Young	N/A	Input	N/A	N/A
RPMI(rep)	N/A	Young	N/A	Input	N/A	N/A
CEM	22897851	Young	Sanda	Input	Cancer Cell	2012
CEM(rep)	22897851	Young	Sanda	Input	Cancer Cell	2012
ECFC	24792117	Brand	Palii	Mock	Cell Stem Cell	2014
MEKA	21571218	Gottgens	Tijssen	Mock	Dev Cell	2011
CD133	21241896	Young	Lawton	Input	Cell	2011
CD34	21795385	Zhao	Hu	Input	Genome Res	2011
CD34(rep)	25394790	Young	Mansour	Input	Science	2014
Erythoid	3034015	Brand	Palii	Mock	EMBO	2011
Erythroid(fetal)	23041383	Orkin	Xu	Input	Dev Cell	2012
Erythroid(adult)	23041383	Orkin	Xu	Input	Dev Cell	2012
K562	Encode	Snyder	N/A	Input	N/A	N/A
K562(rep)	Encode	Snyder	N/A	Input	N/A	N/A

Table 4: Summary of GEO information for each data set

Data Set	GSE	SRR	Antibody
Prima5(brand)	N/A	N/A	Santa Cruz SC-12984
Prima2	GSE29180	SRR372690	Santa Cruz SC-22206
Prima2(rep)	GSE29180	SRR372692	Santa Cruz SC-22206
Prima5	GSE29180	SRR372691	Santa Cruz SC-22206
Prima5(rep)	GSE29180	SRR372693	Santa Cruz SC-22206
Jurkat(rep)	GSE29180	SRR443847	Santa Cruz SC-22206
Jurkat	GSE29180	SRR443848	Santa Cruz SC-22206
Jurkat(brand)	GSE25000	SRR070589	Santa Cruz SC-12984
RPMI	GSE39179	SRR519118	Santa Cruz SC-12984
RPMI(rep)	GSE39179	SRR519120	Santa Cruz SC-12984
CEM	GSE33850	SRR372688	Santa Cruz C-21
CEM(rep)	GSE33850	SRR372689	Santa Cruz C-21
ECFC	GSE53423	SRR1051799	Santa Cruz C-21
MEKA	GSE24674	SRR070379	Santa Cruz SC-12984
CD133	GSE26014	SRR094806	Santa Cruz SC-12984
CD34	GSE26501	SRR189205	Santa Cruz SC-12984
CD34(rep)	GSE59657	SRR1522112	Santa Cruz SC-12984
Erythoid	GSE25000	SRR070589	Santa Cruz SC-12984
Erythroid(fetal)	GSE36985	SRR452947	Santa Cruz SC-12984
Erythroid(adult)	GSE36985	SRR452948	Santa Cruz SC-12984
K562	GSE31477	SRR502380	Santa Cruz SC-12984
K562(rep)	GSE31477	SRR502381	Santa Cruz SC-12984

Table 5: Summary of the naming conventions used throughout the thesis

Data Set	Name	Short Hand
Prima5(brand)	Patient 2 B	P2B
Prima2	Patient 1	P1
Prima2(rep)	Patient 1 (rep)	P1-R
Prima5	Patient 2 A	P2A
Prima5(rep)	Patient 2 A rep	P2A/R
Jurkat(rep)	Jurkat (sandar)	JurkS
Jurkat	Jurkat(sandar rep)	JurkS/R
Jurkat(brand)	Jurkat (brand)	JurkB
RPMI	RPMI	RPMI
RPMI(rep)	RPMI (rep)	RPMI/R
CEM	CEM/C1	CEM
CEM(rep)	CEM/C1 (rep)	CEM
ECFC	ECFC	ECFC
MEKA	Megakaryocyte	MEKA
CD133	CD133/HSC	CD133
CD34	CD34/HSC A	CD34A
CD34(rep)	CD34/HSC B	CD34B
Erythoid	Proerythroblast (Brand)	proebAA
Erythroid(fetal)	Proerythroblast Featal	proebF
Erythroid(adult)	Proerythroblast Adult	proebAA
K562	K562	K562
K562(rep)	K562 (rep)	K562/R

A.2 Peaks Called

Table 6 summarizes the peaks that were identified using MACS with a standard cut off and three control cases, a combined control, each data sets individual control, and no control.

Table 6: Summary of Reads alligned uniquely to hg19 and the reads identified using macs

Data Set	Reads	Peaks No	Peaks Combined	Peaks Single
Prima5(brand)	64400090	41654	11765	18989
Prima2	400006	238	4839	3444
Prima2(rep)	446551	374	6053	4366
Prima5	481923	3919	6910	2987
Prima5(rep)	509859	4224	7655	3404
Jurkat(rep)	5063810	1123	992	615
Jurkat	5529172	5471	5027	3302
Jurkat(brand)	5831704	2937	2644	1716
RPMI	4842091	37477	39253	7352
RPMI(rep)	1690681	25392	29427	9095
CEM	1183173	2500	2935	2824
CEM(rep)	3181126	3873	3725	3643
ECFC	32407916	74614	58254	66438
MEKA	10742953	3841	5042	2040
CD133	6088943	324	109	15
CD34	3791775	1764	1795	628
CD34(rep)	17405262	1542	490	1590
Erythoid	6479544	6466	6132	5015
Erythroid(fetal)	7630154	4859	5318	5371
Erythroid(adult)	2274458	2618	2175	978
K562	22936696	38463	40057	22106
K562(rep)	15343051	38008	39133	21733

A.3 Correlation Between Data Sets

Table 7 and Table 8 show the correlations of all 22 data set heights with one another when defined with a p-value of 5 and 20 respectively.

Table 7: Correlation of TAL1 AFS Peak Pile Up Counts at a Cut Off of 5

	tall_p1	tall_p2_1	tall_p2_2	tall_p3_1	tall_p3_2	jurk_sandar_1	jurk_sandar	jurk	rpmi_1	rpmi_2	cem_1	cem_2	ecfc	meka	cd133	cd34	cd34_new	eryt	eryt_f	eryt_a	k562_1	k562_2
tall_p1	1	0.34	0.36	0.52	0.52	0.43	0.65	0.56	0.60	0.60	0.54	0.61	0.07	0.31	0.28	0.40	0.34	0.13	0.14	0.13	0.08	0.09
tall_p2_1	0.34	1.00	0.98	0.64	0.65	0.58	0.45	0.52	0.24	0.26	0.62	0.55	0.04	0.44	0.59	0.41	0.52	0.24	0.36	0.27	0.08	0.08
tall_p2_2	0.36	0.98	1.00	0.66	0.67	0.58	0.47	0.53	0.25	0.27	0.64	0.57	0.04	0.45	0.59	0.41	0.52	0.24	0.36	0.27	0.08	0.08
tall_p3_1	0.52	0.64	0.66	1.00	0.98	0.59	0.56	0.60	0.36	0.38	0.68	0.64	0.05	0.47	0.56	0.47	0.52	0.25	0.34	0.26	0.10	0.10
tall_p3_2	0.52	0.65	0.67	0.98	1.00	0.59	0.56	0.60	0.35	0.38	0.68	0.64	0.05	0.47	0.57	0.48	0.54	0.24	0.35	0.27	0.09	0.10
jurk_sandar_1	0.43	0.58	0.58	0.59	0.59	1.00	0.72	0.72	0.39	0.36	0.65	0.64	0.08	0.44	0.71	0.54	0.66	0.24	0.46	0.30	0.13	0.12
jurk_sandar	0.65	0.45	0.47	0.56	0.56	0.72	1.00	0.81	0.59	0.59	0.71	0.77	0.07	0.37	0.47	0.42	0.41	0.21	0.30	0.24	0.14	0.14
jurk	0.56	0.52	0.53	0.60	0.60	0.72	0.81	1.00	0.52	0.53	0.72	0.76	0.08	0.56	0.60	0.51	0.55	0.32	0.40	0.31	0.17	0.17
rpmi_1	0.60	0.24	0.25	0.36	0.35	0.39	0.59	0.52	1.00	0.97	0.49	0.57	0.13	0.18	0.20	0.34	0.20	0.12	0.14	0.16	0.12	0.13
rpmi_2	0.60	0.26	0.27	0.38	0.38	0.36	0.59	0.53	0.97	1.00	0.51	0.58	0.13	0.20	0.20	0.34	0.19	0.13	0.14	0.17	0.14	0.15
cem_1	0.54	0.62	0.64	0.68	0.68	0.65	0.71	0.72	0.49	0.51	1.00	0.87	0.08	0.48	0.56	0.44	0.47	0.28	0.37	0.30	0.15	0.15
cem_2	0.61	0.55	0.57	0.64	0.64	0.64	0.77	0.76	0.57	0.58	0.87	1.00	0.09	0.46	0.51	0.44	0.44	0.27	0.35	0.29	0.17	0.17
ecfc.tsa	0.07	0.04	0.04	0.05	0.05	0.08	0.07	0.08	0.13	0.13	0.08	0.09	1.00	0.14	0.14	0.19	0.14	0.05	0.09	0.06	0.04	0.05
meka	0.31	0.44	0.45	0.47	0.47	0.44	0.37	0.56	0.18	0.20	0.48	0.46	0.14	1.00	0.64	0.56	0.63	0.49	0.50	0.42	0.24	0.26
cd133	0.28	0.59	0.59	0.56	0.57	0.71	0.47	0.60	0.20	0.20	0.56	0.51	0.14	0.64	1.00	0.61	0.78	0.35	0.56	0.38	0.17	0.16
cd34	0.40	0.41	0.41	0.47	0.48	0.54	0.42	0.51	0.34	0.34	0.44	0.44	0.19	0.56	0.61	1.00	0.68	0.27	0.41	0.31	0.19	0.19
cd34_new	0.34	0.52	0.52	0.52	0.54	0.66	0.41	0.55	0.20	0.19	0.47	0.44	0.14	0.63	0.78	0.68	1.00	0.32	0.53	0.37	0.20	0.19
eryt	0.13	0.24	0.24	0.25	0.24	0.24	0.21	0.32	0.12	0.13	0.28	0.27	0.05	0.49	0.35	0.27	0.32	1.00	0.71	0.80	0.54	0.58
eryt_f	0.14	0.36	0.36	0.34	0.35	0.46	0.30	0.40	0.14	0.14	0.37	0.35	0.09	0.50	0.56	0.41	0.53	0.71	1.00	0.79	0.58	0.58
eryt_a	0.13	0.27	0.27	0.26	0.27	0.30	0.24	0.31	0.16	0.17	0.30	0.29	0.06	0.42	0.38	0.31	0.37	0.80	0.79	1.00	0.60	0.62
k562_1	0.08	0.08	0.08	0.10	0.09	0.13	0.14	0.17	0.12	0.14	0.15	0.17	0.04	0.24	0.17	0.19	0.20	0.54	0.58	0.60	1.00	0.98
k562_2	0.09	0.08	0.08	0.10	0.10	0.12	0.14	0.17	0.13	0.15	0.15	0.17	0.05	0.26	0.16	0.19	0.19	0.58	0.58	0.62	0.98	1.00

Table 8: Correlation of TAL1 AFS Peak Pile Up Counts at a Cut Off of 20

	tall_p1	tall_p2_1	tall_p2_2	tall_p3_1	tall_p3_2	jurk_sandar_1	jurk_sandar	jurk	rpmi_1	rpmi_2	cem_1	cem_2	ecfc	meka	cd133	cd34	cd34_new	eryt	eryt_f	eryt_a	k562_1	k562_2
tall_p1	1	0.39	0.41	0.58	0.57	0.45	0.68	0.58	0.62	0.63	0.56	0.64	0.00	0.30	0.28	0.40	0.31	0.11	0.12	0.11	0.03	0.04
tall_p2_1	0.39	1.00	0.99	0.80	0.81	0.69	0.53	0.61	0.28	0.30	0.72	0.63	0.06	0.54	0.70	0.51	0.63	0.28	0.44	0.32	0.10	0.10
tall_p2_2	0.41	0.99	1.00	0.81	0.82	0.69	0.55	0.62	0.30	0.32	0.74	0.65	0.06	0.54	0.69	0.52	0.62	0.28	0.44	0.32	0.10	0.10
tall_p3_1	0.58	0.80	0.81	1.00	0.99	0.71	0.65	0.70	0.41	0.43	0.78	0.73	0.06	0.55	0.68	0.56	0.63	0.29	0.42	0.31	0.11	0.11
tall_p3_2	0.57	0.81	0.82	0.99	1.00	0.71	0.65	0.70	0.41	0.43	0.78	0.73	0.06	0.55	0.68	0.57	0.64	0.29	0.42	0.32	0.11	0.11
jurk_sandar_1	0.45	0.69	0.69	0.71	0.71	1.00	0.77	0.79	0.40	0.38	0.72	0.70	0.04	0.53	0.75	0.57	0.70	0.26	0.46	0.31	0.11	0.10
jurk_sandar	0.68	0.53	0.55	0.65	0.65	0.77	1.00	0.84	0.60	0.60	0.74	0.80	0.01	0.40	0.50	0.44	0.44	0.21	0.29	0.23	0.11	0.11
jurk	0.58	0.61	0.62	0.70	0.70	0.79	0.84	1.00	0.53	0.54	0.76	0.79	0.04	0.60	0.66	0.55	0.62	0.32	0.43	0.32	0.15	0.15
rpmi_1	0.62	0.28	0.30	0.41	0.41	0.40	0.60	0.53	1.00	0.97	0.50	0.57	0.04	0.19	0.20	0.35	0.21	0.09	0.12	0.13	0.07	0.08
rpmi_2	0.63	0.30	0.32	0.43	0.43	0.38	0.60	0.54	0.97	1.00	0.52	0.59	0.06	0.20	0.20	0.35	0.20	0.11	0.13	0.15	0.09	0.10
cem_1	0.56	0.72	0.74	0.78	0.78	0.72	0.74	0.76	0.50	0.52	1.00	0.91	0.05	0.52	0.62	0.49	0.53	0.28	0.40	0.32	0.13	0.13
cem_2	0.64	0.63	0.65	0.73	0.73	0.70	0.80	0.79	0.57	0.59	0.91	1.00	0.05	0.50	0.56	0.48	0.49	0.27	0.36	0.30	0.14	0.15
ecfc.tsa	0.00	0.06	0.06	0.06	0.06	0.04	0.01	0.04	0.04	0.06	0.05	0.05	1.00	0.13	0.12	0.17	0.13	-0.01	0.03	-0.01	-0.08	-0.07
meka	0.30	0.54	0.54	0.55	0.55	0.53	0.40	0.60	0.19	0.20	0.52	0.50	0.13	1.00	0.75	0.64	0.71	0.52	0.57	0.47	0.24	0.27
cd133	0.28	0.70	0.69	0.68	0.68	0.75	0.50	0.66	0.20	0.20	0.62	0.56	0.12	0.75	1.00	0.64	0.84	0.39	0.59	0.40	0.15	0.15
cd34	0.40	0.51	0.52	0.56	0.57	0.57	0.44	0.55	0.35	0.35	0.49	0.48	0.17	0.64	0.64	1.00	0.73	0.29	0.42	0.32	0.16	0.17
cd34_new	0.31	0.63	0.62	0.63	0.64	0.70	0.44	0.62	0.21	0.20	0.53	0.49	0.13	0.71	0.84	0.73	1.00	0.36	0.56	0.39	0.18	0.17
eryt	0.11	0.28	0.28	0.29	0.29	0.26	0.21	0.32	0.09	0.11	0.28	0.27	-0.01	0.52	0.39	0.29	0.36	1.00	0.76	0.83	0.54	0.57
eryt_f	0.12	0.44	0.44	0.42	0.42	0.46	0.29	0.43	0.12	0.13	0.40	0.36	0.03	0.57	0.59	0.42	0.56	0.76	1.00	0.82	0.59	0.60
eryt_a	0.11	0.32	0.32	0.31	0.32	0.31	0.23	0.32	0.13	0.15	0.32	0.30	-0.01	0.47	0.40	0.32	0.39	0.83	0.82	1.00	0.61	0.63
k562_1	0.03	0.10	0.10	0.11	0.11	0.11	0.11	0.15	0.07	0.09	0.13	0.14	-0.08	0.24	0.15	0.16	0.18	0.54	0.59	0.61	1.00	0.98
k562_2	0.04	0.10	0.10	0.11	0.11	0.10	0.11	0.15	0.08	0.10	0.13	0.15	-0.07	0.27	0.15	0.17	0.17	0.57	0.60	0.63	0.98	1.00

A.4 Correlation With Technical Factors

To ensure that technical factors, including the number of peaks called in a data set and the number of raw reads aligned uniquely, did not dominate the positioning of data sets along the PCs of interest the correlations were compared. The results for this correlation can be found in Table 9. None of the PCs of interest were seen to have a high correlation with either the number of peaks contributed or the number of aligned unique reads.

Table 9: Summary of the correlation of non biological factors and the principle components found for the AFS analysis of the combind control condition

PCs	Reads	Peaks
PC1	-0.329	0.139
PC2	0.349	0.319
PC3	-0.119	-0.06
PC4	-0.167	0.323
PC5	0.135	0.2
PC6	-0.188	0.336
PC7	-0.061	-0.02
PC8	0.661	0.062
PC9	-0.105	0.012
PC10	0.158	0.092
PC11	-0.036	-0.154
PC12	-0.104	0.023
PC13	-0.324	-0.582
PC14	-0.052	-0.173
PC15	0.029	0.02
PC16	0.044	-0.05
PC17	-0.034	-0.02
PC18	-0.08	-0.185
PC19	-0.139	-0.241
PC20	0.18	0.231
PC21	0.126	0.254
PC22	0.122	0.487

Appendix B

Pilot Analysis

I analyzed a subset of the data in order to pilot the methods that I would be employing. For the pilot scale analysis I compared four data sets. These data sets included a primary eyrthroid sample from the Brand lab two T-ALL cell lines which were a Jurkat sample from the Brand lab and a CEM cell line from the Young lab and CD34+, which represented the HSCs, also from the Young lab. These data sets represented the disparate members of the hematopoietic lineage. Apart from their biological definition these samples also came from different labs and had been sequenced at different times between 2010 and 2014. Taking into account these variations in non biological factors are the basis for my methodology.

For the pilot scale analysis I applied Frequency analysis, the pairwise overlap, along with global analysis and the full AFS method. The results here constitute a more complete analysis of the pilot data than was presented in the Evaluation of Comparison Framework section.

B.1 Pairwise Overlap Values

My initial naive approach to combining data sets was to scale the traditional analysis of two data sets. The result of the scaling was a measure of pairwise overlap between all four cellular conditions. The two T-Cell cell lines I included in the analysis were expected to have the greatest relative overlap and the HSCs were expected to fall

somewhere in between them and the erythroid.

Table 10 shows the absolute number of peaks which overlap in each condition. As expected CEM and Jurkat have the highest overlap with 1,244 peaks or a ratio of 49%.

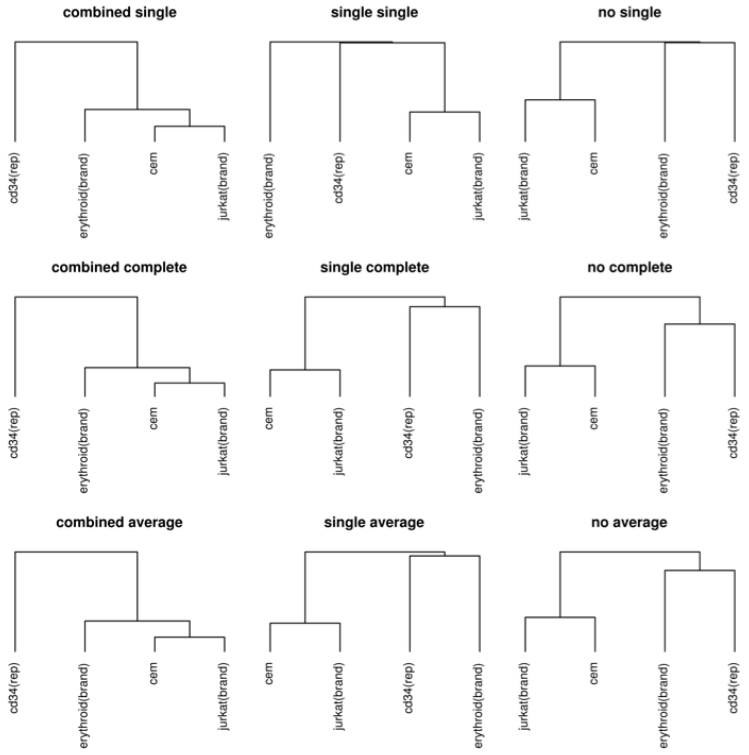
Cell Types	Jurkat	Erythroid	CD34
CEM	1244	300	114
Jurkat		337	92
Erythroid			107

Table 10: The number of peaks identified, using a combined control, in each pairwise case, with a minimal overlap of 1bp.

Previous experiments by Palii et. al. [3] have shown that the erythroid and jurkat cell types have an overlap of 380 peaks when a shared mock was used. This was equivalent to a overlap ratio of 15%. The results shown here using a combined mock for pairwise analysis were slightly more conservative estimating 337 regions intersecting between the Jurkat and Erythroid cellular conditions.

B.2 Hierarchical Clustering the Frequency Data

Figure 29: The results of applying hierarchical clustering to the combinations of the overlaps found using three controls and the three linkages of interest. The results are represented here as dendograms.



Once the pairwise overlap ratios were found for each of the data sets, I applied hierarchical clustering with the ratios being the distance metric and using three linkages (single, complete, and average). The results of applying the three linkages to the three control cases can be seen in Figure 29. For the combined control the T-ALL cell lines grouped closest and were distinctly isolated from both the erythroid and HSCs. These results were anticipated due to the expected high similarity between the T-ALL cell lines.

The use of the individual controls when calling peaks (rather than a combined control) also resulted in a consistent grouping of the T-ALL conditions across the three linkages.

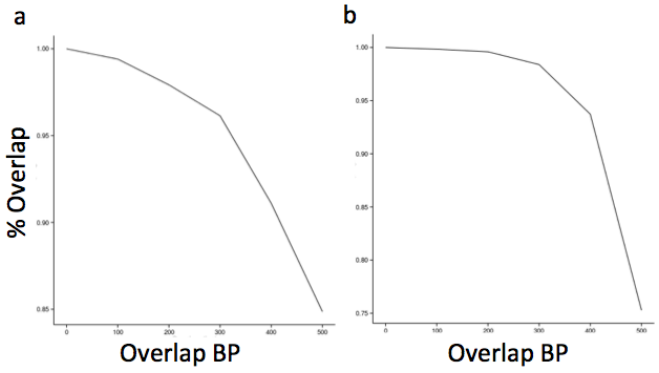
In the case were no control was used when calling peaks, the complete linkage method had two distinct clusters form. With the use of the complete linkage the

Erythroid cellular condition was closer to the CD34 than one of the two T-ALL conditions. In the single linakge case the HSC erythroid cluster does not exits meaning that the erythroid was closer to one of the T-ALL cases than it was to the HSC. This case shows an example of the brittleness of relying on a single value inherent in the single and complete linkages

B.3 Varying the Overlap Ratio

The overlap ratio required for two peaks to be considered overlapping is referred to as the overlap stringency. The selection of overlap stringency was a variable that impacts the representation of overlapping regions. To determine the degree to which changing the stringency of overlap would effect the number of overlapping regions I preformed the pairwise analysis at overlap stringencies ranging from an overlap of 1bp to 500bps.

Figure 30: The effect of changing the overlap stringency on the number of pariwise overlaping regions. Each case was normalized to the lowest stringecy (1bp) a.) The overlap of the Erythroid and Jurkat and the b.) overlap of the Jurkat and CEM.



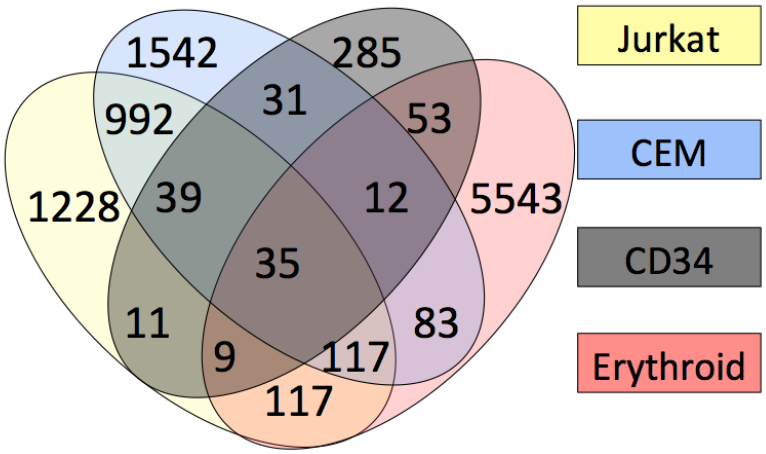
I normalized the values to the initial overlap value. The results of altering the overlap between CEM and Jurkat Can be seen in Figure 30. Increasing the overlap ratio from 1bp to 400bps (out of 600) resulted in less than a 10% loss in the number

of overlapping peaks. The effect of changing the overlap between the Erythroid and Jurkat conditions can be seen in Figure 30 as well. The drop off for the Erythroid Jurkat case was steeper, when compared with the Jurkat CEM case, however after 400bp 90% of the overlapped peaks were still overlapping. Since there was less than a 10% difference between defining overlap as 1bp and defining overlap as 400bps, I set the default to 1bp for the remainder of the investigation.

B.4 The Full Combinations of All Data Sets

Pairwise analysis was useful for defining how cell types relate to one another, however once the number of data sets being compared surpasses two it no longer fully characterized the interaction between the data sets. The number of combinations increases as a power of two for each data set added. For example with three data sets the number of unique entities in the set increased to 8 (including the null case). At this point the three way overlap needed to be taken into account in addition to the two way comparisons. For 4 data sets there are 16 unique bins, the composition of these bins can be visualized in Figure 31. For the full scale analysis the number of data sets increased to 22, thus the number of unique bins increased to 4,194,304.

Figure 31: A venn diagram representation of the peaks shared between four cellular condition, CEM, Jurkat, CD34+ and Erythroid



The 15 intersections which are possible to preform on the 4 data sets are
(1) CEM,
(2) $Jurkat$,
(3) $Erythroid$,
(4) $CD34$,
(5) $CEM \cap Jurkat$,
(6) $CEM \cap Erythroid$,
(7) $CEM \cap CD34$,
(8) $Jrukat \cap Erythroid$,
(9) $Jrukat \cap CD34$,
(10) $Erythroid \cap CD34$,
(11) $CEM \cap Jurkat$
(12)$CEM \cap Erythroid$,
(13) $CEM \cap Jurkat \cap CD34$,
(14) $Jurkat \cap Erythroid \cap CD34$,
(15) $CEM \cap Jurkat \cap Eyrhtroid \cap CD34$.

From these intersections the values of each of the bins can be determined using Equation 2, where L is the number of data sets being analyzed, s is the number of true values in the comparison and N is the dimensional order of A, i.e. the intersection operations have been partitioned into sections based on how many contribute to the argument of interest.

$$\sum_{n=s}^{L} -1^{n-1} \sum_{i=0}^{Li} A^{N=n}[i] \tag{2}$$

An example bin of interest would be $A \cup B \cup \bar{C} \cup \bar{D}$ and would be could be generated via

$A \cup B \cup \bar{C} \cup \bar{D} = A \cup B - A \cup B \cup C - A \cup B \cup D + A \cup B + D + C$

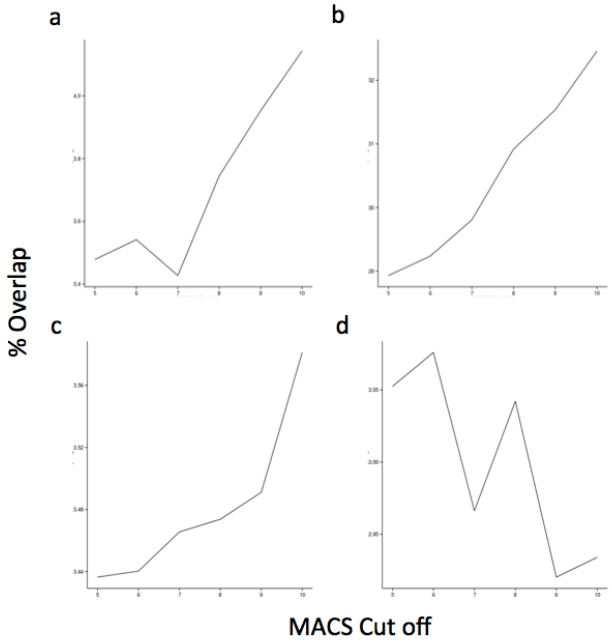
It is feasible to do this analysis for a small set of data, but as the number of data sets increases the number of unique bins increase as a power of two. Even when comparing 4 data sets several of the bins were approaching a size of 0. As the number of possible combinations approaches a million the number of null bins will increase

accordingly.

B.5 Varying the P-Value Cut Off

The statistical cut off used to define peak regions was the peak p-value. By altering the required p-value the regions that were enriched changed. Figure 32 shows the changes in overlap value between the two Leukemic data sets and between the Leukemic and Erythroid conditions. Ideally as stringency increased the overlap between similar cell types would increase and the overlap of significantly different cellular environments would decrease. While there is a clear increase in the CEM and Jurkat overlap of 3% there was also an increase of 1% in the erythroid. While the absolute increase of the Jurkat CEM case was greater the relative increase was less.

Figure 32: The effect of increasing the cut-off stringency on the correlation between cellular conditions. MACS score is equivalent to -log10(pvalue) a.) relation between Jurkat and Erythroid b.) relation between Jurkat and CEM c.) relation between CEM and Erythroid d.) the relation between CD34+ and Jurkat



This large variance due to stringency cut off makes overlap analysis susceptible to significant alteration with small changes in data set quality. In order to avoid these effects, and the combination limitations alternative methods were investigated.

B.6 Global Analysis

Global analysis does away with peak detection so the fragility due to stringency cut off and the scalability limitations does not influence the final analysis. For the purpose of this analysis the I split the genome into non overlapping bins of 1000bp. Bin sizes of 10,000 and 100,000 were also used to gauge the precision needed for accurate analysis. The correlation matrix was generated and the results for the three overlap values, 1000, 10,000 and 1000,000 can be seen in tables 11 , 12, and 13 respectively.

Table 11: The correlation of the global analysis of Erythroid, HSC, and two T-ALL cell lines, generated with a bin size of 1,000bp.

Cell Lines	Erythroid(brand)	CD34(rep)	CEM	Jurkat(brand)
Erythroid(brand)	1	0.314	0.28	0.439
CD34(rep)	0.314	1	0.316	0.392
CEM	0.28	0.316	1	0.439
Jurkat(brand)	0.439	0.392	0.439	1

Table 12: The correlation of the global analysis of Erythroid, HSC, and two T-ALL cell lines, generated with a bin size of 10,000bp.

Cell Lines	Erythroid(brand)	CD34(rep)	CEM	Jurkat(brand)
Erythroid(brand)	1	0.498	0.496	0.737
CD34(rep)	0.498	1	0.51	0.573
CEM	0.496	0.51	1	0.615
Jurkat(brand)	0.737	0.573	0.615	1

Hierarchical clustering was applied to the Dissimilarity matrix (1-corelation). Dissimilarity, as a distance metric, yielded distinct results from those seen using the

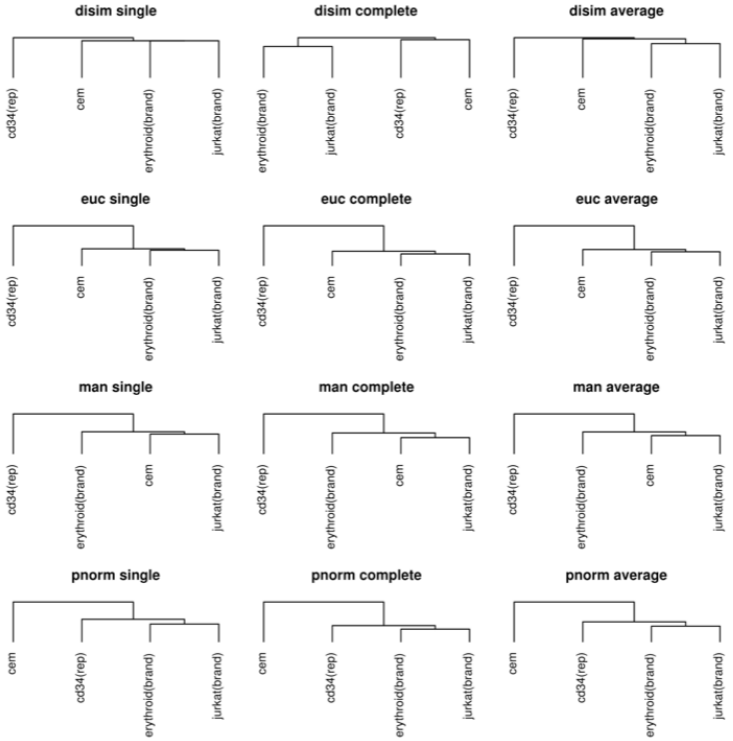
Table 13: The correlation of the global analysis of Erythroid, HSC, and two T-ALL cell lines, generated with a bin size of 100,000bp.

Cell Lines	Erythroid(brand)	CD34(rep)	CEN	Jurkat(brand)
Erythroid(brand)	1	0.543	0.694	0.878
CD34(rep)	0.543	1	0.663	0.599
CEM	0.694	0.663	1	0.763
Jurkat(brand)	0.878	0.599	0.763	1

overlap analysis. The results in dendrogram from can be seen in Figure 33.

The dissimilarity matrix grouped the Erythroid and Jurkat cell types together for each of the cellular conditions. The use of the L10 (pnrom where p is 10) and the L2 (euclidean) grouped the data similarly. Only the use of the manhattan (L1) separated the data the same as the overlap analysis method.

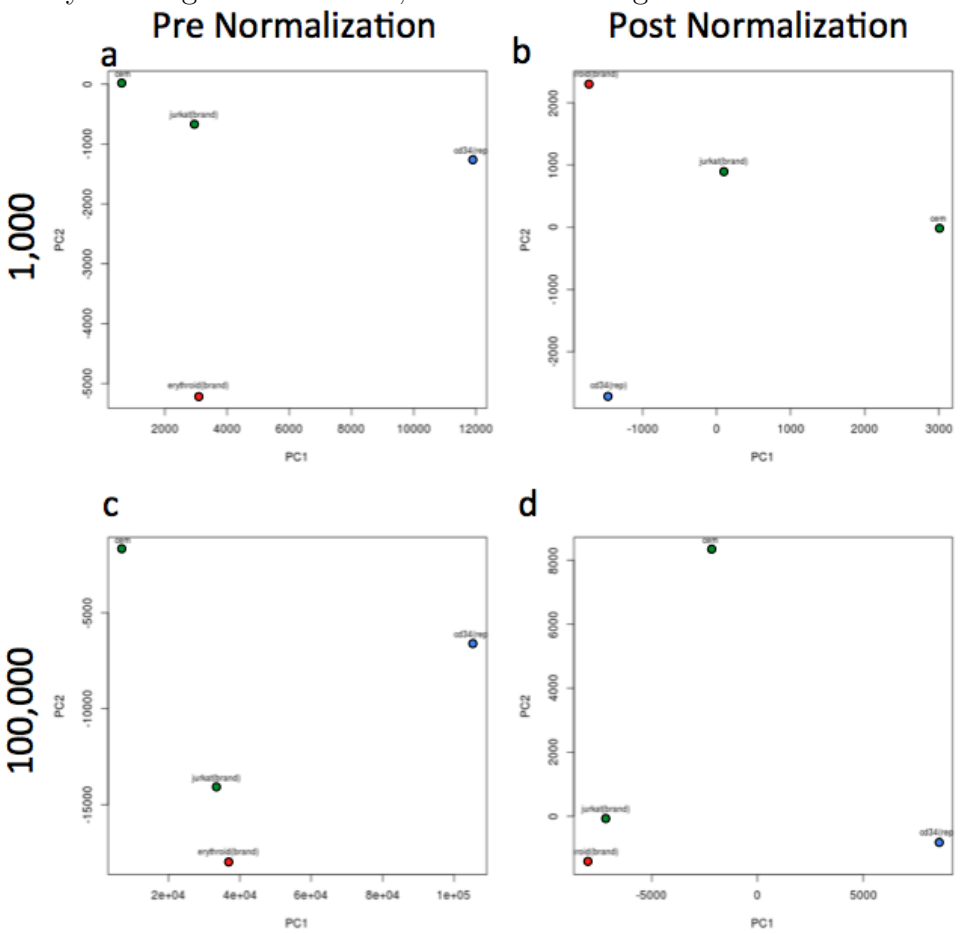
Figure 33: The results of applying heirarchichal clustering to the global analysis results were assesed using a dendrogram visualization. The figure contains all of the linkages and distance metrics used to assess the global analysis data. Man:manhattan, euc: Euclidan, pnorm: p-norm, and disim: dist(1-cor)



After the hierarchical clustering I applied the PCA to the results of the global analysis. The features in each category represented read pile up counter under each of the unique regions in the genome. Crossing the first two eigenvectors with the normalized data can be visualized in Figure 35 a. The first principle component separated the T-ALL cell lines (CEM and Jurkat) and the HSPC data set (CD34). The second principle component separated the cellular conditions capable of self renewal (CEM Jurkat and CD34) from the differentiated erythroid.

The application of quantile normalization as seen in Figure 35 b changes the representation. With the normalization the Jurkat and Eythroid cellular condition (both of which come from the same lab) come closer together. This trend is exacerbated as the bin size is increased from 1,000 to 100,000 as seen in Figures 35 c and d.

Figure 34: The histogram of the data projected onto the first principle component. Global analysis using a bin size of 1,000 was used to generate the data.



B.7 Principle Component Histogram

The projection of the bins onto each of the principle components can be used to generate a histogram. Using this as a visualization technique I investigated if the results were bimodal. If there were two or more clear peaks, in the histogram, we could then apply a linear discriminant to the lower dimensional data in order to isolate regions of expected biological relevance.

Figure 35: The histogram of the data projected onto the first principle component. Global analysis using a bin size of 1,000 was used to generate the data.

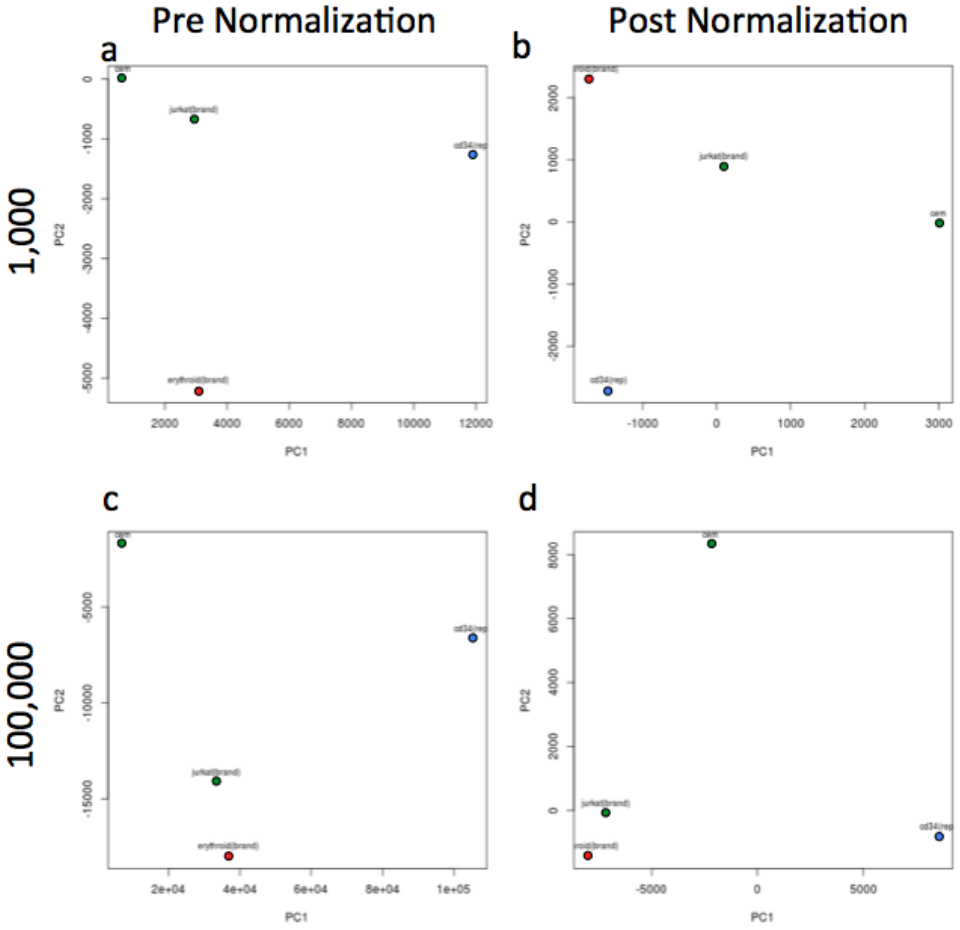


Figure 35 shows the projection of the global analysis with a bin size of 1000 along the first principle component. There was only a single peak seen in the analysis. The relation between axis position and contribution by cellular condition could not be determined explicitly since there was no way to compare a bin directly to the raw data.

B.8 Unified Peak Set

My alternative to analyzing the full genome was to analyze only those regions that were most important to TAL1. Those regions being, the TAL1 ChIP-Seq peaks. To generate a single independent set of regions, I found the union of the peaks found in each data set. The result of this union was the AFS. The AFS for the four pilot data sets, called using a combined mock and a peak width defined as 700bps, contained 10,342 peaks. These peaks are what define the alternate feature space (AFS).

B.9 Applying Global Analysis Techniques to the Unified Matrix

I determined read pile up counts under all peaks in the AFS for each contributing data set. For each peak region the pile up counts yielded a range rather than a binary value, thus a quantifiable relationship could be maintained between the peaks. The resulting matrix was referred to as the unified density matrix (UDM). By defining the regions of interest as the AFS subset of the genome the global analysis was limited to regions that the TF of interest is most likely to bind. This was an attempt to minimize the impact of non specific binding on the comparison signal. The correlations between the values generated using the AFS approach can be seen in Table 14.

Table 14: The correlation within the alternate feature space for the 4 pilot cell types

Cell Lines	Erythroid(brand)	CD34(rep)	CEM	Jurkat(brand)
Erythroid(brand)	1	0.388	0.109	0.197
CD34(rep)	0.388	1	0.579	0.662
CEM	0.109	0.579	1	0.738
Jurkat(brand)	0.197	0.662	0.738	1

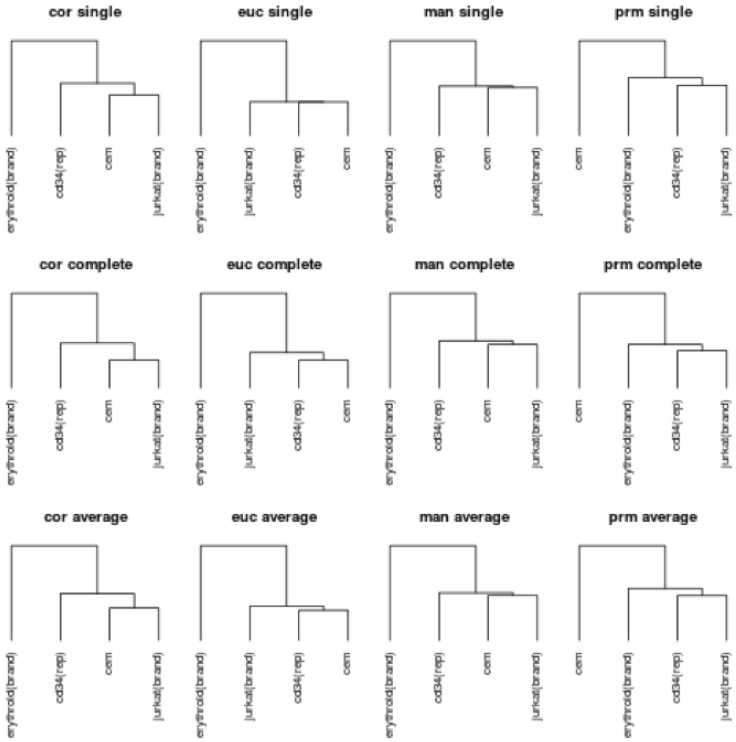
As expected Jurkat and CEM, both being T-ALL cell lines, had a correlation of almost 0.8. The correlation of Erythroid with both T-ALL cell lines was less than

0.2. The relation between CD34 and the T-ALL cell lines ranged between 0.58 and 0.63, less than within the T-ALL group but greater than the relation of T-ALL and Eythroid.

Hierarchical Clustering

The results of the hierarchical clustering of the UDM can be seen in Figure 36. Using Dissimilarity and the L1 as the distance measures yielded the expected biological result where the two T-ALL cell lines group closest. As distance order increases to L2 and L10 the emphasis on shared regions with the highest value increases. The result of this is observed in the increasing similarity between the Erythroid and Jurkat cellular condition. These cell types came from the same laboratory and were done in the same experiment.

Figure 36: The result of applying hierarchical clustering to the AFS combined data was assessed using a dendrogram visualization. The figure contains all of the variants used to assess the AFSanalysis data with a combined control. man:manhattan, euc: Euclidan, pnorm: p-norm, and disim: dist(1-cor)

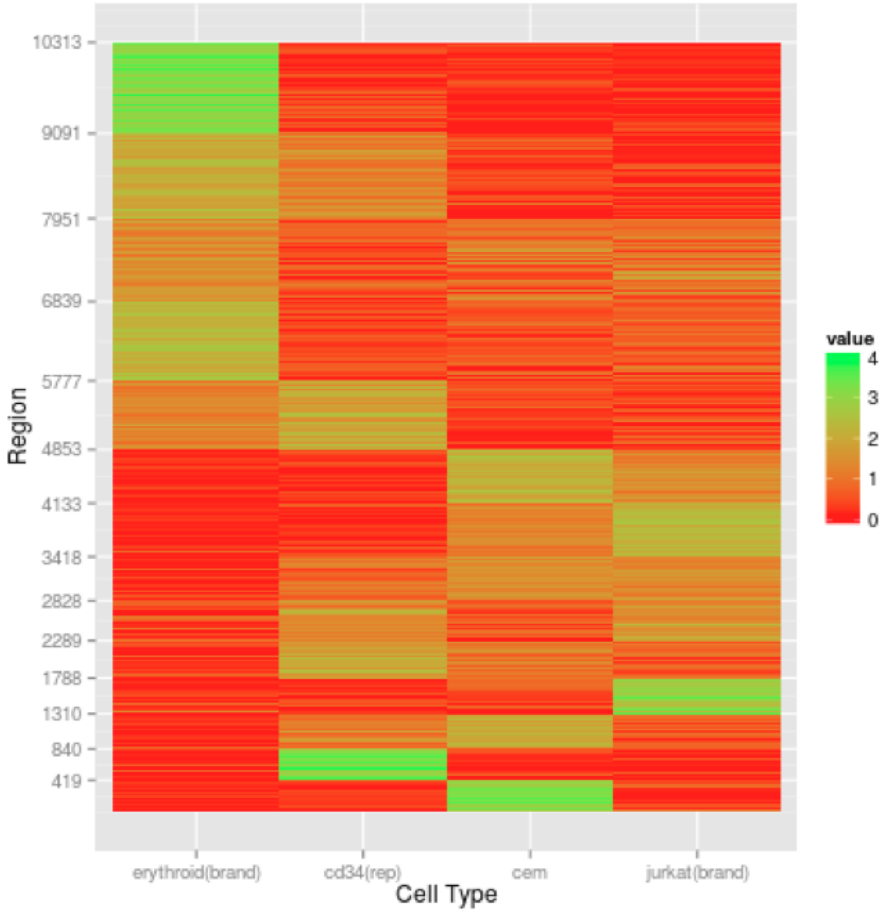


K-means Clustering

The pile up values for each peak were quantile normalized and clustered using k-means clustering. In addition to being quantile normalized by cellular condition each peak group was normalized by its mean.

14 centers were chosen for the representation. This value was chosen relative to the number of data sets being compared. The results can be seen in Figure 37.

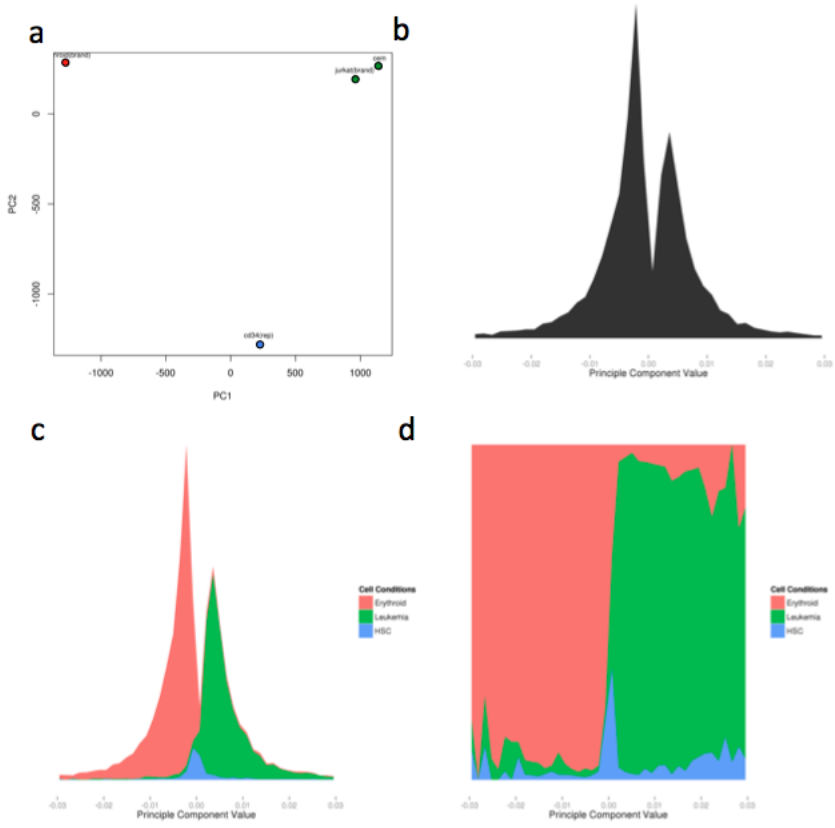
Figure 37: The 10,313 enrichment regions of the unified peak set generated using the combined mock clustered along the y axis using k-means clustering.



From the k-means clusters regions may be isolated based on which cluster they fall into. For example, the cluster which is green in both CEM and Jurkat represents the regions which are important in both cellular conditions. Unfortunately, this analysis only groups the data and it does not give a value on which these regions may be sorted and allows for no adjustment related to biological significance. This results in a loss of quantifiability in the analysis.

PCA

Figure 38: a. The PCA plot, representing the first two principle components, generated using the AFS approach. b the histogram of the afs combined data projected onto the first principle component. c The histogram of the first principle component. Each region relates to a source data set. d the histogram of the first principle component. Each region relates to a source data set. The results are normalized such that each bin sums to one.



The first two eigenvectors for the PCA of the unified data matrix can be seen in Figure 38a. The first principle component separates the Leukemic cell lines from the healthy differentiated Erythroid primary cells. The second principle component

separates the differentiated cellular conditions from the CD34+. These groupings are similar to those seen in the global analysis with a bin size of 1,000. The use of the AFS resulted in a closer grouping of the T-ALL cell lines.

Histogram Contribution

As when doing global analysis the features of the AFS can be projected onto the resulting principle components. Looking at the resulting histogram in Figure 38b we can see two clear peaks present along the first principle component.

The components of the eigenvectors in the AFS, unlike in the global analysis, consist of enriched peak regions. Due to this relation the contribution to the extrema of the vectors can be determined. The relative value of contribution will indicate if there is enrichment of a specific category at either end of the vector.

The contribution to each region based on the data set which provided the peak can be seen in Figure 38 c. There is a clear shift within one standard deviation of the mean of the contribution of each data set transferring from T-ALL to Erythroid.

Appendix C

Global Analysis Method

C.1 Background

While the analysis of a single data set is well defined the comparison of several data sets poses a number of difficulties. When comparing data sets several approaches are used primarily. A global analysis approach, which takes the full genome into account, is presented here.

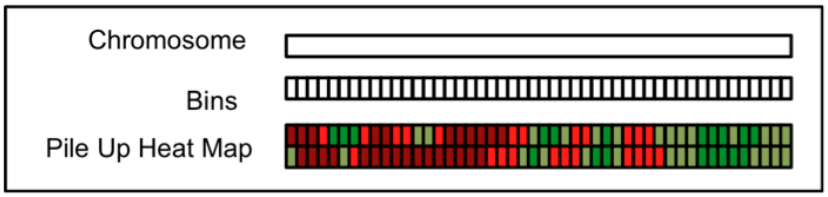
Global analysis does not require the identification of highly enriched regions. Rather the analysis relies on the notion that the global signal is dominated by transcription factor binding and that the correlation of genome wide signal between two data sets provides a measure of similarity. What is not clear is whether this similarity is originates from true transcription factor binding or whether it arises from general chromatin accessibility.

C.2 Method

For the purpose of global analysis I split the genome into equal sizes bins, each independent of one another, i.e. non overlapping. Once the genome was divided I found the pile up count of reads for each of these bins, for each of the data sets. The result was a matrix similar to the UDM. I investigated the efficacy by modifying the size of the bins. Additionally, the impact of the control was determined by applying

the same methodology on the control data set.

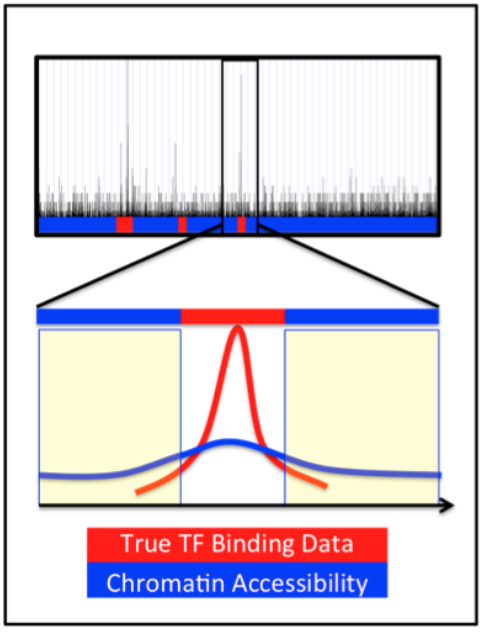
Figure 39: Each chromosome is split into bins of equal size. The read pile up is then determined for each bin for each data set being compared. The resulting matrix can be used to compare the data sets using a variety of methods.



Once the pile up had been determined for each of the data sets independently I determined their global similarity using their correlation. By using the dissimilarity of the data sets (1 - correlation), the data can be investigated using general purpose clustering methods such as hierarchical clustering and k-means clustering.

The limitations of global analysis arise from the non specific background signal. This signal level changes across the genome and between cell types based on the local chromatin accessibility. These issues are represented in Figure 40.

Figure 40: The ChIP-Seq signal contains both specific and non specific binding signal. The specific signal pertains to the substantial binding locations of the protein of interest and the non specific signal comes from either transient binding or fragments that have no protein of interest. ChIP enriched regions shown in red, are locations where the signal is found to be significantly greater than a control, or the surrounding region. Outside of these enriched regions the majority of the signal may come from the non-specific sources, shown in blue.



The red line represents the true binding in the ChIP signal. The blue line represents the background non specific noise. Within the regions that are determined to be statistical enriched the true binding makes up most of the signal. Outside of these regions however the majority of the signal is contributed by background. Since the majority of the genome coverage is outside of the enriched regions, the majority of the total signal stems from non specific binding.

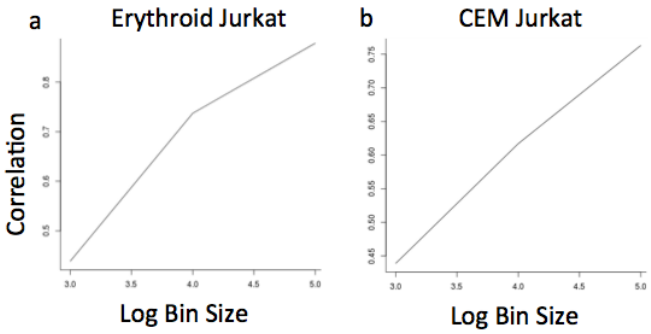
C.3 Results

I analyzed the 22 TAL1 ChIP-Seq data sets using global analysis. To compare the impact of changing bin sizes I repeated the analysis using Bin sizes of 1,000, 10,000 and 100,000. With a bin size of 1,000, 3,095,689 unique regions were generated. Tp put this in context; this was larger than the number of mapped reeds for several of the data sets. The 10,000bp and 100,000bp bin sizes each respectively had 309,579 and 30,970 regions.

C.3.1 Global Correlation Matrix

Figure 41 shows how the correlation between Eythroid and Leukemic data sets change as a function of bin size. For both data sets the correlation increases logarithmically as the size of the bins increase.

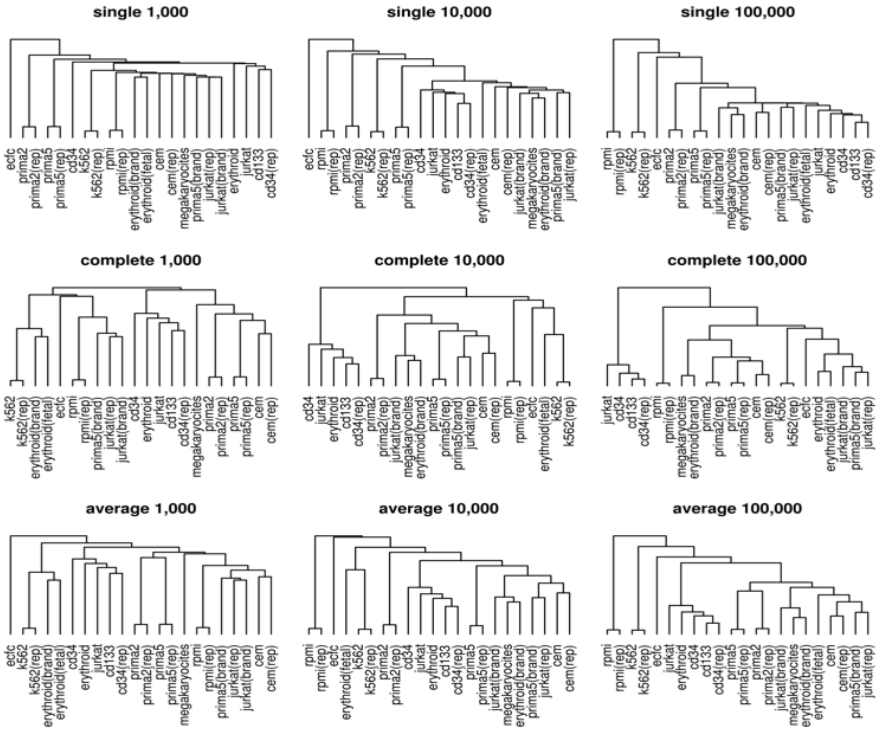
Figure 41: The correlation was assessed for three bin sizes, 1,000 , 10,000 and 100,000. Panel a depicts the correlation of Erythroid and Jurkat as a function of bin size. Panel b depicts the correlation of Jurkat and CEM as a function of bin size.



C.3.2 Hierarchical Clustering

The hierarchical clustering of the global analysis data matrix was preformed using dissimilarity as the measure of distance. The three linkages (single, complete, and average) were compared in addition to the impact of changing the bin size.

Figure 42: Dendrograms generated from global analysis using three bin sizes and dissimilarity as the measure of distance between data sets.



The results of the hierarchical clustering can be seen in Figure 42. For the case where the bin size was set to 1,000 the ECFC clustered separately from the principle Hematopoietic groups. Using the single linkage there is not a clear separation between the Leukemic and Endotheilial cellular conditions. When the average linkage is applied to the 1,000 bin size data there are a few partial groupings seen. All of the erythroid cellular conditions grouped together save one. The HSCs were also grouped closely, however there was also an Erythroid and a Jurkat cellular condition present in their cluster.

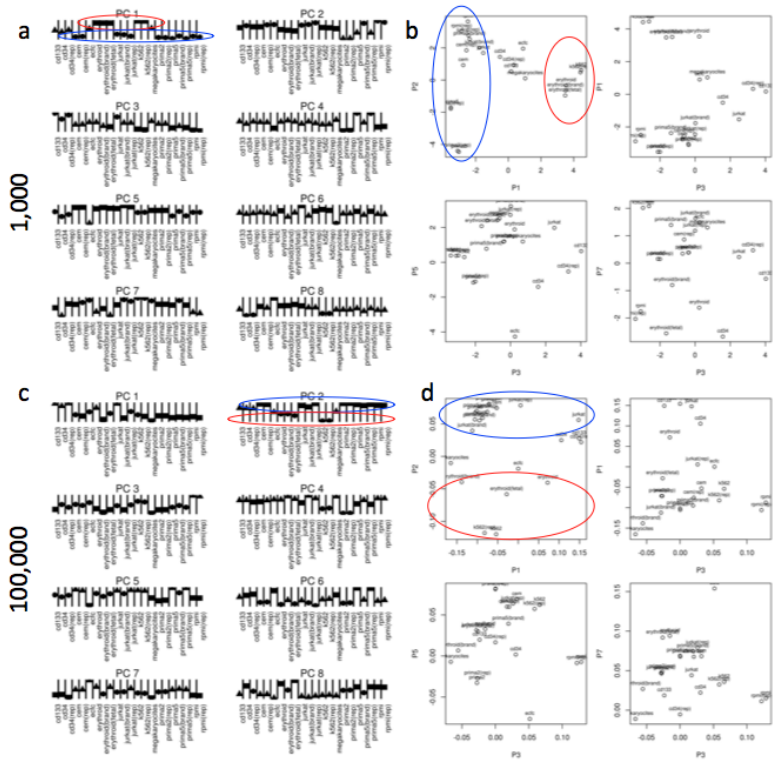
Increasing the bin size caused an observable trend of large data sets moving out of the inner clusters. This trend can is seen when using both the single and the average linkages. This trend is shown by the largest data sets (RPMI and K562) breaking

out of the clusters with their respective cell types and forming branches independent of other clusters.

C.3.3 Principle Component Analysis

I applied the principle component analysis (PCA) to the global read pile up counts in order to try to extract the highest variability of the data in few dimensions. I did this in the hopes of finding a distinct separation vector. Visualizing the first principle component for 1,000 bp bins showed that it clearly separates between erythroid cell types and the Leukemic cell types (Figure 43). As I increased the bin size however, this separation became less clear. As I increased the bin size further to 100,000 (Figure 43) the entire notion of separation along the first axis became apparent in the second principle component.

Figure 43: The results of PCA of the global ChIP signal. a depicts the first eight principle components found using the global analysis method with a bin size of 1,000bp. b was a selection of principle components found using the global analysis method with a bin size of 1,000bp represented in two dimensions. c the first eight principle components found using the global analysis method with a bin size of 100,000bp. d a selection of principle components found using the global analysis method with a bin size of 100,000bp represented in two dimensions.



C.3.4 Correlation With Technical Factors

It was anticipated that the principle components would correlate highly with the number of reads in each data set however the correlation was no greater than what was seen with the overlap analysis. Interestingly there the global analysis results correlated more with the number of peaks though in only one case over 0.41. The results can be seen in Table 15.

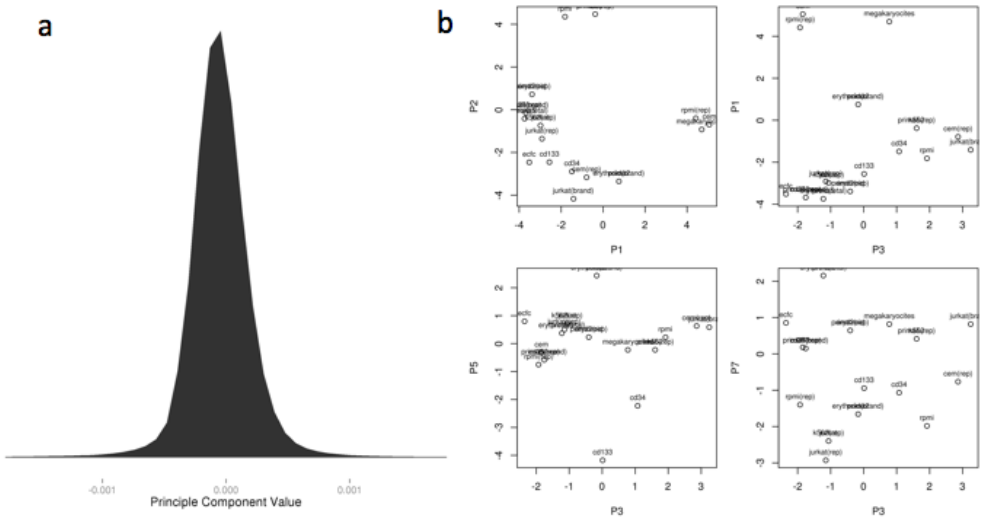
Table 15: Summary of the correlation of non biological factors and the principle components found for the global analysis

PCs	Reads	Peaks
PC1	0.223	0.144
PC2	0.139	0.414
PC3	-0.175	-0.234
PC4	-0.059	-0.254
PC5	-0.091	-0.66
PC6	-0.249	-0.371
PC7	0.358	0.217
PC8	-0.012	0.161
PC9	0.128	0.149
PC10	-0.185	0.084
PC11	0.134	0.018
PC12	-0.353	0.012
PC13	0.039	0.006
PC14	-0.031	-0.004
PC15	0.542	-0.01
PC16	-0.018	-0.004
PC17	0.447	-0.013
PC18	-0.057	0.107
PC19	0.001	0.02
PC20	0.003	0.008
PC21	0.069	0.008
PC22	0.252	0.122

C.3.5 Principle Component Histogram

The projection of the data onto the first principle component yielded the histogram seen in Figure 44. The data was not bimodal and did not provide an inherent means of identifying the relation between the values in the eigenvector and the biological importance.

Figure 44: a the histogram of the first principle component of the global analysis results using a bin size of 1,000. b the control data sets have been swapped with the treatment



C.3.6 Background Influence

The results of changing the Treatment data for the control data when preforming PCA can be seen in Figure 44 was composed using the 12 control data sets using a bin size of 1,000. There is no clear separation seen between the erythroid and leukemia. As the bin size increases to 100,000 it becomes clear that data sets are grouping based on similarity in cellular condition. The CD34 are in close proximity along the second principle component and the K562 replicates are in close proximity along both the first and second principle component.

Appendix D

Overlap Analysis Method

D.1 Background

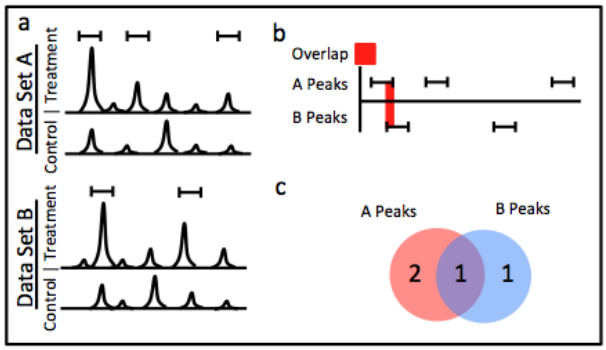
Overlap analysis relies on the identification of regions of high binding confidence. These regions are stringently determined for each data set individually. Once the data sets have been analyzed independently their information must be combined in order to make a comparison. The most naive approach to comparing cell type similarity is to determine the frequency of peak overlap. The result of a binary analysis is quite clear, peaks can either be shared or belong to a single data set. This binary pairwise method can be extended to the comparison of more than one data set, however, the pairwise overlap does not constitute the full characterization of the relationship. While this simple overlap analysis is useful for comparing data sets and identifying which of the largest peaks fall in each category it is limited by its dependence on threshold regions.

D.2 Methods

The comparison of peak overlap, for the purpose of combining data sets, relies on the independent identification of enriched regions. Once identified these regions are intersected between each condition to determine which are shared and which are differential. The ratio of shared to differential provides a clear metric of similarity

between cellular conditions. Figure 45 demonstrates the process of overlap analysis in the binary sense.

Figure 45: An overview of the overlap method. a the two data sets are independently analyzed to identify a set of peak locations for each. b the locations are compared. c a Venn diagram representing the unique regions that are counted and contrasted with the shared.

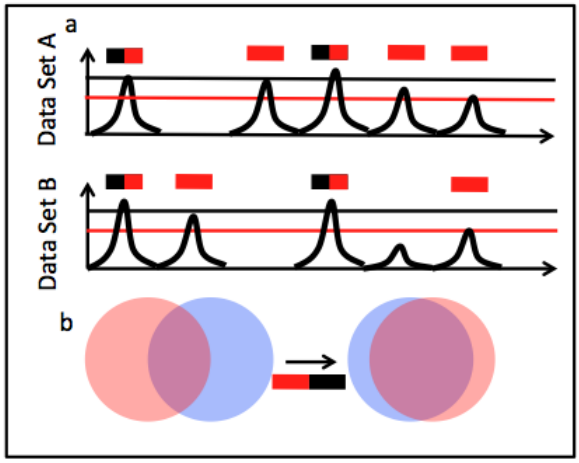


As more data sets are included in the analysis the comparison becomes more complex. While the pairwise comparison of each data set yields useful information that can be used for grouping cell types, it does not completely characterize the data sets. In order to accomplish a complete characterization of each data set the pairwise comparison must be extended exponentially in order to incorporate the increasing number of data sets. This greatly increases the complexity of the analysis.

For this analysis I created a unified set of peaks (the AFS) in order to simplify the analysis. The score for each data a set for each peak can be either a 1 or a 0 based on whether or not it contributed to the unified peak.

Once the data has been reduced into a binary matrix of unified reads hierarchical clustering and PCA were applied to group peak regions for visualization. The similarity of the data sets was visualized using a dendrogram.

Figure 46: The results of comparing two data sets at different cut offs. The red line represents the first cut off and the black line the second. Those peaks with solid red bars are only significant for the first cut off where as those with red and black bars are significant for both. The impact of this shift can be seen in b, where all locations now overlap as the stringency is increased.



The overlap analysis was limited due to its reliance on threasholding. The issue with threasholding can be seen in Figure 46. As the stringency changes the peaks included in each data set change. In most data sets overlap is not independent of stringency [6].

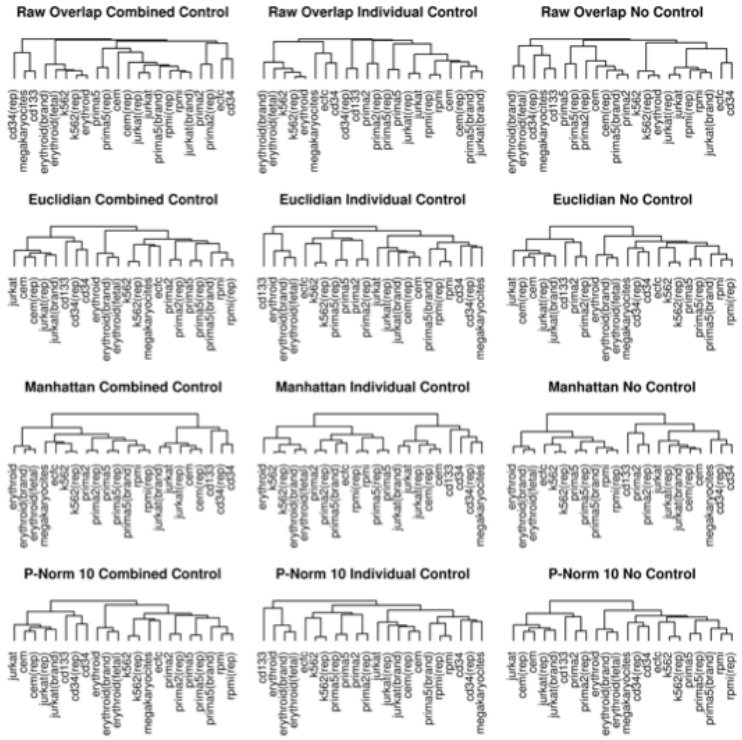
D.3 Results

The overlap methodology was applied to all 22 data sets. The inclusion of the ECFC TAL1 data set was expected to be an outlier in the analysis, thus act as a negative control for similarity. Several data sets with replicates were not combined. The replicates instead were expected to group closely together and act as a positive indicator of similarity.

D.3.1 Hierarchical Clustering

The dendrogram for the overlap matrix of peaks, called using each of the controls, can be seen in Figure 47. I applied the L1, L2, and L10 distance metrics to the overlap matrix to see if the combination of spatial measure and the overlap would yield a greater inter cellular variance.

Figure 47: The results of the hierarchically clustering of the overlap data represented as Dendrograms



Applying the hierarchical clustering to the combined overlap yielded 5 distinct groupings. The Erythroid and k562 cellular conditions grouped together as expected. The leukemic cell lines apart from the primary patient 2 grouped closely as well. The ECFC, which was expected to be an outlier, instead formed a cluster with a CD34+ condition. Applying the L2 distance measure to the matrix split the erythroid and leukemic clusters and moved the ECFC closer to the differentiated cellular condition.

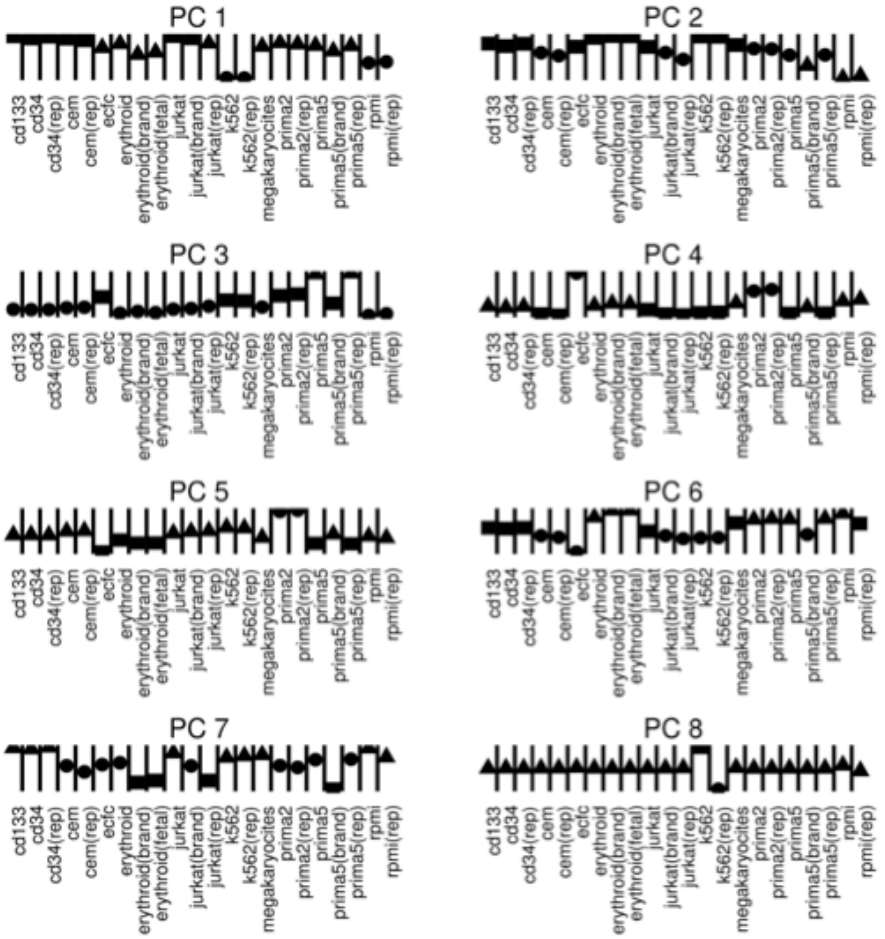
This shift may have been due to the high number of peaks called in the K562 and ECFC condition.

As the control used to identify peaks changed the clustering of the overlap matrix was rearranged. Despite the rearrangement the groupings of the Leukemic and Erythroid conditions remained stable and the ECFC was always seen to be closely related to the HSC cellular condition CD34.

D.3.2 PCA

PCA was applied to the overlap binary matrix. The results of the projection of the data onto the first eight PCs can be seen in Figure 48. As a means of visualizing groupings, k-means clustering was applied to the data along each dimension independently.

Figure 48: The visualization of the first 8 Principle Components.



The first principle component separated the K562 and RPMI data sets from the remaining cellular conditions. The second principle component partially separated the Erythroid conditions from the Leukemic, however the separation was dominated by the RPMI cell line. Additional separation between cellular conditions can be discerned however there are no other vectors of biological interest. Each principle component contains less variance than the previous. By the eighth principle component the internal variance between the K562 replicates was greater than the intercellular variance.

D.3.3 Correlation with Technical Factors

The dot product of the PCA and the data sets were correlated with the mapped read count and the number of peaks called in order to determine if any principle component may be dominated by either factor. The results are presented in Table 16

Table 16: Summary of the correlation of non biological factors and the principle components found for the overlap analysis of the combind control condition

PCs	Reads	Peaks
PC1	-0.166	-0.642
PC2	-0.011	0.186
PC3	-0.132	-0.078
PC4	-0.047	0.062
PC5	-0.219	-0.422
PC6	0.483	0.512
PC7	-0.303	0.175
PC8	-0.146	-0.183
PC9	-0.427	0.179
PC10	-0.584	0.063
PC11	-0.022	-0.023
PC12	0.024	-0.027
PC13	0.113	0.019
PC14	-0.001	0.009
PC15	0.031	0.005
PC16	0.006	-0.001
PC17	0.006	-0.015
PC18	-0.016	0.02
PC19	-0.086	0.015
PC20	0.108	0.007
PC21	0.285	-0.144

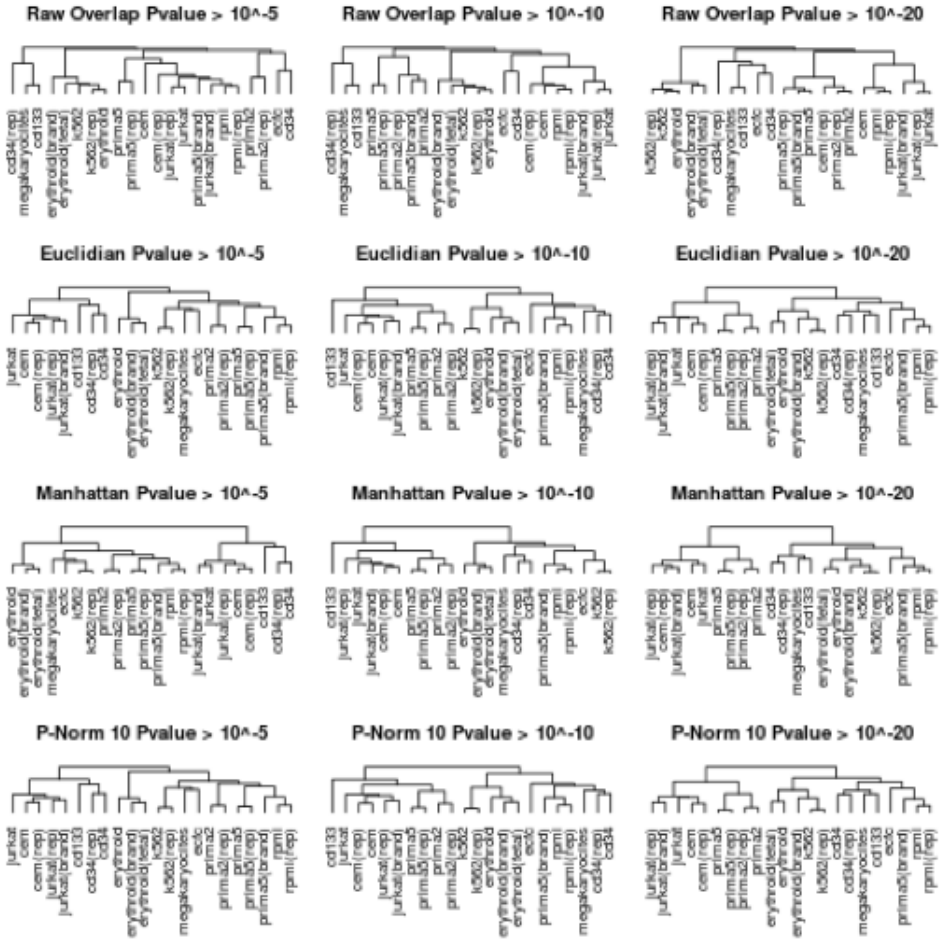
The first principle component indicates the strongest correlation between the data and the number of peaks. Since the first principle component was seen to be the separation between data sets with many peaks (K562 and RPMI) it may be that the overlaps with other cellular conditions were saturated before they were saturated

with each other. The first five principle components do not show strong correlation with the number of uniquely mapped reads.

D.3.4 Changing the Cut Offs and the Overlap Limits

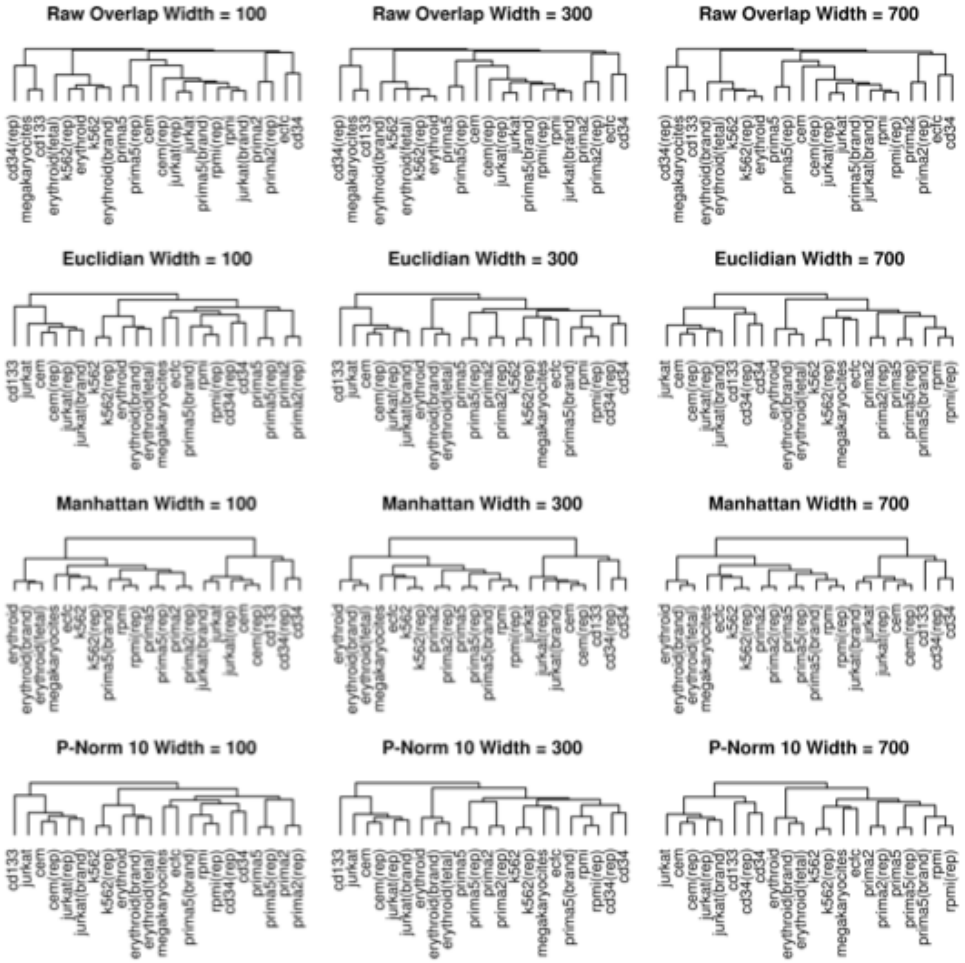
The data was analyzed at p-value cut offs ranging from the default MACS p-value cut off of 10^{-5} to the highly stringent cut off of 10^{-20}. The changes were applied to the overlap values generated from the combined control data. The effect of applying L1,L2 and L10 to each of these stringencies can be seen in Figure 49.

Figure 49: The effect of changing the p-value cut off on the hierarchical clustering based off of the overlap between cellular condition enrichment regions



The overlap stringency was changed by changing the width of the peaks being compared. The widths were initially set to be 700 bp wide. The effects of shortening the widths to 300 and 100 bp can be seen in Figure 50.

Figure 50: The effect of changing the required overlap on the hierarchical clustering based off of the overlap between cellular condition enrichment regions



The use of the overlap as a measure of similarity was highly resilient to the variation in minimal overlap required. There were small variations within the clusters as the requirements became more stringent however there was no change in the overall groupings. Applying the overlap changes to the L1 distance metric also had little impact on the groupings of the cellular conditions. The L2 and L10 metrics however saw slight changes in the proximity of the HSCs and the ECFC conditions.

Appendix E

Biological Results

The biological results presented in the thesis were a summary of the tables that are shown here.

E.1 Gene Ontology

Gene Ontology analysis was preformed for 6 conditions of interest. The GREAT algorithm was used to associate peaks with genes. Once associated, David was used for GO analysis and Revigo was used to visualize the results and to remove the redundant categories. The results can be seen in Tables 17 to 22.

Table 17: The set of all critical GO labels for the ecfc context.

term	ID description	frequency	log10 p-value
GO:0001568	blood vessel development	0.037%	-13.4631
GO:0003001	generation of a signal involved in cell-cell signaling	0.159%	-1.5896
GO:0006468	protein phosphorylation	1.206%	-2.3357
GO:0007155	cell adhesion	0.564%	-8.9030
GO:0022610	biological adhesion	1.210%	-8.8714
GO:0040007	growth	0.083%	-2.5502
GO:0048511	rhythmic process	0.020%	-1.4967
GO:0043473	pigmentation	0.009%	-1.4936
GO:0016477	cell migration	0.069%	-7.3954
GO:0008218	bioluminescence	0.005%	-1.7884
GO:0008283	cell proliferation	0.105%	-3.1715
GO:0016265	death	0.279%	-1.5817
GO:0006914	autophagy	0.099%	-2.1798
GO:0017145	stem cell division	0.003%	-1.5911
GO:0034330	cell junction organization	0.011%	-4.8333

Continued on next page

Table 17 – continued from previous page			
term	ID description	frequency	log10 p-value
GO:0008037	cell recognition	0.015%	-1.6733
GO:0045321	leukocyte activation	0.040%	-2.0414
GO:0001775	cell activation	0.048%	-1.9853
GO:0007167	enzyme linked receptor protein signaling pathway	0.063%	-6.7828
GO:0006928	cellular component movement	0.492%	-6.0325
GO:0006793	phosphorus metabolic process	16.891%	-1.6028
GO:0050999	regulation of nitric-oxide synthase activity	0.001%	-2.2769
GO:0045428	regulation of nitric oxide biosynthetic process	0.002%	-1.3091
GO:0032886	regulation of microtubule-based process	0.011%	-1.5325
GO:0007050	cell cycle arrest	0.011%	-2.3531
GO:0040008	regulation of growth	0.042%	-2.6826
GO:0051674	localization of cell	0.402%	-6.3523
GO:0042127	regulation of cell proliferation	0.080%	-6.6105
GO:0006875	cellular metal ion homeostasis	0.156%	-1.8870
GO:0015718	monocarboxylic acid transport	0.224%	-2.4297
GO:0043067	regulation of programmed cell death	0.134%	-5.7249
GO:0007584	response to nutrient	0.004%	-3.9082
GO:0016192	vesicle-mediated transport	0.381%	-2.3558
GO:0031328	positive regulation of cellular biosynthetic process	0.255%	-3.7986
GO:0042060	wound healing	0.025%	-5.6539
GO:0010324	membrane invagination	0.002%	-2.6286
GO:0009395	phospholipid catabolic process	0.023%	-1.6607
GO:0051606	detection of stimulus	0.147%	-2.1302
GO:0046903	secretion	0.627%	-1.3288
GO:0034329	cell junction assembly	0.009%	-3.7764
GO:0009719	response to endogenous stimulus	0.113%	-2.0272
GO:0051924	regulation of calcium ion transport	0.008%	-1.8569
GO:0019935	cyclic-nucleotide-mediated signaling	0.003%	-1.4426
GO:0009628	response to abiotic stimulus	0.312%	-1.3360
GO:0001666	response to hypoxia	0.019%	-4.6793
GO:0048489	synaptic vesicle transport	0.005%	-2.5807
GO:0051174	regulation of phosphorus metabolic process	0.392%	-1.4747
GO:0045792	negative regulation of cell size	0.001%	-1.7792

Table 18: The set of all critical GO labels for the erythroid context.

term	ID description	frequency	log10 p-value
GO:0002520	immune system development	1.845%	-4.1076
GO:0006512	ubiquitin cycle	0.381%	-1.2812
GO:0022403	cell cycle phase	0.007%	-1.3779
GO:0032236	positive regulation of calcium ion transport via store-operated calcium channel activity	0.381%	-1.1607
GO:0042592	homeostatic process	3.235%	-3.6508
GO:0045767	regulation of anti-apoptosis	0.381%	-1.1726
GO:0045768	positive regulation of anti-apoptosis	0.381%	-1.2806
GO:0046777	protein autophosphorylation	0.414%	-4.1341
GO:0048678	response to axon injury	0.155%	-2.2534
GO:0061024	membrane organization	1.629%	-2.4273
GO:0015804	neutral amino acid transport	0.080%	-3.5202
GO:0008283	cell proliferation	4.075%	-2.1529
GO:0016265	death	4.501%	-1.6654
GO:0010269	response to selenium ion	0.021%	-2.2417
GO:0006783	heme biosynthetic process	0.059%	-3.5189
GO:0042743	hydrogen peroxide metabolic process	0.089%	-1.2051
GO:0016055	Wnt signaling pathway	0.887%	-2.9051
GO:0070482	response to oxygen levels	0.678%	-1.6898

Continued on next page

Table 18 – continued from previous page			
term	ID description	frequency	log10 p-value
GO:0007163	establishment or maintenance of cell polarity	0.358%	-1.2790
GO:0044092	negative regulation of molecular function	1.921%	-1.7684
GO:0007049	cell cycle	3.703%	-1.9539
GO:0030029	actin filament-based process	1.610%	-1.3300
GO:0051301	cell division	1.708%	-1.0198
GO:0001775	cell activation	2.008%	-1.1646
GO:0022402	cell cycle process	2.580%	-1.8561
GO:0006928	cellular component movement	3.830%	-1.0512
GO:0006793	phosphorus metabolic process	17.457%	-4.7261
GO:0042440	pigment metabolic process	0.165%	-1.9133
GO:0048662	negative regulation of smooth muscle cell proliferation	0.075%	-2.4935
GO:0019751	polyol metabolic process	0.210%	-1.8508
GO:0010608	posttranscriptional regulation of gene expression	0.840%	-2.1643
GO:0070085	glycosylation	0.734%	-1.4129
GO:0002440	production of molecular mediator of immune response	0.342%	-1.2135
GO:0030155	regulation of cell adhesion	0.746%	-2.1067
GO:0020027	hemoglobin metabolic process	0.036%	-1.6682
GO:0019935	cyclic-nucleotide-mediated signaling	0.089%	-2.7317
GO:0006897	endocytosis	1.430%	-1.6720
GO:0046488	phosphatidylinositol metabolic process	0.472%	-2.0614
GO:0033013	tetrapyrrole metabolic process	0.173%	-1.8243
GO:0006833	water transport	0.117%	-1.5120
GO:0042044	fluid transport	0.132%	-1.4032
GO:0030518	intracellular steroid hormone receptor signaling pathway	0.228%	-1.1656
GO:0006022	aminoglycan metabolic process	0.332%	-1.1580
GO:0006836	neurotransmitter transport	0.450%	-1.0122
GO:0030522	intracellular receptor signaling pathway	0.684%	-1.3119
GO:0010324	membrane invagination	0.064%	-1.6720
GO:0009064	glutamine family amino acid metabolic process	0.229%	-1.1312
GO:0015791	polyol transport	0.013%	-1.6211
GO:0006282	regulation of DNA repair	0.090%	-1.4507
GO:0010035	response to inorganic substance	1.004%	-1.9020
GO:0007267	cell-cell signaling	2.522%	-1.4034
GO:0021697	cerebellar cortex formation	0.059%	-1.3779
GO:0050808	synapse organization	0.424%	-2.0442
GO:0009100	glycoprotein metabolic process	0.923%	-1.0233
GO:0007259	JAK-STAT cascade	0.253%	-2.4131
GO:0046677	response to antibiotic	0.108%	-1.5941
GO:0016192	vesicle-mediated transport	3.273%	-1.0800

Table 19: The set of all critical GO labels for a subset of the leukemia context that contains a ga Ebox.

term	ID description	frequency	log10 p-value
GO:0006333	chromatin assembly or disassembly	0.277%	-2.1207
GO:0007389	pattern specification process	1.143%	-1.5389
GO:0051098	regulation of binding	0.553%	-2.3203
GO:0016311	dephosphorylation	1.340%	-1.2362
GO:0010765	positive regulation of sodium ion transport	0.047%	-1.2755
GO:0019932	second-messenger-mediated signaling	0.427%	-1.3939
GO:0048872	homeostasis of number of cells	0.558%	-1.3307
GO:0006793	phosphorus metabolic process	17.457%	-1.6476
GO:0006323	DNA packaging	0.361%	-1.6713
GO:0007167	enzyme linked receptor protein signaling pathway	2.356%	-1.3658
GO:0001755	neural crest cell migration	0.109%	-1.4176

Continued on next page

Table 19 – continued from previous page			
term	ID description	frequency	log10 p-value
GO:0007185	transmembrane receptor protein tyrosine phosphatase signaling pathway	0.013%	-1.0412
GO:0019935	cyclic-nucleotide-mediated signaling	0.089%	-1.4988
GO:0051094	positive regulation of developmental process	1.985%	-1.4505
GO:0051099	positive regulation of binding	0.197%	-2.3892
GO:0006796	phosphate-containing compound metabolic process	17.353%	-1.6476
GO:0035556	intracellular signal transduction	6.704%	-1.9761
GO:0065004	protein-DNA complex assembly	0.318%	-2.1071
GO:0030036	actin cytoskeleton organization	1.465%	-1.0636
GO:0002429	immune response-activating cell surface receptor signaling pathway	0.565%	-1.0112
GO:0051090	regulation of sequence-specific DNA binding transcription factor activity	0.562%	-1.2919
GO:0043933	macromolecular complex subunit organization	5.276%	-1.2692
GO:0051091	positive regulation of sequence-specific DNA binding transcription factor activity	0.342%	-1.3230
GO:0030500	regulation of bone mineralization	0.121%	-1.2088
GO:0001667	ameboidal cell migration	0.556%	-1.0501
GO:0044093	positive regulation of molecular function	3.848%	-1.2600
GO:0065003	macromolecular complex assembly	2.983%	-1.7351

Table 20: The set of all critical GO labels for a subset of the leukemia context that contains a gg Ebox and a gata motif.

term	ID description	frequency	log10 p-value
GO:0009791	post-embryonic development	0.281%	-2.3978
GO:0043488	regulation of mRNA stability	0.087%	-3.0330
GO:0043627	response to estrogen	0.539%	-1.8488
GO:0006826	iron ion transport	0.189%	-1.5670
GO:0008211	glucocorticoid metabolic process	0.040%	-1.0692
GO:0030518	intracellular steroid hormone receptor signaling pathway	0.228%	-1.8315
GO:0043523	regulation of neuron apoptotic process	0.513%	-1.3387
GO:0045787	positive regulation of cell cycle	0.253%	-1.0414
GO:0042953	lipoprotein transport	0.042%	-1.0692
GO:0030522	intracellular receptor signaling pathway	0.684%	-1.5382
GO:0002063	chondrocyte development	0.063%	-1.2184
GO:0030325	adrenal gland development	0.070%	-1.0692
GO:0006357	regulation of transcription from RNA polymerase II promoter	3.684%	-1.9114
GO:0048608	reproductive structure development	1.170%	-1.5889
GO:0035295	tube development	1.620%	-1.3357
GO:0010608	posttranscriptional regulation of gene expression	0.840%	-1.4005
GO:0048598	embryonic morphogenesis	1.501%	-1.2545
GO:0010629	negative regulation of gene expression	2.446%	-1.8498
GO:0010628	positive regulation of gene expression	2.987%	-1.1352
GO:0051252	regulation of RNA metabolic process	10.497%	-2.0834
GO:0043487	regulation of RNA stability	0.093%	-2.9199

Table 21: The set of all critical GO labels for the hsc context.

term	ID description	frequency	log10 p-value
GO:0006334	nucleosome assembly	0.220%	-10.5105
			Continued on next page

Table 21 – continued from previous page			
term	ID description	frequency	log10 p-value
GO:0009611	response to wounding	2.056%	-2.7572
GO:0016265	death	4.501%	-3.4172
GO:0040012	regulation of locomotion	1.268%	-2.5089
GO:0051789	response to protein	0.381%	-1.3208
GO:0043691	reverse cholesterol transport	0.032%	-1.4589
GO:0050885	neuromuscular process controlling balance	0.136%	-2.2451
GO:0008283	cell proliferation	4.075%	-2.3649
GO:0045730	respiratory burst	0.035%	-1.0916
GO:0030029	actin filament-based process	1.610%	-2.7433
GO:0055072	iron ion homeostasis	0.273%	-2.1670
GO:0002825	regulation of T-helper 1 type immune response	0.039%	-1.8468
GO:0022406	membrane docking	0.141%	-1.0026
GO:0007626	locomotory behavior	0.515%	-1.6756
GO:0009991	response to extracellular stimulus	1.114%	-2.5187
GO:0007610	behavior	1.517%	-1.4547
GO:0031328	positive regulation of cellular biosynthetic process	3.341%	-5.7051
GO:0001775	cell activation	2.008%	-2.0549
GO:0009719	response to endogenous stimulus	3.718%	-1.2740
GO:0006928	cellular component movement	3.830%	-1.8519
GO:0018212	peptidyl-tyrosine modification	1.147%	-1.0563
GO:0006793	phosphorus metabolic process	17.457%	-2.0728
GO:0050865	regulation of cell activation	0.901%	-2.7731
GO:0051270	regulation of cellular component movement	1.289%	-3.8004
GO:0051056	regulation of small GTPase mediated signal transduction	1.620%	-4.9650
GO:0051094	positive regulation of developmental process	1.985%	-2.7804
GO:0042981	regulation of apoptotic process	3.144%	-4.9736
GO:0042127	regulation of cell proliferation	3.148%	-3.4177
GO:0007270	neuron-neuron synaptic transmission	0.336%	-1.0026
GO:0007229	integrin-mediated signaling pathway	0.418%	-3.9659
GO:0045445	myoblast differentiation	0.113%	-1.2135
GO:0009101	glycoprotein biosynthetic process	0.809%	-1.0730
GO:0030522	intracellular receptor signaling pathway	0.684%	-1.4330
GO:0016192	vesicle-mediated transport	3.273%	-1.6394
GO:0002684	positive regulation of immune system process	20.525%	-2.5512
GO:0001889	liver development	0.340%	-2.5423
GO:0031099	regeneration	0.350%	-1.6730
GO:0015837	amine transport	0.162%	-1.2890
GO:0007267	cell-cell signaling	2.522%	-1.3388
GO:0006816	calcium ion transport	0.973%	-1.4418
GO:0030030	cell projection organization	2.744%	-2.9822
GO:0051674	localization of cell	2.570%	-1.1040
GO:0051338	regulation of transferase activity	1.829%	-3.6543
GO:0007033	vacuole organization	0.138%	-1.4144
GO:0006997	nucleus organization	0.172%	-1.2623
GO:0031100	organ regeneration	0.148%	-1.2729

Table 22: The set of all critical GO labels for the leukemia context.

term	ID description	frequency	log10 p-value
GO:0014070	response to organic cyclic compound	2.318%	-5.8116
GO:0022610	biological adhesion	2.826%	-2.6339
GO:0040007	growth	2.160%	-2.2104
GO:0042981	regulation of apoptotic process	3.144%	-8.3359
GO:0045767	regulation of anti-apoptosis	0.381%	-2.1497
GO:0045768	positive regulation of anti-apoptosis	0.381%	-1.7389

Continued on next page

Table 22 – continued from previous page

term	ID description	frequency	log10 p-value
GO:0016265	death	4.501%	-2.5525
GO:0006470	protein dephosphorylation	0.717%	-1.9378
GO:0045793	positive regulation of cell size	0.030%	-4.0855
GO:0045927	positive regulation of growth	0.359%	-3.9649
GO:0050795	regulation of behavior	0.338%	-1.6116
GO:0009612	response to mechanical stimulus	0.422%	-1.4834
GO:0040012	regulation of locomotion	1.268%	-5.6100
GO:0007610	behavior	1.517%	-3.5660
GO:0051336	regulation of hydrolase activity	3.070%	-4.8517
GO:0009719	response to endogenous stimulus	3.718%	-1.7726
GO:0045321	leukocyte activation	1.587%	-7.6860
GO:0051301	cell division	1.708%	-1.5888
GO:0001775	cell activation	2.008%	-7.1456
GO:0006928	cellular component movement	3.830%	-4.3680
GO:0002712	regulation of B cell mediated immunity	0.113%	-1.6928
GO:0006022	aminoglycan metabolic process	0.332%	-1.7945
GO:0002521	leukocyte differentiation	1.015%	-7.0717
GO:0030155	regulation of cell adhesion	0.746%	-4.6845
GO:0010564	regulation of cell cycle process	1.104%	-2.0437
GO:0051270	regulation of cellular component movement	1.289%	-5.9701
GO:0007167	enzyme linked receptor protein signaling pathway	2.356%	-8.1115
GO:0042127	regulation of cell proliferation	3.148%	-5.5221
GO:0006349	regulation of gene expression by genetic imprinting	0.062%	-1.9989
GO:0010676	positive regulation of cellular carbohydrate metabolic process	0.104%	-1.4778
GO:0000018	regulation of DNA recombination	0.128%	-2.0872
GO:0033555	multicellular organismal response to stress	0.203%	-2.0724
GO:0006816	calcium ion transport	0.973%	-1.6394
GO:0006939	smooth muscle contraction	0.234%	-2.0259
GO:0002684	positive regulation of immune system process	20.525%	-3.1473
GO:0030030	cell projection organization	2.744%	-1.5997
GO:0016311	dephosphorylation	1.340%	-1.5391
GO:0048066	developmental pigmentation	0.124%	-2.0872
GO:0019216	regulation of lipid metabolic process	0.558%	-1.5160
GO:0051674	localization of cell	2.570%	-5.7789
GO:0042596	fear response	0.089%	-1.5691
GO:0030203	glycosaminoglycan metabolic process	0.296%	-1.5323
GO:0051240	positive regulation of multicellular organismal process	1.418%	-3.2237
GO:0009953	dorsal/ventral pattern formation	0.244%	-1.6640
GO:0060021	palate development	0.311%	-3.3367
GO:0009791	post-embryonic development	0.281%	-2.5725
GO:0007568	aging	0.730%	-1.5837
GO:0007267	cell-cell signaling	2.522%	-2.7920
GO:0051056	regulation of small GTPase mediated signal transduction	1.620%	-6.2837
GO:0010629	negative regulation of gene expression	2.446%	-3.1284

E.2 Motif DENovo

Homer was used for the denovo search of DNA motifs for the four contexts of interest, leukemia, erythroid, HSC, and ECFC. A shared central background was used for each analysis. The results of the analysis with for motifs of a fixed width ranging from 6 to 8 can be seen in Tables 23 to 26

Table 23: Summary of all motifs showing the statistics for the top motifs for the ECFC context

motif	order	pvalue	size
GGAAAC	1	35	6
CTTCCG	2	35	6
TTCCGG	3	35	6
CGGGAW	4	35	6
CTCGCA	5	35	6
GATAAG	6	23	6
TTGTTC	7	23	6
TGACCC	8	23	6
CGCCGC	9	23	6
CGGAARC	1	44	7
TGTTTTC	2	44	7
GATAACA	3	44	7
TGSTTCC	4	44	7
CGGGAAA	5	44	7
GCCAGCG	6	26	7
MCTTCCT	7	26	7
CCGGCTG	8	26	7
GACGCGA	9	26	7
CGGAAACA	1	42	8
TTCCGCGG	2	42	8
KGTCCTCG	3	42	8
GTGTTTGC	4	42	8
GCCTTTCC	5	42	8
AAACAATG	6	20	8
CRGATWTC	7	20	8
YCGTCCAG	8	20	8
CCAGGCCA	9	20	8
ACGTTATC	10	20	8
CATACCCT	11	16	8
TTTGACCC	12	16	8
TGACCCTT	13	16	8
GCAKCCGG	14	16	8
GAAATAAC	15	16	8
AGTGCGAC	16	16	8
AGGAAGCS	17	16	8
TTGTTCAG	18	16	8
CCGGCGGG	19	16	8

Table 24: Summary of all motifs showing the statistics for the top motifs for the Erythroid context

motif	order	pvalue	size
TTATCT	1	201	6
CTTATC	2	201	6
TATCGS	3	201	6
GGGCSS	4	201	6
GCAGGC	5	201	6
CTGGCC	6	189	6
AGGCGC	7	189	6

Continued on next page

Table 24 – continued from previous page

motif	order	pvalue	size
CRGCCC	8	189	6
CGCTCC	9	189	6
GCAGCY	10	189	6
TGGGCC	11	57	6
GCGCTC	12	57	6
RGGGMS	13	57	6
CASCMG	14	57	6
GGCCGT	15	57	6
DGGGYG	16	46	6
CTGCTC	17	46	6
CTTATCT	1	240	7
YTATCTG	2	240	7
CCCACCC	3	240	7
GCGCGCG	4	240	7
GGCTGTC	5	240	7
CCAGCTG	6	163	7
CGGGGTG	7	163	7
CGGGCCG	8	163	7
RGCGCCC	9	163	7
GAKGGGC	10	163	7
GAGGGGG	11	47	7
CTTGCCG	12	47	7
GGCTGCT	13	47	7
CGCTGTG	14	47	7
CAGCAGC	15	47	7
CAGGCCC	16	44	7
GCTCTGC	17	44	7
SSCTTATC	1	221	8
GGGKGGGB	2	221	8
GCCTTGTC	3	221	8
CGYTATCT	4	221	8
TATCTGCC	5	221	8
GAGAGGCG	6	171	8
CAGCHGCC	7	171	8
TGCWGCTG	8	171	8
GGCCCGCG	9	171	8
GGCTGTCA	10	171	8
CGGAGCCC	11	52	8
GGCCGGGC	12	52	8
RSBCCCCC	13	52	8
GGCTCCWC	14	52	8
GWGGGTGS	15	52	8
GCTGCGCG	16	44	8
GGGCCGGC	17	44	8
AGCCTKCT	18	44	8

Table 25: Summary of all motifs showing the statistics for the top motifs for the HSC context

motif	order	pvalue	size
CGCGCG	1	79	6
CCACAG	2	79	6
GGCCGG	3	79	6
CGRACG	4	79	6
CAGCCG	5	79	6

Continued on next page

Table 25 – continued from previous page

motif	order	pvalue	size
GCGCAG	6	76	6
CCGCGA	7	76	6
KYTYCG	8	76	6
GGGGGC	9	76	6
GCGGCG	10	76	6
AGGCCG	11	71	6
KGCGTG	12	71	6
CATCGG	13	71	6
CCGCTC	14	71	6
CGGCCG	15	71	6
SSCGCCG	1	81	7
SCGCGCG	2	81	7
CAGAYGS	3	81	7
NCGCCGD	4	81	7
GTGCCCA	5	81	7
NGGCGNH	6	76	7
KSACGGC	7	76	7
BTGTGGC	8	76	7
GGAGGCG	9	76	7
GCWGCGG	10	76	7
CTGGCTG	11	74	7
CTGGGTG	12	74	7
GGGGSCC	13	74	7
GGGCAGC	14	74	7
GGGGGGG	15	74	7
GCAGGGC	16	65	7
CSKTCSN	17	65	7
CSGCGGN	18	65	7
SWNCCGC	19	65	7
GAMTCCG	20	65	7
CGGTTCG	21	58	7
CCGCGCGC	1	86	8
CCCGTCCG	2	86	8
SCGCCGCS	3	86	8
CGGCGACG	4	86	8
SCGGCGGC	5	86	8
CACCGCGG	6	82	8
CGGCTCCG	7	82	8
SCSRCGWS	8	82	8
CVCCCAGG	9	82	8
TGGGGYTC	10	82	8
CTGCCTGC	11	81	8
CGCGCGGG	12	81	8
AGCCSCTS	13	81	8
CCGGCCTC	14	81	8
CAGCTGBY	15	81	8
SDGGCAGC	16	72	8
CTYCGRWK	17	72	8
CGCBCHCY	18	72	8
AKCTGCCC	19	72	8
AGCGGRGC	20	72	8
CSCYCCTC	21	70	8
CACVCCCC	22	70	8
TTCGATTC	23	70	8

Table 26: Summary of all motifs showing the statistics for the top motifs for the Leukemia context

motif	order	pvalue	size
ACCACA	1	172	6
CATCSG	2	172	6
AACCRC	3	172	6
TCGGCA	4	172	6
GCACCS	5	172	6
TGTGCG	6	124	6
GGTTTC	7	124	6
CGAGCG	8	124	6
TGTGGC	9	124	6
CGCCGC	10	124	6
GGCGGC	11	105	6
CACCTG	12	105	6
CGGGGG	13	105	6
AGACGC	14	105	6
CCCCAC	15	105	6
CGCGCG	16	62	6
AMCCRCA	1	197	7
RCCGCAN	2	197	7
CAGCAGC	3	197	7
GGGTGCG	4	197	7
GCASCTG	5	197	7
CATCGGC	6	157	7
GYGGKTT	7	157	7
GCGCCGC	8	157	7
CRGATGC	9	157	7
CCCGAGC	10	157	7
CSGTTKC	11	131	7
GTGGCAG	12	131	7
TCCGCCG	13	131	7
MGTGGTW	14	131	7
AACCCTC	15	131	7
GCGCGCG	16	122	7
GGGGGGG	17	122	7
GHTGTGGT	1	182	8
CAGMTGTK	2	182	8
TGCGCGCG	3	182	8
CACCCACA	4	182	8
AGAGCCGC	5	182	8
TCGGATGC	6	146	8
GGAAGTGG	7	146	8
GGGGGGGG	8	146	8
GMACCCRC	9	146	8
CGCAGGTG	10	146	8
GCGGCABC	11	79	8
VYACCCKC	12	79	8
CTTTGATG	13	79	8
CGGTGCAR	14	79	8
GCGCGCTT	15	79	8
GGTTTGCA	16	72	8
AACCGCCG	17	72	8
AAACCTCA	18	72	8
CGGTTTGC	19	72	8
GGGSCCCC	20	72	8

E.2.1 Assosiation of Motifs Using Stamp

The motifs were annotated using stamp. The top 3 annotations from both the Transfac Family database and the Jaspar Family data base are presented in Tables 27 to 34

Table 27: Summary of all motifs and their assoiated Transfac TFs for the ECFC context

motif	ann	escore	ann	escore	ann	escore
GGAAAC	ETS_c-Ets-1_M00743	3.7401E-06	STAT_STAT6_M00500	0.000025617	STAT_STAT1_M00492	0.000035334
CTTCCG	ETS_GABP_M00341	7.4795E-08	ETS_Elk-1_M00025	2.3885E-07	ETS_c-Ets-1_p54_M00032	2.4705E-07
TTCCGG	ETS_GABP_M00341	1.1003E-07	ETS_c-Ets-1_p54_M00032	2.3685E-07	ETS_Elk-1_M00025	2.451E-07
CGGGAW	fork_E2F-1-DP-1_M00736	1.9429E-06	AP2_ANT_M00501	3.0477E-06	STAT_STAT1_M00224	0.00021735
CTCGCA	NK-2-Nkx_Hmx3_M00433	0.00016965	bHLH_AhR_M00778	0.00030053	CH_Lyf-1_M00141	0.0011085
GATAAG	CH_Evi-1_M00080	2.5985E-08	CC_GATA-2_M00348	3.5983E-08	CC_GATA-1_M00347	4.1107E-08
TTGTTC	HMG_STE11_M01005	6.9859E-08	HMG_SRY_M00148	0.000010951	CC_GR_M00921	0.000016588
TGACCC	CC_CF1_M00111	9.7384E-09	CC_CF1_M00112	1.2559E-08	CC_RORalpha_M01138	3.3597E-07
CGCCGC	AP2_ERF2_M01057	1.1394E-06	trp_Adf-1_M00171	6.3677E-06	trp_Adf-1_M00923	0.000097016
CGGAARC	ETS_c-Ets-1_M00743	8.5378E-08	STAT_STAT1_M00492	2.3801E-06	ETS_c-Ets-1_p54_M00032	3.2398E-06
TGTTTTC	CC_GATA-1_M00346	0.000011937	fork_FOXO3A_M01137	0.000014038	fork_FOXO4_M00476	0.000015926
GATAACA	CC_GATA-1_M00346	1.9747E-10	CC_GATA-1_M00347	1.5559E-07	CC_GATA-2_M00349	7.8771E-07
TGSTTCC	ETS_c-Ets-1_M00743	0.000001652	ETS_Tel-2_M00678	0.0001156	ETS_c-Ets-2_M00340	0.00021283
CGGGAAA	fork_E2F-1_M00940	9.4501E-07	fork_E2F-1-DP-1_M00736	1.2149E-06	STAT_STAT_M00259	2.4041E-06
GCCAGCG	bZIP_HAC1_M00730	2.0737E-08	bHLH_HEB_M00698	0.00025717	Grainyhead_LBP-1_M00644	0.00050991
MCTTCCT	ETS_Tel-2_M00678	6.3828E-09	ETS_Ets_M00971	4.8004E-08	ETS_PEA3_M00655	9.7202E-08
CCGGCTG	bHLH-ZIP_AP-4_M00176	4.3281E-06	bHLH_HEB_M00698	4.7796E-06	Grainyhead_LBP-1_M00644	0.000044581
GACGCGA	bHLH_StuAp_M00263	0.000015044	fork_E2F_M00425	0.000070618	fork_E2F_M00050	0.000087561
CGGAAACA	CH+homeo_AREB6_M00415	0.000021491	ETS_c-Ets-2_M00340	0.000062498	ETS_c-Ets-1_M00743	0.00012206
TTCCGCGG	C6_PDR3_M00752	5.1775E-11	fork_E2F-1_M00940	0.000018451	fork_E2F-1_M00938	0.000048552
KGTCCTCG	CH_Ttk_M00009	0.000058635	ETS_PU.1_M00658	0.00045168	CC_T3R_M00963	0.0010703
GTGTTTGC	bZIP_DBP_M00624	7.0872E-07	fork_HFH4_M00742	1.0149E-06	CH_Sn_M00044	0.000015669
GCCTTTCC	Rel_NF-kappaB_M00208	1.9983E-06	CH_BLIMP1_M01066	0.00001143	Rel_NF-kappaB_M00194	0.000036262
AAACAATG	HMG_Mat1-Mc_M00276	6.3061E-11	HMG_ROX1_M00728	1.4845E-08	HMG_SOX9_M00410	4.3818E-07
CRGATWTC	Rel_c-Rel_M00053	0.000010824	STAT_STAT1_M00492	0.000040183	STAT_STAT5A_M00460	0.00019919
YCGTCCAG	CC_HNF4,_M00967	0.00028142	CH_Ttk_M00009	0.00035351	CC_HNF4_M01032	0.0016514
CCAGGCCA	paired-homeo_Pax-6_M00979	3.5954E-06	CC_PPARalpha-RXRalpha_M00242	0.00013414	bHSH_AP-2_M00915	0.00037604
ACGTTATC	CC_GATA-1_M00346	4.2313E-08	CC_GATA-2_M00348	3.3231E-06	CC_GATA_M00789	5.5402E-06
CATACCCT	bZIP_Tax-CREB_M00115	9.3616E-07	CH_SZF1-1_M01109	2.6189E-06	trp_AtMYB-84_M00970	0.00048524
TTTGACCC	CC_PPAR_M00763	0.00007419	CC_COUP_M00765	0.000076545	CC_HNF4_M00764	0.000082087
TGACCCTT	CC_PPAR,_M00762	8.665E-08	CC_COUPTF_M01036	6.826E-07	CC_CF1_M00112	0.000001185
GCAKCCGG	ETS_c-Ets-1_p54_M00032	3.1171E-07	ETS_c-Ets-1_M01078	0.000015532	bHLH_TAL1_M00993	0.00007608
GAAATAAC	fork_FOXP3_M00992	3.5095E-07	MADS_MEF-2_M00006	3.8655E-07	MADS_MEF-2_M00406	4.1427E-07
AGTGCGAC	NK-2-Nkx_Hmx3_M00433	1.7681E-06	paired-homeo_Pax-3_M00360	0.000055944	fork_E2F_M00425	0.0035694
AGGAAGCS	ETS_c-Ets-1_M00743	2.0445E-09	ETS_Ets_M00971	2.618E-07	ETS_PU.1_M00658	3.3072E-07
TTGTTCAG	HMG_STE11_M01005	0.000028419	POU_SGF-3_M00662	0.000040318	CC_GR_M00921	0.00005494
CCGGCGGG	bHSH_AP-2_M00189	1.1841E-09	CH_Sp1_M00931	0.000011515	CH_Sp1_M00008	0.00001621

Table 28: Summary of all motifs and their assoiated Jaspar TFs for the ECFC context

motif	ann	escore	ann	escore	ann	escore
GGAAAC	ETS_Eip74EF	0.00005816	ETS_ELK1	0.00013804	ETS_ELF5	0.00018228
CTTCCG	ETS_Eip74EF	3.6762E-09	ETS_ELK4	2.284E-08	ETS_GABPA	3.3733E-08
TTCCGG	ETS_Eip74EF	6.1718E-09	ETS_ELK4	1.8417E-08	ETS_ELK1	5.588E-08
CGGGAW	REL_NFKB1	0.0014094	ETS_ETS1	0.0016783	ZN-FINGER_C2H2_MZF1_1-4	0.0019255
CTCGCA	bHLH_Arnt-Ahr	0.0037532	bZIP_Cebpa	0.0052349	bHLH_Hand1-Tcfe2a	0.0081362
GATAAG	ZN-FINGER_C2H2_Evi1	7.7354E-07	bZIP_PEND	0.000079552	ZN-FINGER_GATA_GATA2	0.0007351
TTGTTC	HMG_Sox5	6.3936E-06	HMG_SRY	0.000043172	HMG_SOX9	0.0000466
TGACCC	NUCLEAR_RECEPTOR_usp	9.078E-08	NUCLEAR_RECEPTOR_RXRA-VDR	5.5973E-07	NUCLEAR_RECEPTOR_PPARG-RXRA	2.3447E-06
CGCCGC	AP2_ABI4	0.00020467	bHLH_NHLH1	0.0013737	ZN-FINGER_C2H2_SP1	0.0083805
CGGAARC	ETS_Eip74EF	1.9282E-07	ETS_ELK1	3.4658E-06	ETS_ELK4	9.1121E-06
TGTTTTC	FORKHEAD_FOXD1	1.5729E-06	REL_Dl_2	0.00004018	ZN-FINGER_C2H2_Broad-complex_4	0.0000643
GATAACA	ZN-FINGER_C2H2_Evi1	0.000075783	bHLH-ZIP_Spz1	0.00029612	ZN-FINGER_GATA_GATA2	0.001940
TGSTTCC	ETS_SPI1	0.000043866	HOMEO_CAAT_TLX1-NFIC	0.00046823	ETS_ELK4	0.0008878
CGGGAAA	E2F_TDP_E2F1	0.000064038	REL_dl_1	0.00015283	IPT_TIG_domain_Su_H	0.0004935
GCCAGCG	ZN-FINGER_C2H2_REST	0.0015088	bHLH_Hand1-Tcfe2a	0.0015234	AP2_ABI4	0.0039717
MCTTCCT	ETS_ELF5	2.5662E-07	ETS_SPI1	0.000031292	ETS_Eip74EF	0.0000358
CCGGCTG	bHLH_NHLH1	0.00017269	bHLH_Myf	0.0011415	ETS_ELK4	0.0021735
GACGCGA	E2F_TDP_E2F1	0.00051635	bHLH_Arnt-Ahr	0.0040894	bZIP_bZIP910	0.010871
CGGAAACA	FORKHEAD_FOXD1	0.00010245	REL_dl_1	0.00038782	NUCLEAR_RECEPTOR_NR3C1	0.00043957
TTCCGCGG	ETS_SPIB	0.000066198	ETS_Eip74EF	0.00033324	E2F_TDP_E2F1	0.000757
KGTCCTCG	ETS_SPIB	0.00071706	NUCLEAR_RECEPTOR_RXRA-VDR	0.01005	ZN-FINGER_C2H2_MZF1_1-4	0.029275
GTGTTTGC	FORKHEAD_FOXD1	0.00001894	FORKHEAD_Foxd3	0.00030299	FORKHEAD_FOXF2	0.0010695
GCCTTTCC	ZN-FINGER_C2H2_Klf4	0.000037506	REL_REL	0.0001134	REL_NF-kappaB	0.00036535
AAACAATG	HMG_SOX9	2.0538E-08	HMG_SRY	9.8457E-08	FORKHEAD_Foxq1	5.9407E-07
CRGATWTC	REL_REL	0.000026519	REL_Dl_2	0.000061389	REL_RELA	0.00056325
YCGTCCAG	ZN-FINGER_C2H2_MIZF	0.00082444	ZN-FINGER_C2H2_ZNF354C	0.0051164	ZN-FINGER_C2H2_REST	0.0089213
CCAGGCCA	HOMEO_CAAT_TLX1-NFIC	0.00019558	bZIP_bZIP911	0.0030676	ZN-FINGER_C2H2_sna	0.0064282
ACGTTATC	ZN-FINGER_C2H2_Evi1	0.00046206	ZN-FINGER_GATA_GATA2	0.00086686	ZN-FINGER_GATA_GATA3	0.0018411
CATACCCT	bHLH-ZIP_Spz1	0.0000231	P53_TP53	0.0029037	TEA_TEAD1	0.016119
TTTGACCC	NUCLEAR_RECEPTOR_NR1H2-RXRA	0.00024877	NUCLEAR_RECEPTOR_PPARG-RXRA	0.00056763	bZIP_TCF11-MafG	0.00064548
TGACCCTT	NUCLEAR_RECEPTOR_NR1H2-RXRA	1.0971E-06	NUCLEAR_RECEPTOR_NR2F1	2.3195E-06	NUCLEAR_RECEPTOR_PPARG-RXRA	5.6374E-06
GCAKCCGG	ETS_ELK4	1.9604E-06	ETS_Eip74EF	0.000010556	ETS_GABPA	0.00013724
GAAATAAC	FORKHEAD_FOXI1	0.00033707	REL_dl_1	0.0009202	HOMEO-ZIP_Athb-1	0.0012072
AGTGCGAC	HMG_HMG-1	0.0032046	bZIP_Cebpa	0.0070351	bZIP_Fos	0.0088325
AGGAAGCS	ETS_SPI1	0.000040388	ETS_ELF5	0.0001122	ETS_Eip74EF	0.00018605
TTGTTCAG	HMG_Sox5	0.00017007	ZN-FINGER_C2H2_Broad-complex_4	0.0002757	FORKHEAD_Foxq1	0.00048359
CCGGCGGG	ZN-FINGER_C2H2_SP1	0.00006148	AP2_TFAP2A	0.0054837	bHLH_Arnt-Ahr	0.020793

Table 29: Summary of all motifs and their assoiated Transfac TFs for the Erythroid context

motif	ann	escore	ann	escore	ann	escore
TTATCT	CH_Evi-1_M00011	8.8827E-09	CC_GATA-1_M00347	5.6036E-08	CH_Evi-1_M00079	6.8794E-08
CTTATC	CH_Evi-1_M00080	2.5985E-08	CC_GATA-2_M00348	3.5983E-08	CC_GATA-1_M00347	4.1107E-08
TATCGS	C6_HAP1_M00305	0.000034241	BED_DREF_M00488	0.00030076	CC_GATA-2_M00349	0.00065789
GGGCSS	CH_Sp1_M00931	0.000014739	AP2_ABI4_M00958	0.000018407	CH_Sp1_M00933	0.000020374
GCAGGC	CH_Ttk_M00009	0.000026821	bHLH_HEB_M00698	0.000033979	CH+BTB-POZ_KAISO_M01119	0.000099616
CTGGCC	CH_PacC_M00247	0.00004912	bZIP_bZIP911_M00358	0.00012895	CC_HNF4_M01033	0.0002497
AGGCGC	fork_E2F_M00050	0.00019461	fork_E2F-1_M00939	0.00035088	fork_E2F_M00024	0.00035852
CRGCCC	CH_MTF-1_M00650	0.00037763	CH_PacC_M00247	0.00069814	SMAD_SMAD4_M00733	0.0015795
CGCTCC	CH_MAZ_M00649	0.000005377	paired_Pax-5_M00144	0.00043681	bHLH-ZIP_Spz1_M00446	0.00050005
GCAGCY	Grainyhead_LBP-1_M00644	2.3438E-08	bHLH-ZIP_AP-4_M00927	7.9833E-06	bHLH_myogenin_M00712	0.000013516
TGGGCC	bHLH_PCF2_M00948	2.5071E-08	homeo_Nanog_M01123	0.000018054	CH_Egr-3_M00245	0.00022823
GCGCTC	CH_GAGA_M00723	1.7415E-06	CH_NRSE_M00325	0.000041326	trp_CDC5_M00361	0.00020433
RGGGMS	CH_Sp1_M00933	0.00014402	CH_Sp1_M00196	0.00049342	CH_MAZ_M00649	0.0013768
CASCMG	bZIP_LMAF_M01139	0.00001736	CH_SZF1-1_M01109	0.000021905	bHLH_HEB_M00698	0.00043294
GGCCGT	CH_MTF-1_M00650	1.6603E-06	CH_PacC_M00247	5.6475E-06	trp_GAMYB_M00345	0.00026149
DGGGYG	CH_Zic2_M00449	0.000020647	CH_Zic1_M00448	0.000026563	CC_GBF_M00633	0.000038277
CTGCTC	bZIP_LMAF_M01139	0.000080298	CC_AR_M00962	0.00013197	bHLH_TAL1_M00993	0.00053532
CTTATCT	CC_GATA-2_M00348	2.1594E-10	CC_GATA-1_M00347	4.1304E-10	CC_GATA-2_M00349	6.5133E-10
YTATCTG	CC_GATA-1_M00346	8.6381E-08	LIM_Lmo2_M00278	5.6578E-07	CC_GATA-2_M00349	5.7137E-07
CCCACCC	homeo-PHD_Alfin1_M00479	2.7115E-07	CH_Zic2_M00449	4.0584E-07	CH_Tra-1_M01049	4.3568E-07
GCGCGCG	fork_E2F_M00803	3.7308E-06	CH+BTB-POZ_ZF5_M00333	0.000016799	CH+BTB-POZ_ZF5_M00716	0.000063791
GGCTGTC	SMAD_SMAD3_M00701	7.631E-07	homeo_TGIF_M00418	2.0122E-06	CH_NRSF_M01028	0.000075538
CCAGCTG	bHLH_HEB_M00698	2.1143E-11	bHLH-ZIP_AP-4_M00176	3.9286E-10	bHLH_TAL1_M00993	8.6725E-09
CGGGGTG	bHLH-ZIP_SREBP-1_M00221	3.0151E-06	CC_GBF_M00633	3.5473E-06	bHLH-ZIP_SREBP_M00776	0.000045848
CGGGCCG	CH_MTF-1_M00650	0.000029969	bHLH_PCF2_M00948	0.00089079	bHLH_AhR-Arnt_M00235	0.0012996
RGCGCCC	AP2_ERF2_M01057	0.0011789	CH_Sp1_M00933	0.0074466	CH_Sp1_M00931	0.0075081
GAKGGGC	CH_STRE_M00308	0.0001826	CH_ZNF219_M01122	0.00122	CH_MAZ_M00649	0.0012654
GAGGGGG	CH_ZNF219_M01122	1.6946E-08	CH_MZF1_M00084	1.1842E-07	CH+BTB-POZ_MAZR_M00491	2.4576E-06
CTTGCCG	NF1_NF-1_M00806	0.00048111	fork_E2F-1-DP-1_M00736	0.00084485	fork_E2F-1_M00430	0.001475
GGCTGCT	bHLH_TAL1_M00993	0.000047167	CH_YY1_M00069	0.000053076	CH_SZF1-1_M01109	0.000081369
CGCTGTG	trp_CDC5_M00361	0.000014737	CH_NRSF_M00256	0.00018023	CH_NRSF_M01028	0.00032254
CAGCAGC	bZIP_LMAF_M01139	4.6585E-07	bHLH-ZIP_AP-4_M00927	6.2527E-07	bHLH_HEB_M00698	7.1655E-07
CAGGCCC	bHLH_PCF2_M00948	0.0016313	bHLH_AhR-Arnt_M00237	0.0044436	P53_p53_M00034	0.0068643
GCTCTGC	homeo_Eve_M00629	1.0968E-07	CH_Ttk_M00009	0.000078312	trp_CDC5_M00361	0.000087013
SSCTTATC	CC_GATA-2_M00348	1.7698E-08	CC_GATA-1_M00347	2.1762E-08	CH_Evi-1_M00080	3.7086E-08
GGGKGGGB	CH_RREB-1_M00257	8.0309E-07	CH+BTB-POZ_MAZR_M00491	1.4274E-06	homeo-PHD_Alfin1_M00479	1.8435E-06
GCCTTGTC	P53_p53_M00761	0.00013871	CC_SF1_M00727	0.00017082	CH_Evi-1_M00079	0.00019699
CGYTATCT	LIM_Lmo2_M00278	4.2822E-09	CC_GATA-1_M00346	5.2755E-06	CH_Evi-1_M00011	0.000018399
TATCTGCC	CH_SZF1-1_M01109	9.8649E-06	bHLH_TAL1_M00993	0.000023664	bHLH_E2A_M00973	0.000056448
GAGAGGCG	CH_Sry-beta_M00666	0.000068678	CH+BTB-POZ_GZF1_M01069	0.00080994	CH_MAZ_M00649	0.001668

Continued on next page

Table 29 – continued from previous page						
motif	ann	escore	ann	escore	ann	escore
CAGCHGCC	bHLH-ZIP_AP-4_M00927	1.7415E-07	bHLH_myogenin_M00712	5.8782E-07	bHLH_HEB_M00698	2.6365E-06
TGCWGCTG	bHLH_HEB_M00698	7.0621E-07	bHLH-ZIP_AP-4_M00927	8.3474E-07	bHLH_myogenin_M00712	6.8565E-06
GGCCCGCG	bHLH_PCF2_M00948	0.000038132	bHSH_AP-2_M00800	0.000041892	C6_PDR3_M00752	0.0002724
GGCTGTCA	homeo_TGIF_M00418	7.9201E-08	SMAD_SMAD3_M00701	6.6749E-06	homeo_MEIS1_M00419	0.00011918
CGGAGCCC	CH_ZID_M00085	0.00029674	Rel_NF-kappaB_M00051	0.00057957	NF1_myogenin_M00056	0.00067223
GGCCGGGC	P53_p53_M00034	0.000020768	CH_MTF-1_M00650	0.000022215	CH_Sp1_M00931	0.00011389
RSBCCCCC	CH_ADR1_M00048	0.00094887	CC_Churchill_M00986	0.0012619	CH_STRE_M00308	0.0016948
GGCTCCWC	CH_ZID_M00085	0.00036399	CH_Sry-beta_M00666	0.0018636	CH_NRSE_M00325	0.0022871
GWGGGTGS	CH_CACCC-binding_M00721	1.0586E-06	CH_Roaz_M00467	8.9368E-06	CH_GLI_M01037	0.000012143
GCTGCGCG	bHLH_HEN1_M00068	4.9441E-06	fork_E2F_M00803	8.1443E-06	bHLH_HEN1_M00058	0.000014639
GGGCCGGC	CH_MTF-1_M00650	0.000024602	CH_Sp1_M00933	0.000072497	CH_Sp1_M00931	0.00024824
AGCCTKCT	bHLH_HEB_M00698	0.00016403	CH+BTB-POZ_KAISO_M01119	0.00038495	CH_Ttk_M00009	0.0010335

Table 30: Summary of all motifs and their assoiated Jaspar TFs for the Erythroid context

motif	ann	escore	ann	escore	ann	escore
TTATCT	ZN-FINGER_C2H2_Evi1	4.1077E-07	ZN-FINGER_GATA_GATA3	0.00011711	ZN-FINGER_GATA_GATA2	0.00073555
CTTATC	ZN-FINGER_C2H2_Evi1	7.7354E-07	bZIP_PEND	0.000079552	ZN-FINGER_GATA_GATA2	0.00073555
TATCGS	ZN-FINGER_GATA_GATA2	0.00044097	ZN-FINGER_GATA_GATA3	0.003755	MADS_SRF	0.026339
GGGCSS	AP2_ABI4	0.00023102	TRP-CLUSTER_Myb	0.010151	E2F_TDP_E2F1	0.010413
GCAGGC	ZN-FINGER_C2H2_sna	0.0023548	bHLH_Arnt-Ahr	0.01382	bHLH_Hand1-Tcfe2a	0.014514
CTGGCC	bHLH_Hand1-Tcfe2a	0.00023934	bZIP_bZIP911	0.00069968	ZN-FINGER_C2H2_REST	0.0025674
AGGCGC	E2F_TDP_E2F1	0.00013555	AP2_ABI4	0.0051908	ZN-FINGER_C2H2_ZEB1	0.0067788
CRGCCC	bZIP_MAF_Mafb	0.019638	bHLH-ZIP_Spz1	0.059157	ZN-FINGER_C2H2_MIZF	0.061909
CGCTCC	AP2_ABI4	0.0065557	PAIRED_Pax5	0.010213	ZN-FINGER_DOF_Dof3	0.017089
GCAGCY	bHLH_NHLH1	7.8034E-06	bHLH_Myf	0.0001333	bZIP_MAF_Mafb	0.014755
TGGGCC	MADS_SRF	0.00031502	AP2_TFAP2A	0.021066	AP2_ABI4	0.021868
GCGCTC	ZN-FINGER_C2H2_REST	0.0042882	E2F_TDP_E2F1	0.0084727	AP2_ABI4	0.019172
RGGGMS	ZN-FINGER_C2H2_MZF1_1-4	0.0019882	AP2_ABI4	0.0055427	AP2_TFAP2A	0.0082547
CASCMG	bHLH_Myf	0.00029016	bZIP_CREB1	0.0027107	ZN-FINGER_C2H2_sna	0.0032659
GGCCGT	TRP-CLUSTER_Myb	0.000069645	ZN-FINGER_C2H2_MIZF	0.00088569	TRP-CLUSTER_GAMYB	0.0025491
DGGGYG	ZN-FINGER_C2H2_ZEB1	0.0016715	ZN-FINGER_C2H2_RREB1	0.0018153	ZN-FINGER_C2H2_sna	0.001981
CTGCTC	bHLH_Myf	0.008668	bZIP_PEND	0.012902	ETS_SPIB	0.017466
CTTATCT	ZN-FINGER_C2H2_Evi1	2.5494E-08	bZIP_PEND	0.00081074	ZN-FINGER_GATA_GATA2	0.0019402
YTATCTG	ZN-FINGER_GATA_GATA3	0.000054523	ZN-FINGER_C2H2_Evi1	0.00041178	ZN-FINGER_GATA_GATA2	0.0021447
CCCACCC	ZN-FINGER_C2H2_RREB1	0.00037455	ZN-FINGER_C2H2_SP1	0.00050944	PAIRED-HOMEO_Pax4	0.00066465
GCGCGCG	E2F_TDP_E2F1	0.0061421	ZN-FINGER_C2H2_MIZF	0.01469	bHLH_Arnt-Ahr	0.015693
GGCTGTC	ZN-FINGER_C2H2_REST	0.000031504	REL_REL	0.0051057	bHLH-ZIP_Spz1	0.0059884

Continued on next page

Table 30 – continued from previous page						
motif	ann	escore	ann	escore	ann	escore
CCAGCTG	bHLH_NHLH1	5.0655E-06	bHLH_TAL1-TCF3	0.000021333	bHLH_Myf	0.000042081
CGGGGTG	ZN-FINGER_C2H2_RREB1	0.00082729	ZN-FINGER_C2H2_SP1	0.0008974	bZIP_CREB1	0.00096089
CGGGCCG	AP2_TFAP2A	0.0089179	P53_TP53	0.010903	TRP-CLUSTER_Myb	0.014268
RGCGCCC	E2F_TDP_E2F1	0.02364	TRP-CLUSTER_GAMYB	0.051178	ZN-FINGER_C2H2_SP1	0.092334
GAKGGGC	MADS_SRF	0.0017215	ZN-FINGER_C2H2_MZF1_5-13	0.0033676	ZN-FINGER_C2H2_sna	0.0097553
GAGGGGG	ZN-FINGER_C2H2_MZF1_5-13	2.6287E-06	PAIRED-HOMEO_Pax4	0.00024515	ZN-FINGER_C2H2_SP1	0.0053646
CTTGCCG	E2F_TDP_E2F1	0.0035003	HOMEO_CAAT_TLX1-NFIC	0.010481	ETS_ELK4	0.023488
GGCTGCT	bHLH_Myf	0.00062534	bHLH_NHLH1	0.0019988	bHLH-ZIP_Spz1	0.0041063
CGCTGTG	ZN-FINGER_C2H2_REST	0.0005239	T-BOX_T	0.0024072	bHLH-ZIP_Spz1	0.0030726
CAGCAGC	bHLH_Myf	1.9668E-07	bHLH_NHLH1	5.0939E-06	ZN-FINGER_C2H2_REST	0.00013081
CAGGCCC	AP2_ABI4	0.0080074	P53_TP53	0.0081687	ZN-FINGER_C2H2_sna	0.024204
GCTCTGC	ZN-FINGER_DOF_MNB1A	0.0065741	ZN-FINGER_DOF_PBF	0.0086285	ZN-FINGER_DOF_Dof2	0.012465
SSCTTATC	ZN-FINGER_C2H2_Evi1	1.0694E-06	bZIP_PEND	0.00010828	ZN-FINGER_GATA_GATA2	0.00083629
GGGKGGGB	ZN-FINGER_C2H2_RREB1	6.5175E-06	ZN-FINGER_C2H2_SP1	0.00013806	PAIRED-HOMEO_Pax4	0.00030459
GCCTTGTC	HMG_Sox17	0.00046791	NUCLEAR_RECEPTOR_RORA_1	0.0012229	ZN-FINGER_C2H2_Evi1	0.0041706
CGYTATCT	ZN-FINGER_GATA_GATA3	0.000092104	ZN-FINGER_C2H2_Evi1	0.00095603	ZN-FINGER_GATA_GATA2	0.0033292
TATCTGCC	ZN-FINGER_GATA_GATA2	0.0033777	TRP-CLUSTER_Myb	0.0037465	bHLH-ZIP_Spz1	0.007621
GAGAGGCG	ETS_SPIB	0.0022546	AP2_ABI4	0.020145	ZN-FINGER_C2H2_ZEB1	0.028856
CAGCHGCC	bHLH_Myf	1.0119E-06	bHLH_NHLH1	2.6238E-06	ZN-FINGER_C2H2_REST	0.00064046
TGCWGCTG	bHLH_NHLH1	0.000010368	bHLH_Myf	0.000031136	ZN-FINGER_C2H2_REST	0.00010223
GGCCCGCG	AP2_TFAP2A	0.00062474	P53_TP53	0.0027745	ZN-FINGER_C2H2_SP1	0.014706
GGCTGTCA	ZN-FINGER_C2H2_REST	0.00040677	bHLH_Hand1-Tcfe2a	0.0033945	bHLH-ZIP_Spz1	0.0041585
CGGAGCCC	AP2_ABI4	0.000008749	NUCLEAR_RECEPTOR_PPARG	0.0087086	NUCLEAR_RECEPTOR_usp	0.012531
GGCCGGGC	P53_TP53	0.000063323	AP2_TFAP2A	0.0019195	ZN-FINGER_C2H2_SP1	0.0063535
RSBCCCCC	ZN-FINGER_C2H2_MZF1_1-4	0.0015342	NUCLEAR_RECEPTOR_usp	0.014021	ZN-FINGER_C2H2_Macho-1	0.014978
GGCTCCWC	ZN-FINGER_C2H2_ZNF354C	0.00013289	TRP-CLUSTER_IRF1	0.024837	TEA_TEAD1	0.026681
GWGGGTGS	T-BOX_T	0.0031205	ZN-FINGER_C2H2_ZEB1	0.003935	ZN-FINGER_C2H2_sna	0.004163
GCTGCGCG	bHLH_NHLH1	0.00006608	E2F_TDP_E2F1	0.0145	AP2_ABI4	0.023382
GGGCCGGC	AP2_ABI4	0.00048209	ETS_Eip74EF	0.028951	TRP-CLUSTER_Myb	0.060314
AGCCTKCT	ZN-FINGER_C2H2_sna	0.021586	bZIP_PEND	0.034515	ZN-FINGER_C2H2_Staf	0.049724

Table 31: Summary of all motifs and their assoiated Transfac TFs for the HSC context

motif	ann	escore	ann	escore	ann	escore
CGCGCG	fork_E2F_M00803	5.9315E-07	fork_E2F-1_M00939	2.8617E-06	fork_E2F_M00918	3.8448E-06
CCACAG	runt_Osf2_M00731	0.000014071	runt_AML_M00769	0.000017244	runt_PEBP_M00984	0.000023844
GGCCGG	CH_PacC_M00247	0.00016561	CH_MTF-1_M00650	0.00065976	CH_YY1_M00069	0.0016319
CGRACG	C6_GAL4_M00049	0.0016899	trp_c-Myb_M00004	0.0022375	C6_CAT8_M00732	0.0024956

Continued on next page

Table 31 – continued from previous page

motif	ann	escore	ann	escore	ann	escore
CAGCCG	CH_SZF1-1_M01109	4.1456E-06	bHLH_HEB_M00698	0.00030985	bHLH-ZIP_AP-4_M00927	0.00053655
GCGCAG	bHLH_HEN1_M00068	5.9778E-06	bHLH_HEN1_M00058	0.000020673	bZIP_Nrf-1_M00652	0.000038903
CCGCGA	fork_E2F-1_M00430	5.974E-08	fork_E2F_M00427	7.2509E-07	fork_E2F_M00426	7.5094E-07
KYTYCG	fork_E2F-1_M00938	0.00074416	C6_HAP1_M00305	0.0024263	CH_NRSE_M00325	0.0044413
GGGGGC	CH_KROX_M00982	3.4392E-07	CH_Egr_M00807	6.0177E-07	CH+BTB-POZ_MAZR_M00491	8.7863E-07
GCGGCG	AP2_ERF2_M01057	3.9818E-06	bHSH_AP-2_M00189	0.000095779	trp_Adf-1_M00171	0.00026244
AGGCCG	CH_MTF-1_M00650	0.00026957	CH_PacC_M00247	0.00044504	C6_UAY_M00391	0.00097103
KGCGTG	NK-2-Nkx_Hmx3_M00433	1.2032E-06	bHLH_AhR_M00778	3.7786E-06	CH_Egr-1_M00243	5.7753E-06
CATCGG	AP2_ANT_M00501	0.000052625	CH+BTB-POZ_RP58_M00532	0.00023519	bHLH_TAL1_M00993	0.00048229
CCGCTC	paired_Pax-1_M00326	6.0272E-07	C6_FACB_M00390	0.000025006	CH_Sp1_M00931	0.00042581
CGGCCG	CH_MTF-1_M00650	0.00026957	CH_PacC_M00247	0.00044504	C6_UAY_M00391	0.00097103
SSCGCCG	bHSH_AP-2_M00189	0.00013346	trp_Adf-1_M00171	0.00013809	AP2_ERF2_M01057	0.00045334
SCGCGCG	fork_E2F_M00803	3.7308E-06	CH+BTB-POZ_ZF5_M00716	0.000063791	fork_E2F-1_M00939	0.0001035
CAGAYGS	bHLH_Hand1-E47_M00222	2.4332E-06	CH+BTB-POZ_RP58_M00532	0.000003488	bHLH_Tal-1beta-ITF-2_M00070	0.000013404
NCGCCGD	AP2_ERF2_M01057	0.0013207	CH_PacC_M00247	0.017707	CH_MTF-1_M00650	0.017923
GTGCCCA	bHLH_PCF2_M00948	0.000016255	CC_HNF4_M01033	0.000046524	MADS_SRF_M00152	0.00005386
NGGCGNH	AP2_ERF2_M01057	0.0005111	fork_E2F-1_M00430	0.004746	CH_Sp1_M00931	0.0050055
KSACGGC	homeo_TGIF_M00418	0.00010099	bZIP_ATF2-c-Jun_M00041	0.00055056	CH_MTF-1_M00650	0.0016321
BTGTGGC	runt_Osf2_M00731	7.1589E-06	bZIP_TRAB1_M00507	0.000027736	runt_PEBP_M00984	0.000062403
GGAGGCG	CH+BTB-POZ_ZF5_M00333	0.0000423	CH_KROX_M00982	0.00007779	CH_MAZ_M00649	0.000082087
GCWGCGG	bHLH_HEN1_M00058	5.7452E-06	bHLH-ZIP_AP-4_M00005	0.0000154	Grainyhead_LBP-1_M00644	0.000081595
CTGGCTG	SMAD_SMAD4_M00733	9.3126E-06	bHLH_Hand1-E47_M00222	0.00096494	CH_YY1_M00069	0.0012252
CTGGGTG	CH_GLI_M01037	5.0216E-06	CC_GBF_M00633	0.000022022	Grainyhead_CP2_M00072	0.000061162
GGGGSCC	bHLH_PCF2_M00948	4.5559E-06	CH_Egr_M00807	0.000049281	CH_KROX_M00982	0.000086155
GGGCAGC	CC_HNF4_M01033	0.000020465	bHLH_E47_M00002	0.000042592	SMAD_SMAD4_M00733	0.000071106
GGGGGGG	CH+BTB-POZ_MAZR_M00491	2.0524E-09	CH_ZNF219_M01122	2.6169E-08	CH_WT1_M01118	1.1732E-06
GCAGGGC	homeo_Eve_M00629	1.2814E-06	CH_Ttk_M00009	1.7627E-06	bHLH_E2A_M00973	0.00025777
CSKTCSN	CH_Ttk_M00009	0.013717	CH_ID1_M01021	0.015753	bHLH_PCF5_M00952	0.01728
CSGCGGN	AP2_ERF2_M01057	0.0004653	C6_PDR3_M00752	0.0013843	fork_E2F-1_M00430	0.0014791
SWNCCGC	CH_WT1_M01118	0.00051132	fork_E2F-1-DP-1_M00736	0.00061903	paired_Pax-5_M00144	0.0015125
GAMTCCG	bZIP_MAF_M00983	0.00089725	Rel_NF-kappaB_M00051	0.0043155	bZIP_AP-1_M00173	0.0043561
CGGTTCG	trp_GAMYB_M00345	0.00030914	C6_LEU3_M00306	0.0034354	trp_AtMYB-15_M00969	0.0044975
CCGCGCGC	fork_E2F_M00803	0.00028055	fork_E2F_M00425	0.0020243	CH+BTB-POZ_ZF5_M00716	0.0020488
CCCGTCCG	CH_Sp1_M00931	0.000055274	CH_Sp1_M00933	0.000070314	CH_Sp1_M00196	0.00042751
SCGCCGCS	AP2_ERF2_M01057	6.1284E-06	fork_E2F-1_M00428	0.00028808	fork_E2F-1_M00940	0.00061602
CGGCGACG	trp_Adf-1_M00171	1.2955E-06	AP2_ERF2_M01057	0.000070931	CH_Egr_M00807	0.00035245
SCGGCGGC	bHSH_AP-2_M00189	3.5946E-06	AP2_ERF2_M01057	0.000030113	AP2_ABI4_M00958	0.00021257
CACCGCGG	CH_NRSF_M01028	1.8598E-07	C6_PDR3_M00752	2.4218E-07	CH_NRSE_M00325	2.5869E-07
CGGCTCCG	NF1_myogenin_M00056	0.00037605	CH_ZID_M00085	0.00063058	paired_Pax-5_M00143	0.0010146
SCSRCGWS	bZIP_TAF-1_M00369	0.0015139	bZIP_ATF6_M00483	0.0017691	bZIP_CPRF-3_M00370	0.0021517
CVCCCAGG	CH_CACCC-binding_M00721	0.00001296	bHSH_AP-2_M00189	0.00018508	CH_NGFI-C_M00244	0.0002089
TGGGGYTC	bHLH-ZIP_SREBP-1_M00749	0.00001506	bHLH-ZIP_SREBP-1_M00221	0.00087246	CC_ARP-1_M00155	0.0010593
CTGCCTGC	trp_Adf-1_M00923	0.000013989	ETS_NERF1a_M00531	0.00004826	ETS_Tel-2_M00678	0.000062286
CGCGCGGG	CH+BTB-POZ_ZF5_M00716	1.0157E-06	fork_E2F_M00803	0.000026378	CH_Sp1_M00931	0.00015381

Continued on next page

Table 31 – continued from previous page						
motif	ann	escore	ann	escore	ann	escore
AGCCSCTS	CH_STRE_M00308	0.00064797	CH_MZF1_M00084	0.00070043	AP2_ERF2_M01057	0.0016235
CCGGCCTC	bHLH_E2A_M00804	0.00050971	CH_Sp1_M00933	0.00065541	CH_PacC_M00247	0.0018333
CAGCTGBY	bHLH_TAL1_M00993	7.8562E-09	bHLH_HEB_M00698	8.6559E-07	bHLH-ZIP_AP-4_M00927	9.097E-07
SDGGCAGC	bHLH_E2A_M00973	0.000027034	bHLH_E47_M00002	0.00013421	bHLH_E2A_M00804	0.00048823
CTYCGRWK	trp_IRF-7_M00453	0.0014779	trp_GAMYB_M00345	0.0021537	HMG_TCF-4_M00671	0.010518
CGCBCHCY	CH_Egr_M00807	0.00006851	CH_KROX_M00982	0.000074252	CH+BTB-POZ_ZF5_M00333	0.00017134
AKCTGCCC	Grainyhead_LBP-1_M00644	0.000058756	CC_HNF4_M01033	0.00011928	SMAD_SMAD4_M00733	0.00018611
AGCGGRGC	paired_Pax-1_M00326	1.7443E-06	paired_Pax-5_M00144	0.00010242	CH_Sp1_M00931	0.0022246
CSCYCCTC	CH_ZNF219_M01122	3.6554E-06	ETS_PU.1_M00658	0.000015473	CC_VDR_M00444	0.000041191
CACVCCCC	CC_GBF_M00633	3.7353E-06	CH_KROX_M00982	6.3351E-06	CH_Sp1_M00933	0.000021165
TTCGATTC	trp_IRF-7_M00453	4.9178E-06	homeo_CDP_M00095	0.000049453	homeo_Clox_M00103	0.00016256

Table 32: Summary of all motifs and their assoiated Jaspar TFs for the HSC context

motif	ann	escore	ann	escore	ann	escore
CGCGCG	E2F_TDP_E2F1	0.00047806	bHLH_Arnt-Ahr	0.0074301	bHLH-ZIP_Mycn	0.033379
CCACAG	RUNT_RUNX1	0.00034263	ZN-FINGER_C2H2_ZNF354C	0.0013748	IPT_TIG_domain_Su_H	0.0015852
GGCCGG	P53_TP53	0.0048621	AP2_TFAP2A	0.005652	ETS_Eip74EF	0.017714
CGRACG	ZN-FINGER_C2H2_MIZF	5.2426E-07	ETS_Eip74EF	0.0022379	ETS_GABPA	0.0036089
CAGCCG	bHLH_NHLH1	0.0029558	ZN-FINGER_C2H2_sna	0.004655	ETS_ELK4	0.0081453
GCGCAG	bHLH_NHLH1	0.0002393	E2F_TDP_E2F1	0.0075281	bZIP_Cebpa	0.009021
CCGCGA	E2F_TDP_E2F1	0.00029696	NUCLEAR_RECEPTOR_PPARG	0.01924	ZN-FINGER_C2H2_MIZF	0.023267
KYTYCG	ETS_Eip74EF	0.0054083	ETS_ELK4	0.012541	ETS_GABPA	0.01569
GGGGGC	AP2_ABI4	0.00031297	ZN-FINGER_C2H2_MZF1_5-13	0.00041015	ZN-FINGER_C2H2_Roaz	0.000669
GCGGCG	AP2_ABI4	0.00017038	E2F_TDP_E2F1	0.0088571	bHLH_NHLH1	0.0088754
AGGCCG	TRP-CLUSTER_Myb	0.0052604	NUCLEAR_RECEPTOR_HNF4A	0.030544	NUCLEAR_RECEPTOR_RORA_1	0.053468
KGCGTG	bHLH_Arnt-Ahr	1.2404E-06	PAIRED_Pax6	0.00026264	bHLH_Arnt	0.0013504
CATCGG	bHLH_TAL1-TCF3	0.00044982	NUCLEAR_RECEPTOR_RXRA-VDR	0.0016265	ZN-FINGER_GATA_Gata1	0.0021798
CCGCTC	ETS_SPIB	0.0035773	ZN-FINGER_C2H2_MZF1_5-13	0.010735	ZN-FINGER_C2H2_MIZF	0.016632
CGGCCG	ZN-FINGER_C2H2_MIZF	0.00066039	TRP-CLUSTER_Myb	0.0052604	bZIP_bZIP911	0.058909
SSCGCCG	AP2_ABI4	0.00062851	E2F_TDP_E2F1	0.0038079	ZN-FINGER_C2H2_SP1	0.030531
SCGCGCG	E2F_TDP_E2F1	0.0061421	bHLH_Arnt-Ahr	0.015693	bHLH-ZIP_MYC-MAX	0.017088
CAGAYGS	bHLH_TAL1-TCF3	0.000016868	bHLH_Hand1-Tcfe2a	0.00011051	ZN-FINGER_C2H2_sna	0.00036469
NCGCCGD	AP2_TFAP2A	0.06742	AP2_ABI4	0.11478	TRP-CLUSTER_Myb	0.1324
GTGCCCA	AP2_ABI4	0.00014701	IPT_TIG_domain_Su_H	0.0027509	MADS_SRF	0.0031194
NGGCGNH	E2F_TDP_E2F1	0.0077481	ZN-FINGER_C2H2_SP1	0.011696	AP2_ABI4	0.039131
KSACGGC	ZN-FINGER_C2H2_REST	0.0018404	bZIP_bZIP910	0.0051178	bZIP_TGA1a	0.0059164
BTGTGGC	RUNT_RUNX1	0.00017577	ZN-FINGER_C2H2_ZNF354C	0.0034777	E2F_TDP_E2F1	0.0048686

Continued on next page

Table 32 – continued from previous page

motif	ann	escore	ann	escore	ann	escore
GGAGGCG	AP2_ABI4	0.010492	ZN-FINGER_C2H2_ZEB1	0.013928	ZN-FINGER_C2H2_sna	0.015911
GCWGCGG	bHLH_NHLH1	0.000033982	bHLH_Myf	0.0043172	ETS_SPIB	0.021246
CTGGCTG	bHLH_Hand1-Tcfe2a	0.00011865	T-BOX_T	0.030798	bZIP_MAF_Mafb	0.037074
CTGGGTG	ZN-FINGER_C2H2_Roaz	0.00021919	T-BOX_T	0.00027963	ZN-FINGER_C2H2_SP1	0.0019754
GGGGSCC	ZN-FINGER_C2H2_Roaz	0.0021284	ZN-FINGER_C2H2_MZF1_1-4	0.003513	AP2_ABI4	0.0041396
GGGCAGC	AP2_ABI4	0.0001827	P53_TP53	0.0011666	ZN-FINGER_C2H2_REST	0.0043467
GGGGGGG	PAIRED-HOMEO_Pax4	4.8058E-06	ZN-FINGER_C2H2_SP1	0.000026725	ZN-FINGER_C2H2_RREB1	0.000062971
GCAGGGC	ZN-FINGER_C2H2_SP1	0.0019573	ZN-FINGER_C2H2_sna	0.0035832	AP2_TFAP2A	0.0039284
CSKTCSN	ZN-FINGER_C2H2_MIZF	0.017453	ZN-FINGER_GATA_GATA2	0.09234	REL_NF-kappaB	0.094256
CSGCGGN	ZN-FINGER_C2H2_MIZF	0.0049905	ZN-FINGER_C2H2_SP1	0.028958	ETS_SPIB	0.045273
SWNCCGC	ETS_ELK4	0.0066502	ETS_Eip74EF	0.0088987	bZIP_CREB1	0.0097023
GAMTCCG	TEA_TEAD1	0.0087493	ZN-FINGER_C2H2_ZNF354C	0.017818	TRP-CLUSTER_IRF1	0.028112
CGGTTCG	TRP-CLUSTER_GAMYB	0.0029943	TRP-CLUSTER_MYB.ph3	0.0051393	TRP-CLUSTER_IRF1	0.023971
CCGCGCGC	E2F_TDP_E2F1	0.024348	bHLH_Arnt-Ahr	0.0724	bHLH-ZIP_Mycn	0.1621
CCCGTCCG	ZN-FINGER_C2H2_MIZF	0.00001375	ETS_GABPA	0.0026394	ETS_Eip74EF	0.0027359
SCGCCGCS	AP2_ABI4	0.000079609	E2F_TDP_E2F1	0.0011004	bHLH_NHLH1	0.011176
CGGCGACG	NUCLEAR_RECEPTOR_ESR1	0.00090868	NUCLEAR_RECEPTOR_PPARG	0.0025121	AP2_ABI4	0.0052405
SCGGCGGC	ZN-FINGER_C2H2_SP1	0.0004616	TRP-CLUSTER_GAMYB	0.0085228	ZN-FINGER_C2H2_MZF1_5-13	0.011014
CACCGCGG	NUCLEAR_RECEPTOR_PPARG	0.0037357	NUCLEAR_RECEPTOR_ESR1	0.0038205	ZN-FINGER_C2H2_sna	0.006509
CGGCTCCG	AP2_ABI4	0.0018855	bHLH_NHLH1	0.0020359	PAIRED_Pax5	0.0025576
SCSRCGWS	bZIP_EMBP1	0.0050779	bHLH-ZIP_MYC-MAX	0.015103	bHLH-ZIP_MAX	0.021248
CVCCCAGG	ZN-FINGER_C2H2_Roaz	0.000032625	AP2_TFAP2A	0.0085447	ZN-FINGER_C2H2_sna	0.013982
TGGGGYTC	REL_Dl_2	0.0026521	REL_NFKB1	0.0047985	RUNT_RUNX1	0.0078562
CTGCCTGC	ETS_Eip74EF	0.0015245	ZN-FINGER_C2H2_sna	0.0077146	ETS_GABPA	0.026165
CGCGCGGG	P53_TP53	0.012024	ZN-FINGER_C2H2_SP1	0.022858	bHLH_Arnt-Ahr	0.027177
AGCCSCTS	bHLH_Myf	0.0010013	AP2_ABI4	0.016998	ZN-FINGER_C2H2_sna	0.028866
CCGGCCTC	ETS_Eip74EF	0.028077	bZIP_bZIP911	0.045187	ZN-FINGER_C2H2_MIZF	0.048911
CAGCTGBY	bHLH_Myf	6.6561E-08	bHLH_TAL1-TCF3	1.7642E-06	bHLH_NHLH1	0.000029748
SDGGCAGC	ETS_SPI1	0.0026357	ETS_Eip74EF	0.0057567	bHLH_Myf	0.0070248
CTYCGRWK	TRP-CLUSTER_GAMYB	0.022383	ETS_ELK4	0.026335	HMG_HMG-1	0.055307
CGCBCHCY	AP2_ABI4	0.0059942	ZN-FINGER_C2H2_RREB1	0.01749	PAIRED_Pax5	0.038525
AKCTGCCC	AP2_ABI4	0.0055416	bHLH_Myf	0.0067928	NUCLEAR_RECEPTOR_HNF4A	0.0088852
AGCGGRGC	PAIRED_Pax5	4.0765E-06	ETS_SPIB	0.000068728	ZN-FINGER_C2H2_MZF1_5-13	0.0058846
CSCYCCTC	ZN-FINGER_C2H2_Klf4	0.0012208	ZN-FINGER_C2H2_MZF1_5-13	0.0022071	ETS_SPIB	0.0045174
CACVCCCC	ZN-FINGER_C2H2_RREB1	0.000017675	bHLH_Arnt-Ahr	0.00049124	PAIRED-HOMEO_Pax4	0.0061281
TTCGATTC	ZN-FINGER_C2H2_Broad-complex_2	0.0012149	TRP-CLUSTER_IRF1	0.010721	HMG_HMG-1	0.011534

Table 33: Summary of all motifs and their assoiated Transfac TFs for the Leukemia context

motif	ann	escore	ann	escore	ann	escore
ACCACA	runt_AML1a_M00271	1.265E-07	runt_PEBP_M00984	1.639E-07	runt_AML_M00769	1.6894E-07
CATCSG	ETS_PEA3_M00655	0.000017533	ETS_c-Ets-1_p54_M00032	0.000061038	ETS_c-Ets-1_M01078	0.000310
AACCRC	runt_AML1a_M00271	1.1475E-06	runt_AML1_M00751	5.0256E-06	runt_core-binding_M00722	0.000030
TCGGCA	bZIP_LMAF_M01139	0.0010361	bHLH_E12_M00693	0.0023517	AP2_ERF2_M01057	0.0024
GCACCS	AP2_ABI4_M00958	0.000009768	CH_Sn_M00060	0.0000452	CH+homeo_AREB6_M00414	0.000054
TGTGCG	NK-2-Nkx_Hmx3_M00433	0.00028585	fork_E2F_M00803	0.001563	C6_PDR3_M00752	0.0018
GGTTTC	trp_IRF-1_M00062	8.3871E-07	CH+homeo_AREB6_M00415	3.8816E-06	Grainyhead_Grainyhead-Elf-1-NTF-1_M00951	0.000010
CGAGCG	WRKY_ZAP1_M00735	0.00011073	paired_Pax-1_M00326	0.00097	homeo_CDP_M00104	0.0011
TGTGGC	bZIP_TRAB1_M00507	6.2592E-06	runt_Osf2_M00731	7.9214E-06	bZIP_bZIP910_M00356	0.0001
CGCCGC	AP2_ERF2_M01057	2.9894E-07	trp_Adf-1_M00171	0.00038279	CH_ZBRK1_M01105	0.0011
GGCGGC	AP2_ERF2_M01057	8.9194E-09	trp_Adf-1_M00171	0.000037072	CH_ZBRK1_M01105	0.0007
CACCTG	bHLH_MyoD_M00184	2.4826E-08	bHLH_E12_M00693	3.3127E-08	CH+homeo_AREB6_M00414	5.4625E-
CGGGGG	CH_KROX_M00982	2.1131E-06	CH_MOVO-B_M01104	4.9448E-06	CC_Churchill_M00986	0.0000
AGACGC	CH+BTB-POZ_GZF1_M01069	3.7484E-06	bHLH_Hand1-E47_M00222	0.000012121	fork_Whn_M00332	0.0000
CCCCAC	bHLH-ZIP_SREBP-1_M00221	5.415E-08	homeo-PHD_Alfin1_M00479	4.3747E-07	bHLH-ZIP_SREBP-1_M00749	0.0000
CGCGCG	fork_E2F_M00803	6.4026E-07	CH+BTB-POZ_ZF5_M00716	0.000010376	fork_E2F_M00024	0.0000
AMCCRCA	runt_CBF_M01080	6.7039E-06	bHLH-ZIP_AP-4_M00005	0.000019361	runt_AML1a_M00271	0.0000
RCCGCAN	AP2_ERF2_M01057	0.0014886	C6_PDR3_M00752	0.0023273	fork_E2F-1_M00430	0.0041
CAGCAGC	bZIP_LMAF_M01139	3.4341E-07	bHLH_HEB_M00698	7.3165E-06	bHLH_myogenin_M00712	8.4436E-
GGGTGCG	CH+BTB-POZ_ZF5_M00333	0.00028724	T-box_TBX5_M01044	0.0002929	Grainyhead_LBP-1_M00644	0.0003
GCASCTG	bHLH_myogenin_M00712	8.8345E-10	bHLH-ZIP_AP-4_M00927	1.0324E-09	LIM_Lmo2_M00277	1.0662E-
CATCGGC	bHLH_TAL1_M00993	0.000049966	AP2_ANT_M00501	0.00033425	trp_v-Myb_M00227	0.0011
GYGGKTT	runt_core-binding_M00722	1.5263E-06	CH_Kr_M00021	0.000098494	runt_AML_M00769	0.0001
GCGCCGC	AP2_ABI4_M00958	1.5781E-06	AP2_ERF2_M01057	7.8731E-06	trp_Adf-1_M00171	0.00018
CRGATGC	ETS_c-Ets-1_M00743	4.9734E-06	ETS_PEA3_M00655	0.000012052	CH+BTB-POZ_RP58_M00532	0.0000
CCCGAGC	WRKY_ZAP1_M00735	2.2256E-06	AP2_ANT_M00501	0.00015556	P53_p53_M00034	0.0002
CSGTTKC	fork_E2F-1_M00938	0.00006314	trp_GAMYB_M00345	0.000085466	trp_MYB.Ph3_M00218	0.0002
GTGGCAG	bZIP_HBP-1b_M00697	0.000012097	bZIP_bZIP910_M00356	0.000071829	bZIP_ABF1_M00399	0.000
TCCGCCG	AP2_ERF2_M01057	2.7089E-06	bHSH_AP-2_M00189	9.2989E-06	fork_E2F-1_M00430	0.0000
MGTGGTW	runt_core-binding_M00722	4.4615E-07	runt_AML1_M00751	7.492E-07	runt_AML1a_M00271	1.1789E-
AACCCTC	CH_KR_M01089	4.9622E-06	CH_MAZ_M00649	0.00011145	trp_AtMYB-77_M00968	0.0002
GCGCGCG	CH+BTB-POZ_ZF5_M00333	4.2091E-07	fork_E2F_M00803	3.3325E-06	CH+BTB-POZ_ZF5_M00716	0.0000
GGGGGGG	CH+BTB-POZ_MAZR_M00491	2.5919E-09	CH_ZNF219_M01122	3.4993E-08	CH_MAZ_M00649	1.7648E-
GHTGTGGT	runt_AML1_M00751	1.4063E-07	runt_AML_M00769	2.0605E-07	runt_PEBP_M00984	2.2052E-
CAGMTGTK	CH+BTB-POZ_RP58_M00532	1.0005E-08	bHLH-ZIP_AP-4_M00176	3.4325E-06	bHLH-ZIP_AP-4_M00005	6.1481E-
TGCGCGCG	fork_E2F_M00803	4.8117E-06	fork_E2F_M00425	0.000012258	fork_E2F-1_M00939	0.0000
CACCCACA	CC_GBF_M00633	6.0121E-08	CH_NGFI-C_M00244	4.2046E-06	CH_GLI_M01037	5.9655E-
AGAGCCGC	AP2_ERF2_M01057	0.000064592	CH_Sry-beta_M00666	0.00027631	Grainyhead_LBP-1_M00644	0.00
TCGGATGC	ETS_PEA3_M00655	0.000011693	ETS_c-Ets-1_M00743	0.000014319	ETS_c-Ets-1_p54_M00032	0.0000
GGAAGTGG	ETS_NERF1a_M00531	2.5432E-08	ETS_GABP_M00341	3.5952E-08	ETS_c-Ets-1_M00339	1.4801E-
GGGGGGGG	CH+BTB-POZ_MAZR_M00491	1.658E-11	CH_ZNF219_M01122	2.8964E-10	CH_Sp1_M00931	2.3453E-
GMACCCRC	CH_Kr_M00021	0.000064732	Grainyhead_Grainyhead-Elf-1-NTF-1_M00951	0.0003943	CH_GLI_M01037	0.0011
CGCAGGTG	bHLH_myogenin_M00712	1.9715E-10	bHLH_E12_M00693	1.3238E-09	bHLH_E47_M00002	3.7507E-

Continued on next page

Table 33 – continued from previous page						
motif	ann	escore	ann	escore	ann	escore
GCGGCABC	AP2_ERF2_M01057	0.00011902	trp_Adf-1_M00171	0.00018409	bHLH_E12_M00693	0.000240
VYACCCKC	CH_MAZ_M00649	0.00017966	GCM_GCM_M00270	0.00076468	CC_VDR_M00444	0.002072
CTTTGATG	HMG_LEF1,_M00978	2.3634E-11	HMG_TCF-4_M00671	4.0074E-10	HMG_LEF1_M01022	6.9099E-1
CGGTGCAR	bHLH_E12_M00693	0.000001962	CH_Sn_M00060	0.000019307	CH_MTF-1_M00650	0.000040
GCGCGCTT	CH+BTB-POZ_ZF5_M00333	1.7164E-06	fork_E2F_M00803	0.000014662	NK-2-Nkx_Hmx3_M00433	0.000073
GGTTTGCA	bZIP_DBP_M00624	0.000013637	CH_Sn_M00044	0.000019789	POU_Oct-1_M00342	0.000020
AACCGCCG	trp_GAMYB_M00345	5.8997E-09	trp_c-Myb_M00004	1.1565E-06	AP2_ABI4_M00958	0.000023147
AAACCTCA	bZIP_ROM_M00700	7.2122E-06	runt_AML_M00769	0.00021754	runt_core-binding_M00722	0.00028857
CGGTTTGC	bZIP_DBP_M00624	1.0415E-06	bZIP_ABF1_M00401	0.000017406	CH_Sn_M00044	0.000096
GGGSCCCC	bHLH_PCF2_M00948	0.000010024	CH+BTB-POZ_HIC1_M01073	0.0012994	CH_Egr_M00807	0.001310

Table 34: Summary of all motifs and their assoiated Jaspar TFs for the Leukemia context

motif	ann	escore	ann	escore	ann	escore
ACCACA	RUNT_RUNX1	5.7621E-08	bHLH-ZIP_MYC-MAX	0.00023168	bHLH-ZIP_MAX	0.00097891
CATCSG	TRP-CLUSTER_Myb	0.000034483	ETS_ELK4	0.00017312	ETS_Eip74EF	0.00039626
AACCRC	RUNT_RUNX1	0.00012708	bHLH-ZIP_MYC-MAX	0.00054922	bHLH-ZIP_MAX	0.0013868
TCGGCA	HOMEO_CAAT_TLX1-NFIC	0.0038204	ZN-FINGER_C2H2_REST	0.007094	bHLH_Myf	0.028862
GCACCS	ZN-FINGER_C2H2_REST	0.000069726	HOMEO_CAAT_TLX1-NFIC	0.0001179	ZN-FINGER_C2H2_ZEB1	0.00039311
TGTGCG	IPT_TIG_domain_Su_H	0.034925	RUNT_RUNX1	0.050063	T-BOX_T	0.055672
GGTTTC	TRP-CLUSTER_IRF1	6.2212E-07	TRP-CLUSTER_IRF2	0.000032726	REL_dl_1	0.00023744
CGAGCG	ZN-FINGER_DOF_Dof3	0.030796	PAIRED_Pax6	0.060934	ZN-FINGER_DOF_MNB1A	0.068024
TGTGGC	RUNT_RUNX1	0.00023441	bZIP_bZIP911	0.00069408	E2F_TDP_E2F1	0.00075674
CGCCGC	E2F_TDP_E2F1	0.0051556	AP2_ABI4	0.0076941	ZN-FINGER_C2H2_MIZF	0.0077395
GGCGGC	AP2_ABI4	0.00004419	TRP-CLUSTER_Myb	0.0019506	bHLH_Myf	0.013824
CACCTG	ZN-FINGER_C2H2_sna	4.4299E-11	ZN-FINGER_C2H2_ZEB1	1.0052E-06	bHLH-ZIP_Mycn	0.000014686
CGGGGG	AP2_TFAP2A	0.00026785	ZN-FINGER_C2H2_MZF1_5-13	0.00046797	ZN-FINGER_C2H2_SP1	0.0009202
AGACGC	bHLH_Arnt-Ahr	0.0011376	bZIP_bZIP910	0.0034471	bZIP_TGA1a	0.0037555
CCCCAC	ZN-FINGER_C2H2_ZNF354C	0.00085421	IPT_TIG_domain_Su_H	0.0010179	ZN-FINGER_C2H2_MZF1_1-4	0.001235
CGCGCG	E2F_TDP_E2F1	0.0015614	bHLH_Arnt-Ahr	0.0089698	bHLH-ZIP_Mycn	0.03189
AMCCRCA	RUNT_RUNX1	3.9034E-06	TRP-CLUSTER_GAMYB	0.0034433	ETS_ELK4	0.0062236
RCCGCAN	ZN-FINGER_C2H2_MIZF	0.011609	ZN-FINGER_C2H2_ZNF354C	0.057167	TRP-CLUSTER_GAMYB	0.058846
CAGCAGC	bHLH_Myf	6.6967E-07	bHLH_NHLH1	0.000047956	ZN-FINGER_C2H2_REST	0.00023303
GGGTGCG	bHLH_NHLH1	0.00011495	T-BOX_T	0.00073986	ZN-FINGER_C2H2_ZEB1	0.0017041
GCASCTG	bHLH_NHLH1	1.1466E-07	ZN-FINGER_C2H2_sna	3.1251E-07	bHLH_Myf	0.000010447
CATCGGC	TRP-CLUSTER_Myb	0.00038175	NUCLEAR_RECEPTOR_RXRA-VDR	0.0029713	ZN-FINGER_GATA_Gata1	0.0045638
GYGGKTT	RUNT_RUNX1	0.002255	TRP-CLUSTER_GAMYB	0.00275	ZN-FINGER_C2H2_RREB1	0.0062399
GCGCCGC	AP2_ABI4	0.00018563	E2F_TDP_E2F1	0.00071461	bHLH_NHLH1	0.0092759

Continued on next page

Table 34 – continued from previous page						
motif	ann	escore	ann	escore	ann	escore
CRGATGC	bHLH_TAL1-TCF3	0.00024871	bHLH_NHLH1	0.00077259	ETS_ELK4	0.001142
CCCGAGC	P53_TP53	0.00051086	AP2_TFAP2A	0.0043915	bHLH-ZIP_Mycn	0.041794
CSGTTKC	TRP-CLUSTER_Myb	0.00012335	TRP-CLUSTER_GAMYB	0.00027937	TRP-CLUSTER_IRF1	0.00096085
GTGGCAG	HOMEO_CAAT_TLX1-NFIC	0.0010179	bHLH_Hand1-Tcfe2a	0.0057757	ZN-FINGER_C2H2_ZNF354C	0.0067184
TCCGCCG	ZN-FINGER_C2H2_SP1	0.00058251	AP2_ABI4	0.0054437	E2F_TDP_E2F1	0.010097
MGTGGTW	RUNT_RUNX1	0.000014065	bHLH-ZIP_MYC-MAX	0.00012299	bHLH-ZIP_MAX	0.00040518
AACCCTC	bHLH-ZIP_Spz1	0.0007909	TRP-CLUSTER_Myb	0.0014223	REL_dl_1	0.0024918
GCGCGCG	E2F_TDP_E2F1	0.0020809	bHLH_Arnt-Ahr	0.0040551	AP2_ABI4	0.010014
GGGGGGG	ZN-FINGER_C2H2_SP1	8.0451E-06	PAIRED-HOMEO_Pax4	0.000014409	ZN-FINGER_C2H2_RREB1	0.000021514
GHTGTGGT	RUNT_RUNX1	1.4452E-07	bHLH-ZIP_MYC-MAX	0.00029211	bHLH-ZIP_MAX	0.00087348
CAGMTGTK	bHLH_TAL1-TCF3	1.0638E-06	bHLH_NHLH1	0.000036024	ZN-FINGER_C2H2_sna	0.00039585
TGCGCGCG	E2F_TDP_E2F1	0.0016927	P53_TP53	0.015222	bHLH_Arnt-Ahr	0.019388
CACCCACA	ZN-FINGER_C2H2_RREB1	0.00005053	RUNT_RUNX1	0.00080632	IPT_TIG_domain_Su_H	0.0011062
AGAGCCGC	TRP-CLUSTER_ARR10	0.0034915	bHLH_NHLH1	0.026109	E2F_TDP_E2F1	0.062227
TCGGATGC	ETS_ELK4	0.00057952	ETS_Eip74EF	0.001072	ZN-FINGER_GATA_GATA2	0.0043781
GGAAGTGG	ETS_GABPA	9.4052E-06	ETS_ELK4	0.000018732	ETS_SPI1	0.000046438
GGGGGGGG	PAIRED-HOMEO_Pax4	1.7549E-06	ZN-FINGER_C2H2_RREB1	2.3983E-06	ZN-FINGER_C2H2_SP1	0.000085989
GMACCCRC	NUCLEAR_RECEPTOR_RXRA-VDR	0.00070899	REL_dl_1	0.0010418	REL_NFKB1	0.0039687
CGCAGGTG	ZN-FINGER_C2H2_sna	8.7801E-10	ZN-FINGER_C2H2_ZEB1	6.8494E-06	bHLH_NHLH1	0.000034757
GCGGCABC	bHLH_NHLH1	0.00080455	AP2_ABI4	0.00088062	HOMEO_CAAT_TLX1-NFIC	0.0015404
VYACCCKC	bHLH_TAL1-TCF3	0.0018302	ZN-FINGER_C2H2_MZF1_5-13	0.0045631	bHLH-ZIP_Spz1	0.0081315
CTTTGATG	NUCLEAR_RECEPTOR_NR1H2-RXRA	0.00022123	NUCLEAR_RECEPTOR_PPARG-RXRA	0.00038872	NUCLEAR_RECEPTOR_NR2F1	0.00063834
CGGTGCAR	HOMEO_CAAT_TLX1-NFIC	0.00026601	ZN-FINGER_C2H2_REST	0.003307	NUCLEAR_RECEPTOR_NR2F1	0.0044006
GCGCGCTT	ZN-FINGER_DOF_Dof3	0.0012832	ZN-FINGER_DOF_MNB1A	0.0031719	ZN-FINGER_DOF_PBF	0.0042561
GGTTTGCA	bZIP_Ddit3-Cebpa	0.000052978	T-BOX_T	0.0019874	REL_Dl_2	0.0060733
AACCGCCG	TRP-CLUSTER_GAMYB	3.8521E-07	TRP-CLUSTER_Myb	3.2709E-06	RUNT_RUNX1	0.0023322
AAACCTCA	bZIP_bZIP910	0.00018598	bZIP_TGA1a	0.00031427	RUNT_RUNX1	0.0019535
CGGTTTGC	TRP-CLUSTER_MYB.ph3	0.0020233	REL_Dl_2	0.003052	bZIP_Ddit3-Cebpa	0.0039598
GGGSCCCC	AP2_ABI4	0.000045542	NUCLEAR_RECEPTOR_usp	0.0046471	ZN-FINGER_C2H2_Roaz	0.0054784

Appendix F

Cross Condition ChIP-Seq Analysis

F.1 Background

Cross Condition ChIP-Seq Analysis (CCCA) is the R package that I wrote to preform the analysis presented in this thesis. It consists of 4 key components. The first component generates the AFS. It requires a list of peaks outputted by MACS in xls format. From the AFS the UDM generator is the second key component. The UDM is formed from a score for each peak that was determined for each data set being analyzed. In the case of my analysis this score was the raw read pile up count under the peak. The third component applies quantile normalization to the UDM and PCA to the resulting normalized matrix. The resulting PCs are then used for downstream analysis.

The downstream analysis portion forms the final component of the anlaysis. CCCA couples several techniques with the PCs of the UDM. For motif analysis the biostrings and the hg19 genome package is used to generate the basepair information. I provide wrappers for homer and stamp, which can be used for motif DENovo and association respectively. Please note that homer 2.0 must be installed on the system in order to use the wrapper. I have also included functionality for finding known motifs and preforming analysis for preferred distance. For gene assessment I have provided

the GREAT algorithm reimplemented in R that can be used to associate peaks with means.

F.2 Application

F.2.1 Set Up

Make sure that you have devtools installed and download CCCA.

```
devtools :: install_github("alexjgriffith/CCCA")

library(CCCA)
library( parallel )
library(ggplot2)
library(grid)
library(BSgenome.Hsapiens.UCSC.hg19)
```

F.2.2 Making the AFS and UDM

If the list of categories, peakfiles and raw data files is available the AFS and UDM can be formed using a single command.

```
categories<-c("jurk","eryt","cd34","cem_1")

peakFiles<-paste0("peakFiles/",categories,"_combined_mock.bed")
readFiles<-paste0("readFiles/",categories,"_unique_nodupes.bed")

env<-generate(peakFiles,rawFiles,macsCutOff=3,
              cl=20,categories=categories)

env$prc<-pca(env$heights)

sepFun<-sepAxis(list(name="Erythroid",pc=1,sd=2,fun=">"),
                list(name="Leukemia",pc=1,sd=-2,fun="<"),
```

```
                list (name="HSC",pc=2,sd=2,fun=">"))

center<-(normalize(env$prc$eigenVectors[,1])<inner) &
    (normalize(env$prc$eigenVectors[,1])>-inner) &
        (normalize(env$prc$eigenVectors[,2])>-inner) &
            (normalize(env$prc$eigenVectors[,4])>-inner) &
                (normalize(env$prc$eigenVectors[,4])<inner)

env$reg<-cbind(sepFun(env),center)
```

F.2.3 Creating Name Change Functions

In order for multiple names to be applied to each categories (so a Jurkat data set can be both jurk_sandar and leukemia) their must be mapping between categories. I have provided a basic tool for making a map that lets the user map from the data set specific code name to a non unique value.

```
swapFun<-createSwapFun(
          "jurk Jurkat eryt Erythroid cd34 CD34 cem_1 CEM")

swapFunB<-createSwapFun(
          "jurk Leukemia eryt Erythrod cd32 HSC cem_1 Leuekemia")

swapFunC<-createSwapFun(
          "Leukemia blue Erythroid red HSC orange")
```

F.2.4 Creating the Dendrogram

The dendrogram is a useful visualization for hierarchical clustering. I use the ARS Dendrogram package to colour code the dendogram with a predefined number of cuts.

```
dend(x,norm=pass,n=3,linkage="average",
     colours=c("green","red","blue"))
```

F.2.5 Ploting the PCs

I provide a utility for plotting the dot product of the normalized data and the eigen-vectors (PCs)

```
plotPCs(env$prc$eigenVectors,c(1,2),env$prc$normData,categories)

matr<-pca2Matr(env$prc)
plotPCMat2D(matr,c("PC1","PC2"),
            categories,swapFun,swapFunB,swapFunC)
```

F.2.6 Generating the Contribution Histogram

The contribution histogram benefits from the association of peaks with their source data sets. The histogram can be used to visualize the projection of all AFS peaks onto a single dimension.

```
p<-stackedContribWrapper(env$heights,env$over,
                         categories, list(1),swapFunB,swapFunC)
```

F.2.7 Motif Analysis

The preliminalry motif analysis I preformed relied on identification of peaks using Homer 2.0 and the annotation of these motifs using Stamp. Here I show how to use the wrappers I wrote in conjunction with Biostrings to generate a database of motifs and their annotations in R.

```
env<-addFasta(env)
 motifFile<-"test.motif"
motifs<-with(env,{
    homerWrapper(fasta,
                 reg[,"Leukemia"],
                 r[,"Center"],
                 "inst/lib/homer-4.7/bin/homer2",motifFile,
                 opts=paste0(
```

```
                    "-S 25 -len ",len," > /tmp/homerTrash 2>&1 "))
})
ann<-stampWrapper(motifFile,"TRANSFAC_Fams","Leukemia_Center")
```

F.2.8 Gene Analysis

Peaks were associated with the REFSeq list of coding genes using the GREAT algorithm implemented in R.

```
geneFile<-"hg19.RefSeqGenes.csv"
geneList<-read.delim(geneFile)
chrom<-as.character(geneList$chrom)
tss<-as.numeric(geneList$txStart)
strand<-geneList$strand
levels(strand)<-c(-1,1)
strand<-as.numeric(strand)
regions<-genomicRegions(chrom,tss,strand,1000,5000,1000000)
genes<-with(env,{geneMatrix(over,reg,regions,geneList)})
```

F.3 Future Directions

In the future my analysis would benefit from a downstream analysis that could take advantage of the PCs acting as a distance measure or a rank rather than just a tool to partition the peaks. One such method that could be applied is kmerHMM. This method defines a HMM using the oligotides present under the peak. Each oligotide is weighted either by rank or an alternative measure.

Bibliography

[1] V. Teif and K. Rippe, "Statistical-mechanical lattice models for protein-DNA binding in chromatin," *Journal of Physics*, vol. 22, no. 414105, pp. 1–14, 2010.

[2] K. Omari, S. Hoosdally, K. Tuladhar, D. Karia, E. Hall-Ponsele, O. Platonova, P. Vyas, R. Patient, C. Porcher, and E. Mancini, "Structural basis for LMO2-driven recruitment of the SCL:E47bhh heterodimer to hematopoietic-specific transcriptional targets," *Cell Reports*, vol. 4, pp. 135–147, 2013.

[3] C. Palii, C. Perez-Iratxera, Z. Yao, Y. Cao, F. Dai, J. Davison, H. Atkins, D. Allan, J. Dilworth, R. Gentleman, S. Tapscott, and M. Brand, "Differential genomic targeting of the transcription factor TAL1 in alternate hematopoietic lineages," *EMBO*, vol. 20, pp. 494–509, 2011.

[4] C. Porcher, W. Swat, K. Rockwell, Y. Fujiwara, and F. Alt, "The T Cell leukemia oncoprotein SCL/Tal1 is essential for development of all hematopoietic lineages," *Cell*, vol. 86, pp. 47–57, 1996.

[5] E. Lecuyer and T. Hoang, "SCL: From the origin of hematopoiesis to stem cells and leukemia," *Experimental Hematology*, vol. 32, pp. 11–24, 2004.

[6] A. Bardet, Q. He, J. Xeitlinger, and A. Stark, "A computational pipeline for comparative ChIP-Seq analysis," *Nature Protocols*, vol. 7, no. 1, pp. 45–51, 2011.

[7] T. Bailey, P. Krajewski, I, Ladunga, C. Lefebvre, Q. Li, T. Liu, P. Madrigal, C. Taslim, and J. Zhang, "Practical guidelines for the comprehensive analysis of chip-seq data," *PLOS Computational Biology*, vol. 9, no. 11, pp. 1–8, 2013.

[8] Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nusbaum, R. Myers, M. Brown, W. Li, and X. Liu, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biology*, vol. 9, no. 9, p. R137, 2008.

[9] M. Toscano, O. Navarro-Montero, V. Ayllon, V. Ramos-Mejia, X. Guerro-Carreno, C. Bueno, T. Romero, M. Lamolda, M. Cobo, F. Martin, P. Menendez, and P. Real, "SCL/TAL1-mediated trascriptional network enhances megakaryocytic specification of human embryonic stem cells," *Molecular Therapy*, vol. 23, no. 1, pp. 158–170, 2015.

[10] W. Wu, C. Morrsissy, C. Keller, T. Mishra, M. Pimkin, G. Blobel, M. Weiss, and R. Hardison, "Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis," *Genome Research*, vol. 24, pp. 1945–1962, 2014.

[11] O. Tamplin, E. Durand, L. Carr, S. Childs, E. Hagedorn, P. Li, A. Yzaguirre, N. Speck, and L. Zon, "Hematopoietic stem cell arrival triggers dynamic remodeling of the perivascular niche," *CELL*, vol. 160, pp. 241–252, 2015.

[12] D. Wu, D. Bittencourt, M. Stallcup, and K. Siegmund, "Identifying differential transcription factor binding in ChIP-Seq," *Frontiers in Genetics*, vol. 6, pp. 1–11, 2015.

[13] C. Porcher, E. Liao, Y. Fujiwara, L. Zon, and S. Orkin, "Specification of hematopoietic and vascular development by the bHLH transcription factor SCL without direct DNA binding," *Development*, vol. 126, pp. 4604–4615, 1999.

[14] J. Lacombe, S. Herblot, S. Sutterlin, A. Haman, S. Barakat, N. Iscove, G. Sauvageau, and T. Hoang, "SCL regulates the quiescence and the long-term competence of hematopoietic stem cells," *Blood*, vol. 115, no. 4, pp. 792–803, 2010.

[15] M. Koller and B. Palsson, "Tissue engineering: Reconstitution of human hematopoiesis ex vivo," *Biotechnology and Bioengineering*, vol. 42, pp. 909–930, 1993.

[16] T. Miyamoto, H. Iwasaki, B. Reizis, M. Ye, T. Graf, I. Weissman, and K. Akashi, "Myeloid or lymphoid promiscuity as a critical step in hematopoietic lineage commitment," *Dev. Cell*, vol. 3, no. 1, pp. 137–147, 2002.

[17] K. Choi, M. Kennedy, A. Kazarov, J. Papadimitriou, and G. Keller, "A common precursor for hematopoietic and endothelial cells," *Development*, vol. 125, pp. 725–732, 1998.

[18] M. Kondo, K. Akashi, and I. Weissman, "Identification of clonogenic common lymphoid progenitors in mouse bone marrow," *Cell Press*, vol. 91, pp. 661–667, 1997.

[19] K. Akashi, D. Traver, T. Miyamoto, and I. Weissman, "A clonogenic common myeloid progenitor that gives rise to all myeloid lineages," *Nature*, vol. 404, pp. 193–197, 2000.

[20] T. Asahra, A. Kawamtoto, and H. Masuda, "Concise review: Circulating endothelial progenitor cells for vascular medicine," *Stem Cells*, vol. 29, pp. 1650–1655, 2011.

[21] J. Isner and T. Asahra, "Angiogensis and vasculogensesis as therapeutic strategies for postnatal neovascularization," *Journal of Clinical Investigation*, vol. 103, no. 9, pp. 1231–1236, 1999.

[22] C. Palii, B. Vulesevic, S. Fraineau, E. Pranckeviciene, A. Griffith, A. Chu, H. Faralli, Y. Li, B. McNeill, J. Sun, T. Perkins, J. Dilworth, C. Perez, E. Suronen, D. Allan, and M. Brand, "Trichostatin A enhances vascular repair by injected human endothelial progenitors through increasing expression of TAL1 dependent genes," *CELL Stem Cell*, vol. 14, pp. 644–657, 2014.

[23] N. Minegishi, J. Ohta, H. Yamagiwa, N. Suzuki, S. Kawauchui, Y. Zhou, S. Takahashi, N. Hayashi, J. Engel, and M. Yamamoto, "The mouse GATA2 gene is expressed in the para-aortic planchnopleura and aorta-gonads and mesonephros region," *EMBO*, vol. 21, no. 24, pp. 6700–6708, 2002.

[24] M. Endoh, M. Ogawa, S. Orkin, and S. Nishikawa, "SCL/TAL1 dependent process determines a competence to select the definitive hematopoietic lineage prior to endothelial differentiation," *EMBO*, vol. 21, no. 24, pp. 6700–6708, 2002.

[25] J. Dignam, R. Lebovitz, and R. Roeder, "Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei," *Nucleic Acids Research*, vol. 11, no. 5, pp. 1475–1489, 1983.

[26] D. Reinberg, G. Orphanides, R. Ebright, S. Akoultchev, J. Carcama, H. Cho, P. Cortes, R. Drapkin, O. Flores, I. Ha, J. Inostroza, S. Kim, T. Kim, P. Kumar, T. Legrange, G. LeRoy, H. Lu, D. Ma, E. Maldonado, A. Merino, F. Mermelstein, I. Olave, M. Sheldon, R. Shiekhattar, N. Stone, X. Sun, L. Weis, K. Yeung, and L. Zawel, "The RNA polymerase II general transcription factors past, present, and future," *Cold Spring Harbor Symposium on Quantitative Biology*, vol. 73, pp. 83–105, 1998.

[27] R. Roeder, "The role of general initation factors in transcription by RNA polymerase II," *TIBS Reviews*, vol. 21, pp. 327–334, 1996.

[28] M. Slattery, T. Zhou, L. Yang, A. Dantas, R. Gordan, and R. Rohs, "Absence of simple code: how transcription factors read the genome," *CELL Future Reviews*, vol. 39, no. 9, pp. 381–399, 2014.

[29] M. Levine and R. Tjian, "Transcription regulation and animal diversity," *Nature Review*, vol. 424, no. 10, pp. 147–151, 2003.

[30] T. Whitfield, J. Wang, P. Collins, C. Patridge, S. Aldred, N. Trinklein, R. Myers, and Z. Weng, "Functional analysis of transcription factor binding sites in human promoters," *Genome Biology*, vol. 13, p. R50, 20012.

[31] K. Zaret and J. Carroll, "Pioneer transcriptoin factors: Establishing competence for gene expression," *Genes and Development Reviews*, vol. 25, pp. 2227–2241, 20011.

[32] H. Hsu, L. Huang, J. Tsan, W. Funk, W. Wright, J. Hu, R. Kingston, and R. Baer, "Preferred sequences for DNA recognition by the TAL1 helix-loop helix proteins," *Molecular and Cellular Biology*, vol. 14, no. 2, pp. 1256–1265, 1994.

[33] J. Boyes, P. Byfield, Y. Nakatani, and V. Ogryzko, "Regulation of activity of the transcription factor GATA1 by acetylation," *Nature*, vol. 396, pp. 594–598, 1998.

[34] S. Shen, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature Genetics*, vol. 31, pp. 64–68, 2002.

[35] S. Whiteside and S. Goodbourn, "Signal transduction and nuclear targeting: regulation of transciption factor activity by subcellular localization," *Journal of Cell Science*, vol. 104, pp. 949–955, 1993.

[36] M. Babu, N. Luscombe, L. Aravind, M. Gerstein, and S. Teichmann, "Structure and evolution of transcriptional regulatory networks," *Current Opinion in Structural Biology*, vol. 14, pp. 283–291, 2004.

[37] M. Dawson and T. Kouzarides, "Cancer epigenetics: From mechanism to therapy," *CELL*, vol. 150, pp. 12–27, 2012.

[38] P. Rodriguez, E. Bonte, J. Krijgsveld, K. Kolodziej, B. Guyot, A. Heck, P. Vyas, E. Boer, F. Grosveld, and J. Strouboulis, "Gata-1 forms distinct activating and repressive complexes in erythroid cells," *EMBO*, vol. 24, pp. 2354–2366, 2005.

[39] T. Sanda, L. Lawton, I. Barrasa, Z. Fan, H. Kohlhammer, A. Gutierrez, W. Ma, J. Tatarek, Y. Ahn, M. Kelliher, C. Jamieson, L. Staudt, R. Young, and A. Look, "Core transcriptional regulatory circuit controlled by the TAL1 complex in human T-Cell acute lymphoblastic leukemia," *Cancer Cell*, vol. 22, pp. 209–221, 2012.

[40] N. Wilson, S. Foster, X. Wang, K. Knezevic, J. Schutte, O. Kaimakis, P. Chilarska, S. Kinston, W. Ouwehand, E. Dzierzak, J. Primanda, M. Brujin, and B. Gottgens, "Combinatorial transcriptional control in blood stem/progenitor

cells: Genome-wide analysis of ten major transcriptional regulators," *CELL Stem Cell*, vol. 7, pp. 532–544, 2010.

[41] M. Valtieri, A. Tocci, M. Gabbianelli, L. Luchetti, B. Masella, L. Vitelli, R. Botta, U. Testa, G. Condorelli, and C. Pechle, "Enforced TAL1 expression stimulates primitive, erythroid, and megakaryocytic progenitors but blocks the granulopoietic differentiation program," *Cancer Research*, vol. 58, pp. 562–569, 1998.

[42] P. Brunet, F. Armstrong, V. Duval, M. Rouyex, N. Goardon, P. Romeo, and F. Pflumio, "Low SCL/TAL1 expression reveals its major role in adult hematopoietic myeloid progenitors and stem cells," *Blood*, vol. 107, no. 9, pp. 2998–2004, 2006.

[43] C. Calkhoven, C. Muller, R. Martin, G. Krosl, T. Hoang, and A. Leutz, "Translational control of SCL isoform expression in hematopoietic lineage choice," *Genes and Development*, vol. 17, pp. 959–964, 2003.

[44] G. Condorelli, F. Facchiano, M. Valtieri, E. Proietti, L. Vitelli, V. Lulli, K. Hueber, C. Peschle, and C. Croce, "T-Cell directed TAL1 expression induces T-Cell malignancies in transgenic mice," *Cancer Research*, vol. 56, pp. 5113–5119, 1996.

[45] M. Kekkuher, D. Seldin, and P. Leder, "TAL1 indicates T-Cell acute lymphoblastic leukemia accelerated by casein kinase II alpha," *EMBO*, vol. 15, no. 19, pp. 5160–5166, 1996.

[46] J. Nelson, O. Denisenko, and K. Bomsztyk, "Protocol for the fast chromatin immunoprecipitation(ChIP) method," *Nature Protocols*, vol. 1, no. 2, pp. 179–185, 2006.

[47] J. Bartlett and D. Stirling, "A short history of the polymerase chain reaction," *Methods in Molecular Biology*, vol. 226, pp. 3–6, 2003.

[48] A. Aszodi, "MULTOVL: fast multiple overlaps of genomic regions," *Bioinformatics*, vol. 28, pp. 3318–3319, 2012.

[49] M. Xu, L. Steiner, H. Bogardus, T. Mishra, V. Schulz, R. Hardison, and P. Gallagher, "Identification of biologically relevant enhancers in human erythroid cells," *Journal of Biochemistry*, vol. 288, no. 12, pp. 8433–8444, 2013.

[50] M. Mansour, B. Abraham, L. Anders, A. Berezovskaya, A. Gutierrez, A. Durbin, J. Etchin, L. Lawton, S. Sallan, L. Silverman, M. Loh, S.Hunger, T. Sanda, R. Young, and T. Look, "An onncogenic super-enhancer formed through somatic mutation of noncoding intergenic element," *Science*, vol. 12, no. 346, pp. 1373–1377, 2014.

[51] A. Quinlan and I. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, pp. 841–842, 2010.

[52] E. Keogh and A. Mueen, "Curse of dimensionality," in *Encyclopedia of Machine Learning*. New York: Springer US, 2010, pp. 257–258.

[53] R. Cattell, "The data box: Its ordering of total resources in terms of possible relational systems." in *Handbook of multivariate experimental psychology*. Chicago: Rand-McNally, 1966, pp. 67–128.

[54] N. Novershtern, A. Subramanian, L. Lawton, R. Mak, N. Haining, M. McConkey, N. Habib, N. Yosef, C. Chang, T. Shay, G. Frampton, A. Drake, I. Leskov, B. Nilsson, F. Preffer, D. D. andm J. Evans, T. Liefeld, J. Smutko, and G. Ebert, "Densely interconnected transcriptional circuits control cell states in human hematopoiesis," *CELL*, vol. 114, pp. 296–309, 2011.

[55] G. Hu, D. Schones, K. Cui, R. Ybarra, D. Northrup, Q. Tang, L. Gattinoni, N. Restifo, S. Huang, and K. Zhao, "Regulation of nucleosome landscape and transcription factor targeting at tissue specific enhancers by BRG1," *Genome Research*, vol. 21, pp. 1650–1658, 2011.

[56] C. Palii, C. Perex-Iratxera, Z. Yao, Y. Cao, F. Dai, J. Davison, H. Atkins, D. Allan, J. Dilworth, R. Gentleman, S. Tapscott, and M. Brand, "Differential genomic targeting of the transcription factor TAL1 in alternate hematopoietic lineages," *EMBO*, vol. 20, pp. 494–509, 2011.

[57] J. Xu, Z. Shao, K. Glass, D. Bauer, L. Pinello, B. Handel, S. Hou, H. Stamatoyannopoulos, H. Mikkola, G. Yuan, and S. Orkin, "Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis," *Developmental Cell*, vol. 23, pp. 796–811, 2012.

[58] M. Snyder, M. Gerstein, S. Weissman, P. Farnham, and K. Struhl, "Encode data project: Stanford," *Unpublished*, 2011.

[59] M. Tijssen, A. Cvejic, A. Joshi, R. Hannah, R. Ferrera, A. Forrai, D. Bellissimo, S. Oram, P. Smethrust, N. Wilson, X. Wang, K. Ottersbach, D. Stemple, A. Green, W. Ouwehand, and B. Gottgens, "Genome-wide analysis of simultaneous GATA1/2, RUNX1, FL1, and SCL binding in megakaryocytes identifies hematopoietic regulators," *Development Cell*, vol. 20, no. 5, pp. 597–609, 2011.

[60] D. Wheeler, T. Barrett, D. Benson, S. Bryant, K. Canese, V. Chetvernin, D. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. Lipman, T. Madden, D. Maglott, V. Miller, J. Ostell, K. Pruitt, G. Schuler, M. Shumway, E. Sequeira, S. Scherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. Tatusov, T. Tatusova, L. Wagner, and E. Yaschenko, "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 36, pp. D13–D21, 2008.

[61] H. Li and R. Durbin, "Fast and accurate long read alignment with burrows-wheeler transform," *Bioinformatics*, vol. 26, no. 5, pp. 589–595, 2010.

[62] S. Heinz, S. Benner, N. Spann, N. Bertolino, Y. Lin, P. Lasio, J. CHeng, C. Murre, H. Sigh, and C. Glass, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B-Cell identities." *Mol. Cell*, vol. 28, no. 38, pp. 576–589, 2010.

[63] X. Zhao and et al., "JASPAR 2013: An extensively expanded and updated open-access database of transcription factor binding profiles," *Nucleic Acids Research*, vol. TBA, no. TBA, p. TBA, 2013. [Online]. Available: TBA

[64] S. Mahony and P. Benos, "STAMP: a web tool for exploring DNA-binding motif similarities," *Nucleic Acids Research*, vol. 35, pp. 253–358, 2007.

[65] C. McLean, D. Bristor, M. Hiller, S. Clarke, B. Schaar, C. Lowe, A. Wenger, and G. Bejerano, "GREAT imporoves functional interpretation of cis-regulatory regions," *Computational Biology*, vol. 5, no. 28, pp. 495–503, 2010.

[66] C. Trapnell, L. Pachter, and S. Salzberg, "Tophat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, p. 11051111, 2009.

[67] S. Anders, A. Reyes, and W. Huber, "Detecting differential usage of exons from RNA-Seq data," *Bioinformatics*, vol. 22, pp. 2008–2017, 2012.

[68] A. Gaspar-Maia, A. Alajem, F. Polesso, R. Sidharan, M. Mason, M. Heidersbach, J. Romalho, M. McManus, K. Plath, E. Meshorer, and M. Ramalho-Santos, "ChD1 regulates open chromatin and pluripotency of embryonic stem cells," *Nature*, vol. 460, pp. 863–866, 2009.

[69] J. Tober, A. Yzaguirre, E. Piwarzyk, and N. Speck, "Distinct temporal requirements for RUNX1 in hematopoietic progenitors and stem cells," *Development*, vol. 140, no. 18, pp. 2765–3776, 2013.

[70] J. Wnag, Q. Sun, Y. Morita, H. Jiang, A. Gross, A. Lechel, K. Hildener, L. Guachalla, A. Gompf, D. Harmann, A. Schambach, T. Wuestefeild, D. Dauch, H. Schrezenmeier, W. Hofmann, H. Nakauchi, Z. Ju, H. L., Zender, and L. Rudolph, "A differentiation checkpoint limits hematopoietic stem cell self-renewal response to DNA damage," *CELL*, vol. 160, pp. 1001–1014, 2011.

[71] J. Iwasaki, C. Somoza, J. Shigematsu, E. Diprez, J. Iwasaki, S. Mizuro, Y. Arinobu, K. Geary, P. Zhang, T. Dayaram, M. Fenyius, S. Elf, S. Chan, P. Kastner, C. Uettner, R. Murray, D. Tenenand, and K. Akashi, "Distinctive and indispensable roles of PU.1 in maintenance of hematopoietic stem cells and their differentiation," *Blood*, vol. 106, no. 5, pp. 1590–1600, 2005.

[72] U. Blank and S. Karlesson, "The role of SMAD signaling in hematopoiesis and transitional hematology," *Leukemia*, vol. 25, pp. 1379–1388, 2011.

[73] A. Mullen, D. Orlando, H. Newman, J. Loven, R. Kumar, S. Bilodeau, J. Reddy, M. Guenther, R. Dekoter, and R. young, "Master transciption factors determine cell-type specific responses to TGFb signaling," *CELL*, vol. 147, pp. 565–576, 2011.

[74] R. Supek, M. Bosnjak, N. Skunca, and T. Smuc, "Revigo sumarizes and visualizes long lists of gene ontology terms," *PLOS One*, vol. 6, no. 7, p. e21800, 2011.

[75] K. Liang and S. Keles, "Detecting differential binding of transcription factors with ChIP-Seq," *Bioinformatics*, vol. 28, pp. 121–122, 2012.

[76] Z. Shao, Y. Zang, G. Yuan, S. Orkin, and D. Waxman, "MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets," *Genome Biology*, vol. 13, p. R16, 2012.