# Functional Characterization of the Binding Properties of the Hematopoietic Transcription Factor PU.1

DISSERTATION ZUR ERLANGUNG DES
DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN
DER UNIVERSITÄT REGENSBURG

vorgelegt von

**Julia Minderjahn**

aus Regensburg

im Jahr 2018

# Functional Characterization of the Binding Properties of the Hematopoietic Transcription Factor PU.1

DISSERTATION ZUR ERLANGUNG DES
DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN
DER UNIVERSITÄT REGENSBURG

vorgelegt von

**Julia Minderjahn**

aus Regensburg

im Jahr 2018

The work presented in this thesis was carried out at the Clinic and Polyclinic of Internal Medicine III at the University Hospital Regensburg from May 2013 to October 2018.

Die vorliegende Arbeit entstand in der Zeit von Mai 2013 bis Oktober 2018 in der Klinik und Polyklinik für Innere Medizin III des Klinikums der Universität Regensburg.

Das Promotionsgesuch wurde eingereicht am: 05.10.2018

Die Arbeit wurde angeleitet von: Prof. Dr. Michael Rehli

Unterschrift:

________________________________

(Julia Minderjahn)

“Ever tried. Ever failed. No matter. Try Again. Fail again. Fail better.”

(Samuel Beckett)

# Table of Contents

# LIST OF FIGURES

# LIST OF TABLES

# 1 SUMMARY

Transcription factors are defined through their ability to recognize and bind specific sequence motifs in genomic DNA. Such sequence motifs are very small compared to the length of the human genome, which is why there is a large proportion of non-functional motifs, where binding of a transcription factor is undesired. The myeloid and B cell-specific transcription factor PU.1 provides a well-suited model to study global dynamic binding processes. This transcription factor is a central regulator of hematopoietic cell differentiation and plays diverse roles in different hematopoietic lineages by regulating cell-type specific genes. How this master transcription factor gains access to its binding sites in the context of chromatin is only partially understood yet. Here, I analyzed its motif cooperativeness and epigenetic regulation and used a transient mRNA-transfection model to study the *de novo* binding of PU.1 in the lymphatic leukemia cell line CTV-1 that neither expresses PU.1 nor its related ETS-factors SPIB and SPIC. Introduction of PU.1 rapidly initiated a gene expression program (as measured by RNA-sequencing) dominated by myeloid genes which were correlated with PU.1 expression across hematopoietic lineages. ATAC-sequencing revealed extensive remodeling of the chromatin upon PU.1 expression, which was partially associated with the deposition of the histone modification H3K27ac and the enhanced expression of neighboring genes. *De novo* remodeled sites were significantly associated with clusters of PU.1 sites and/or higher motif scores, suggesting that homotypic binding sites and high affinity consensus sequences are responsible for a large fraction of *de novo* remodeled PU.1 binding sites. Moreover, shared sites between PU.1 and its ETS-family members ETS-1 and FLI-1 seemed to enhance PU.1's chromatin remodeling capacity in the lymphoid cell line, likely by establishing novel ETS-dependent co-associations in less wide-open chromatin regions. PU.1 binding in pre-existing open chromatin, however, was predominantly found at single PU.1 binding sites with lower motif scores and many surrounding consensus motifs for other transcription factor families, including GATA and RUNX, which likely enable PU.1 binding at low affinity sites. Titration of PU.1 levels and the analysis of several deletion mutants showed that the efficient binding of PU.1 to *de novo* remodeled sites was dependent on PU.1 concentration, reduced in the absence of the glutamine-rich domain and even more diminished in the absence of the acidic domain, suggesting that the latter is required for accessing binding sites in closed chromatin. *In vivo* proximity-dependent biotinylation analysis (BioID) uncovered the association of PU.1 with several components of the SWI/SNF family of chromatin remodeling complexes, including ARID1A, SMARCD2 and SMARCA4 (BRG1) among others. These interactions were specifically lost in the PU.1 mutant lacking the acidic transactivation domain.

In conclusion, we could show that the *de novo* binding of PU.1 to nuclear DNA induces rapid and marked changes in the chromatin landscape of the lymphatic CTV-1 cell line, which requires the acidic transactivation domain and its interaction with the SWI/SNF remodeling complex.

# 2 INTRODUCTION

The majority of the cells of a multicellular organism comprises the same genetic information, however the specific gene expression profiles differ from cell to cell (Eeckhoute et al. 2009). Cell type-specific transcriptional programs establishing those distinct gene expression profiles are mainly regulated by transcription factors (TFs), which bind *cis*-regulatory DNA elements like promotors and enhancers in a sequence-dependent manner (Choukrallah and Matthias 2014). Besides basal TFs, which are expressed ubiquitous in every cell type (Zehavi et al. 2015), eukaryotes possess a diversity of cell type- and tissue-specific TFs, which regulate the transcription of cell type- and tissue-specific genes by either activating or repressing the basal transcription machinery. The vast majority of TFs recognize specific DNA motifs that frequently range from 6-12 base pairs (bp) in length, suggesting that most sequence-specific TFs will have million potential binding sites throughout the genome (Heinz et al. 2015). Nevertheless, the access to their potential binding sites is highly restricted and only a relatively small proportion of regulatory elements are effectively bound. Furthermore, since gene regulation is cell type-specific, a single TF shows distinct binding profiles in different cell types. This indicates that the interaction between the TFs and their target DNA motifs alone is not sufficient to explain their transcriptional output. Indeed, besides DNA sequence preferences the main component controlling cell type-specific TF occupancy is the dynamic state of the chromatin structure on nucleosome or higher-order structure levels (Choukrallah and Matthias 2014). The positioning of nucleosomes throughout the genome modifies the availability of binding sites to TFs and the basal transcription machinery, therefore affecting almost all DNA-dependent processes such as transcription, replication and recombination in a highly cell-specific manner (Tsompana and Buck 2014). The ability to overcome this restrictive structure of the chromatin may be a key feature of master TFs or so called pioneering factors. Specific features of those master regulators which enable them to overcome chromatin restriction remain to be defined, but may include their capability to recruit cofactors like chromatin remodeling complexes, as well as epigenetic modifiers to create active chromatin states. In addition, they may more efficiently compete with nucleosomes for DNA binding, either autonomously or in a cooperative fashion with additional factors. This project makes use of the well-studied master regulator of the hematopoietic system PU.1 to address the central question of how a TF is able to influence the expression profiles of diverse human cell types in cooperation with the cellular enzymes that shape chromatin accessibility.

## 2.1 Chromatin Accessibility & Transcription

It remains a key question in biology, how DNA-recognizing proteins interact with enzymes to either activate or silence regulatory elements in a given cellular context. In principle and as already stated above, transcriptional regulation occurs on two joint levels. The first one involves TFs as well as the basal transcription machinery, and the second involves the chromatin structure and its regulators. Chromatin features that are involved in this epigenetic regulation can be divided into three general classes. The first class involves specific histone modifications that are associated with altered activity states for both promoters and enhancers, as well as direct modifications of the genomic DNA sequence. The second class involves interactions of *cis*-regulatory DNA elements like long-range interactions between enhancers and their targets, which occur on a genome-wide scale and shaping the nuclear architecture. The third class deals with the accessibility of the chromatin and the enzymes mediating this process, which inhibit the access to the underlying DNA sequence and impede TF binding (Voss and Hager 2014).

### 2.1.1 Chromatin- & DNA-Modifications

The packaging of the DNA in the nucleus is one of the major features that regulates differential gene expression in eukaryotic cells. The basic subunits of chromatin are so-called nucleosomal core particles. Each nucleosome consists of 145-147 bp of DNA wrapped around an octamer comprised of two copies of each of the canonical histones H3, H4, H2A and H2B. These four proteins interact in an ordered manner during nucleosome assembly to establish the modular nature of the nucleosome. The core particles are connected by a short stretch of linker DNA, forming a structure resembling beads on a string, where the 'nucleosome-beads' are usually found every 200 bp (Lawrence et al. 2016). Linker histone H1 family members are another key component of chromatin. They bind to the nucleosomal core particle around the DNA entry and exit sites, stabilizing both, nucleosome structure and higher-order chromatin architecture (Hergeth and Schneider 2015). In general, this packaging of the DNA into nucleosomes impedes transcription, by physical obstruction as well as by bending the DNA, hence reducing its accessibility for TFs (Lawrence et al. 2016). To distinguish inactive or closed versus active or open chromatin, the terms hetero- and euchromatin respectively have been established (Bell et al. 2011). In heterochromatin, the DNA is tightly coiled, thus preventing the access of the transcriptional machinery. Heterochromatin is usually found in regions of repetitive DNA and often associated with the nuclear envelope. In the euchromatin state, the DNA is less tightly coiled, allowing the access of TFs and chromatin remodelers (Sanders and Mason 2016). The switch from an inactive, heterochromatin state to an active, accessible state of regulatory DNA elements is established in an active manner. As part of this, post-translational modifications (PTMs) of histones can affect nucleosome stability by altering the interactions within nucleosomes or within neighboring nucleosomes. Hence, PTMs influence all DNA-based processes, including chromatin folding, nucleosome remodeling, and transcriptional gene expression (Lawrence et al. 2016). Amino acids of the N-terminal histone tail can be acetylated, phosphorylated or methylated among other modifications (Bannister and Kouzarides 2011). Most of these PTMs are reversible, since the cell possesses specified histone-modifying enzymes for adding and removing these marks. These enzymes can be grouped into two major classes: writers like histone acetyl transferases (HATs) and histone methyltransferases (HMTs) and erasers such as histone deacetylases (HDACs) and histone demethylases (KDMs; Choukrallah and Matthias 2014). PTMs act in a combinatorial fashion to regulate transcriptional activity. Some modifications are associated with transcriptional activation and others with transcriptional repression, depending on their relative position in histones.

DNA methylation of cytosines (C) within CG (cytosine-guanine) dinucleotides (CpGs) provides an additional mechanism regulating gene expression and genome architecture. In the mammalian system, most of the CpGs are methylated, but their appearance seems to be rarer than statistically predicted. In part, this can be explained by the spontaneous deamination of the 5'-methyl cytosines (5mC) that results in thymine thereby creating G/T mismatches, which can be repaired by DNA glycosylases such as the methyl-CpG binding domain protein 4 (MBD4) and the thymine DNA glycosylase (TDG) followed by base excision repair (BER) pathways (Moore et al. 2013). Vertebrate CpG islands (CGIs) which are short interspersed DNA sequences, represent an exception to this rule and are GC-rich, CpG-rich, and predominantly unmethylated. CGIs are mostly associated with promotor regions, hence representing sites for transcription initiation (Deaton and Bird 2011). Nucleosome-associated DNA sequences however, show an increased rate of methylation when compared with the more accessible linker regions between nucleosomes. Gene bodies of highly expressed genes on the other hand, are heavily methylated, while active gene regulatory elements show a lower degree of methylation. Methylation patterns outside of CGIs strongly vary between different cell types and correlate with the resulting gene expression profile (Yin et al. 2017). When present, DNA methylation can block TF binding, either directly through interference with their base recognition or indirectly by recruiting methylation-specific binding proteins involved in gene repression (Domcke et al. 2015, Moore et al. 2013). Such proteins include methyl-CpG binding-domain (MBD) proteins, which are thought to be the readers of DNA methylation patterns and which recruit HDACs promoting local chromatin condensation. Moreover, emerging evidence suggests that some TFs lacking a MBD are also able to interact with methylated DNA (Zhu et al. 2016). The cell type-specific TF CCAAT/enhancer-binding protein-α (C/EBPα) for example was shown to specifically bind to methylated promotors in primary keratinocyte cultures from newborn mice leading to their activation upon differentiation (Rishi et al. 2010). A recent systematic approach to globally analyze the effect of DNA methylation on TF binding unveiled major differences in TF binding specificities to unmethylated and CpG-methylated DNA sequences. For example, several factors, which regulate embryonic stem cell (ESC) self-renewal like PRDM4, Nanog, and the octamer-binding TF 4 (OCT4), were shown to be capable of binding to methylated CpGs, which may in part explain their ability to reprogram differentiated cells towards a pluripotent state. Furthermore, this study could show that certain TF families differ in their CpG-methylation sensitivity. For instance, basic leucine zipper (bZIP)- and E26 transformation-specific (ETS)-family TFs were inhibited by methylated CpGs, whereas NFAT (nuclear factor of activated T cells) factors preferred to bind to methylated CpG sequences (Yin et al. 2017). In addition, the major enzymes setting and erasing DNA methylation were also shown to interact with several cell type-specific TFs, thus influencing the methylation pattern among various cell types (Suzuki et al. 2017). The hematopoietic master TF PU.1 for example, was shown to be involved in the recruitment of the *de novo* DNA methyltransferase 3B (DNMT3B) to hypermethylated promotors. Moreover, this factor also recruited the ten-eleven translocation methyl cytosine dioxygenase 2 (TET2), an enzyme involved in active DNA demethylation, to genes that become demethylated during monocyte (MO) to osteoclast (OC) differentiation, suggesting that this TF is able to directly interact with the main enzymes involved in DNA methylation and demethylation (de la Rica et al. 2013).

## 2.1.2 *Cis*-regulatory DNA Elements

### 2.1.2.1 Promoters

Transcription of a gene in eukaryotes is a highly complex process that requires precise coordination in the assembly of *trans*-acting factors through the recognition of various types of *cis*-regulatory DNA sequences. Promoters generally refer to DNA regions that allow accurate initiation of transcription of a certain gene, by directing RNA Polymerase II (RNA Pol II) to initiate transcription at the transcription start site (TSS; Hardison and Taylor 2012). A promoter is typically comprised of proximal, core and downstream elements (see **Figure 2-1**). The core promoter is defined as a stretch of DNA of around 50-100 bp surrounding the TSS of a gene that directly interacts with the transcription machinery (Kim and Shiekhattar 2015). Canonical core promotor elements include the TATA-box, the Initiator (Inr), the TFIIB recognition element (BRE), the downstream promotor element (DPE), and the downstream core element (DCE). However, most active promoters in a given cell are GC- and CpG-rich regions that lack TATA-boxes and tend to support initiation of transcription at a broad range of positions within a roughly 100 bp interval (Hardison and Taylor 2012). Although promoters can be structurally diverse, they all share the same function, which is the accurate initiation of messenger RNA (mRNA) transcription. The assembly of the transcription machinery at the core promoter is initiated by TFs that not only bind close to the TSS, but also to DNA elements ~100–200 bp upstream of the core promoter sequence. The binding of TFs leads to an increasing rate of transcription by facilitating the recruitment or assembly of the basal transcription machinery onto the core promoter or by mediating the recruitment of specific distal regulatory DNA sequences to the core promoter (Akbari et al. 2008), which will be further explored in the next chapter.

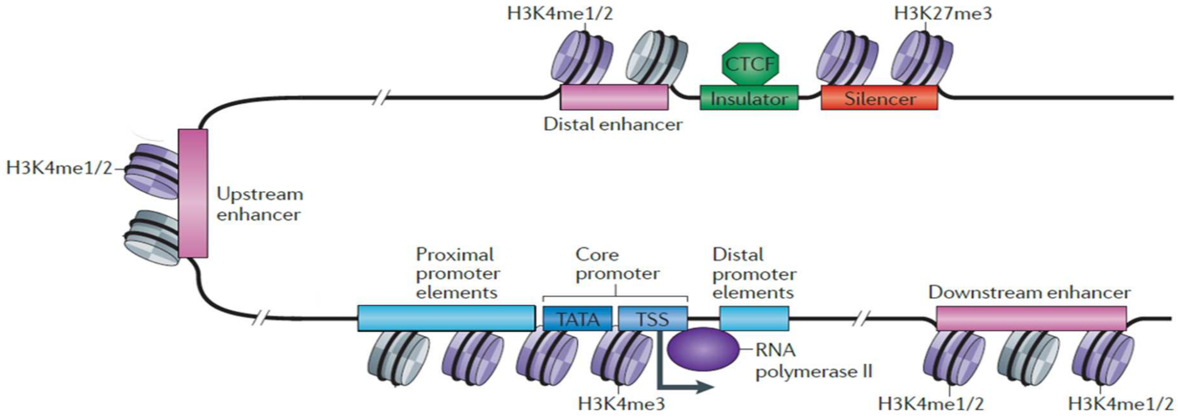


**Figure 2-1 - Schematic view of transcriptional regulatory elements**

The promoter usually consists of proximal, core and downstream elements. Multiple enhancers located distant from the promoter and interspersed with silencer and insulator elements influence transcriptional regulation. Regulatory DNA elements are marked by certain histone modifications reflecting either active (H3K4me1/2/3 i.e.) or passive (H3K27me3 i.e.) states (see next chapter; adapted from Ong and Corces 2011).

### 2.1.2.2 Enhancers, Silencers & Boundary Elements

Enhancers are one of the primary determinants of cell identity. Enhancers refer to DNA sequences, which can increase basal transcription levels from gene promoters and TSSs that can be located at any distance from their target gene varying between hundreds of bases up to megabases (Heinz et al. 2015). These *cis*-acting DNA sequences are able to increase the transcription of genes through cooperative and synergistic binding of TFs, DNA-binding effectors and chromatin-modifying complexes (Huang 2016). Enhancers function in an orientation-independent manner and harbor distinct chromatin features including increased chromatin accessibility, characteristic histone modifications, DNA hypomethylation, and bidirectional transcription (Bulger and Groudine 2011). Their most important feature however, is their ability to function as integrated platforms for TF binding (Buecker and Wysocka 2012). The main characteristics of enhancers are depicted in **Figure 2-2**. Enhancer states can be classified as inactive, primed, poised or active. An inactive enhancer is essentially located in compact chromatin and is devoid of TF binding and histone modifications. Primed enhancers are characterized by closely bound sequence-specific TFs that establish a nucleosome-free region, but still need to be activated upon distinct signaling pathways. Poised enhancers are primed enhancers, which harbor repressive epigenetic chromatin marks instead. Enhancer activation is a result of TF and nucleosome remodeler binding. TFs for instance, are able to recruit co-regulators such as the HAT p300, which is accompanied by the covalent modification of histone tails in enhancer-associated nucleosomes. Active enhancers are associated with acetylation of lysine 27 of histone 3 (H3K27ac) or H3K9ac, which often coincides with the presence of actively transcribing RNA Pol II (Heinz et al. 2015). These regulatory elements are often interspersed by silencers, which can be marked by histone 3 lysine 27 tri-methylation (H3K27me3). Silencers on the other hand, repress transcription mainly by recruiting repressive TFs or by blocking DNA-binding of additional activators. Furthermore, so-called boundary or insulator elements that are bound by regulatory proteins such as the CCCTC-binding factor (CTCF) also reduce the gene expression of a target gene by blocking the interaction of distal enhancers with their target promotors (see **Figure 2-1**). At active promotors, the interaction between the promotor and an enhancer is often mediated by cohesin, which stabilizes the DNA loop formed between those two (see **Figure 2-2**). However, many studies have shown that cohesin and the DNA-binding zinc-finger TF CTCF also co-localize on chromatin. Here, they lie together at the anchors of the loops and form so-called topologically associated domains (TAD), suggesting that cohesin is also implicated in the formation of insulator elements (Rao et al. 2017). In flies and mammals, CTCF is mainly enriched at accessible regions that do not function as active enhancers (Shlyueva et al. 2014). Thus, in contrast to enhancers, silencers and boundary elements usually harbor a repressive phenotype and create barriers between enhancers and other up- or downstream-located regulatory DNA elements like promotors (see **Figure 2-1**).

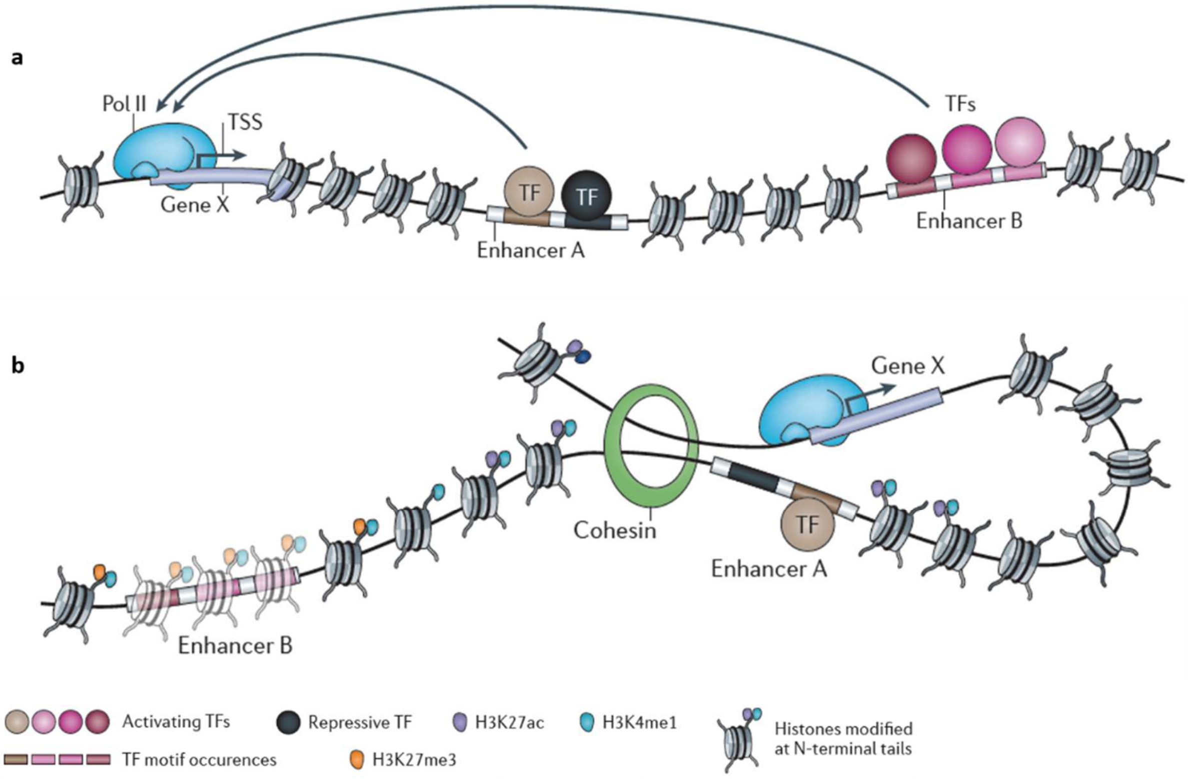


**Figure 2-2 - Enhancer features**

**a |** Enhancers are distinct genomic regions that contain TF binding sites (TFBSs) and have the potential to enhance basal transcription levels from gene promoters and TSS. **b |** Active enhancers (Enhancer A) are bound by activating TFs and are looped into proximity to their target promotors. Cohesins and other protein complexes are known to mediate this process. Furthermore, active enhancers are characterized by a depletion of nucleosomes and flanking nucleosomes show specific histone marks like H3K4me1 and H3K27ac. Inactive enhancers however, are associated with repressive marks like H3K27me3 (**b**) and repressive TF binding (**a**) (adapted from Shlyueva et al. 2014).

## 2.1.3 Transcription Factor Binding & Chromatin Remodeling

TFs can bind and regulate different targets at the same time and corresponding targets on the other hand can be regulated by several TFs. To unveil the complex mechanisms, which underlie differential gene regulation, the interaction of TFs with their potential target genes in different cells types and tissues, in combination with the interaction of cell type-specific TFs with each other is of major scientific interest (MacQuarrie et al., 2011). Many studies were carried out over the recent years to analyze the genome-wide and cell type-specific binding properties of TFs (Carroll et al. 2005; Eeckhoute et al. 2009; Heinz et al. 2010; John et al. 2011; Lefterova et al. 2010; Lupien et al. 2008; Mullen et al. 2011; Palii et al. 2011; Pham et al. 2012; Siersbaek et al. 2011). Those studies unveiled two key properties of TF binding. On the one hand, there are factors, which are expressed in the same cell type and tend to bind nearby cooperative DNA elements. On the other hand, a single TF itself can be expressed in distinct cell types and bind to various sequence motifs depending on the developmental status of the cell. This phenomenon might be explained by specific protein-protein interactions, which can influence the DNA structure to improve the binding of a cell type-specific TF. Besides, there is evidence for a close relationship between TF binding and the local occurrence or lack of various epigenetic modifications creating poised or activated chromatin states. This includes histone methylation and acetylation, diverse patterns of histone variants or DNA methylation. Those cell type-specific epigenetic signatures appear frequently on promotor-distal enhancer elements, which are usually marked by histone 3 lysine 4 mono- or di-methylation (H3K4me1/me2) and H3K27ac in an active state as already stated in the previous chapter. These distal and cell type-specific elements regulate the gene expression profile of a cell and can be bound cooperatively by different cell type-specific master regulators (Pham et al. 2012). Moreover, the cooperative competition of one or more factors with nucleosomes is of major functional importance to establish cell type-specific regulatory elements (Heinz et al. 2010).  In addition, pioneer factors, which are able to overcome chromatin restriction, at least on nucleosome level, are of special interest here, because they are thought to be capable of recruiting other cell type-specific TFs to those binding sites they just made accessible. Taking together, TFs must exploit various strategies to induce the local reorganization of the chromatin on the nucleosomal structure level to be able to stably interact with their specific regulatory DNA elements buried in chromatin. The disruption of the nucleosome structure that allows for TF binding is known to be coupled with the recruitment of specific adenosine triphosphate (ATP)-dependent remodeling complexes. Those proteins can recognize and bind to modified histones promoting transcription by several molecular mechanisms including nucleosome sliding and displacement, histone disassembly and the substitution of histone variants (Voss and Hager 2014; see **Figure 2-3**).

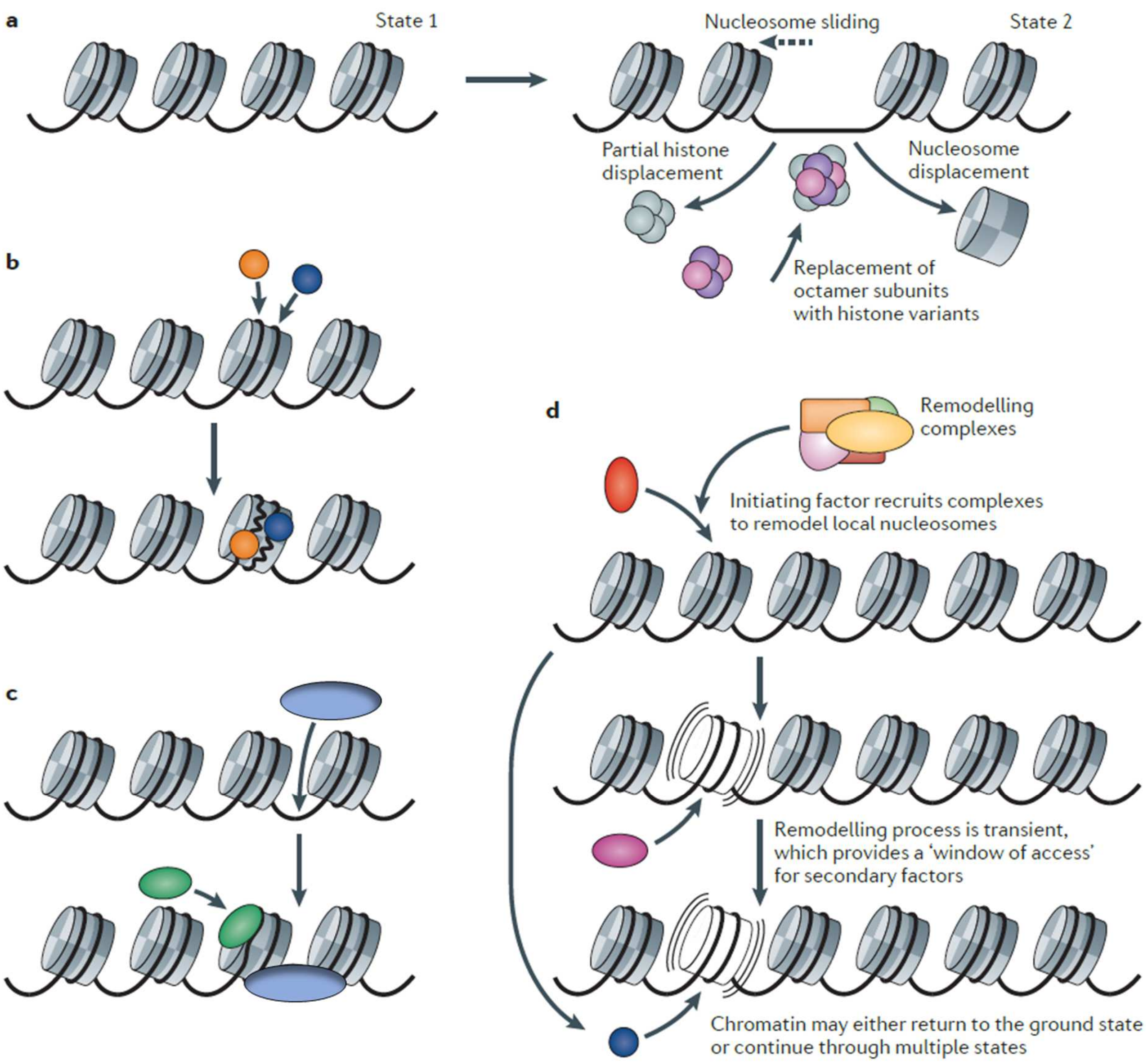

a
State 1
Nucleosome sliding
State 2
Partial histone displacement
Nucleosome displacement
Replacement of octamer subunits with histone variants
b
c
d
Remodelling complexes
Initiating factor recruits complexes to remodel local nucleosomes
Remodelling process is transient, which provides a 'window of access' for secondary factors
Chromatin may either return to the ground state or continue through multiple states

**Figure 2-3 - Mechanisms involved in chromatin remodeling**

**a |** Mechanisms allowing for increased local access of TFs to their binding sites include nucleosome sliding and displacement, histone displacement as well as replacement of octamer subunits with specific histone variants. **b |** A common mechanisms leading to TF binding to closed chromatin structures is thought to involve the cooperative nucleosome attack of two factors (orange and blue circles), generating enough free energy for binding to overcome the histone-DNA contacts in the nucleosome. **c |** Furthermore, pioneering factors (shown in blue) are thought to harbor special binding properties, which allow them to interact with closed chromatin, at least on nucleosome level, thus pinching the histone-DNA contacts and facilitating the binding of additional TFs (shown in green). The molecular mechanisms allowing for the modulation of the nucleosomal structure though, are not clarified yet. **d |** An alternative concept includes proteins which recruit multiple chromatin remodeling complexes to induce transient chromatin conformation changes allowing for secondary factors (red, pink and blue circles) to bind their binding sites (adapted from Voss and Hager 2014).

Chromatin remodeling complexes are large multi-protein assemblies typically comprised of an ATPase (ATP triphosphatase)-containing motor protein and several accessory proteins (Erdel et al. 2011). Based on their subunit composition and biochemical activity, remodeling complexes can be divided into the four major classes SWI/SNF, ISWI, CHD and INO80 (Wilson and Roberts 2011). Among those, the SWI/SNF-complex was the first one initially discovered in yeast. With more than 20% of cancers harboring mutations to one subunit of this 15-subunit complex, it may be one of the most important remodeling complexes in humans (Kadoch and Crabtree 2015).

In mammals, the SWI/SNF family can be divided into the BAF (BRG1-associated factor, also known as SWI/SNF-A) and the polybromo BRG1-associated factor (PBAF, also known as SWI/SNF-B) complexes. These complexes harbor one of two mutually exclusive catalytic ATPase subunits, which are either BRM (brahma homologue, also known as SMARCA2) or BRM/SWI2-related gene 1 (BRG1, also known as SMARCA4). Furthermore, they are comprised of several variant subunits, which contribute to the targeting and assembly of the nucleosomes as well as to the regulation of the lineage-specific functions of these complexes. For instance the AT-rich interactive domain containing proteins 1A (ARID1A, also known as BAF250A) and 1B (ARID1B) are mutually exclusive and only present in BAF complexes, whereas BAF200 and BRD7 (bromodomain-containing 7) subunits are only present in PBAF complexes (Wilson and Roberts 2011). Their well-conserved bromodomain leads to the recruitment of these factors to acetylated histones, where they initiate the sliding of the histone octamer along the DNA or even promote the complete removal of the octamer from the DNA (Hassan et al. 2002). The predicted remodeling mechanism and the induction of TF binding by the SWI/SNF complex are depicted in **Figures 2-4** and **2-5**, respectively.

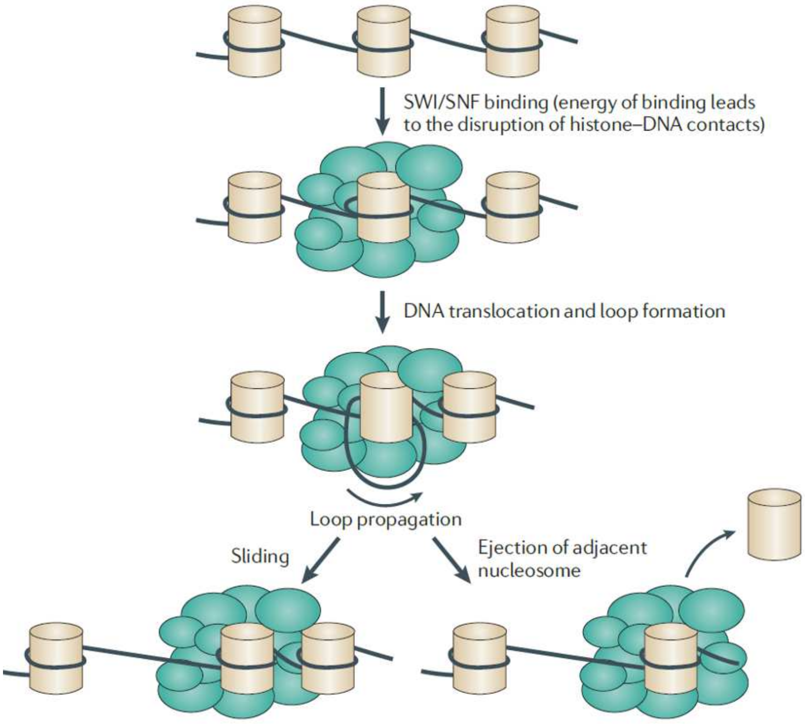

SWI/SNF binding (energy of binding leads to the disruption of histone–DNA contacts)
DNA translocation and loop formation
Loop propagation
Sliding
Ejection of adjacent nucleosome

**Figure 2-4 - SWI/SNF remodeling**

After binding of the SWI/SNF complex, the histone-DNA contacts are disrupted and a loop of DNA that propagates around the nucleosome is formed. This either leads to the sliding of a nucleosome, or as a consequence to the complete ejection of an adjacent nucleosome (adapted from Wilson and Roberts 2011).

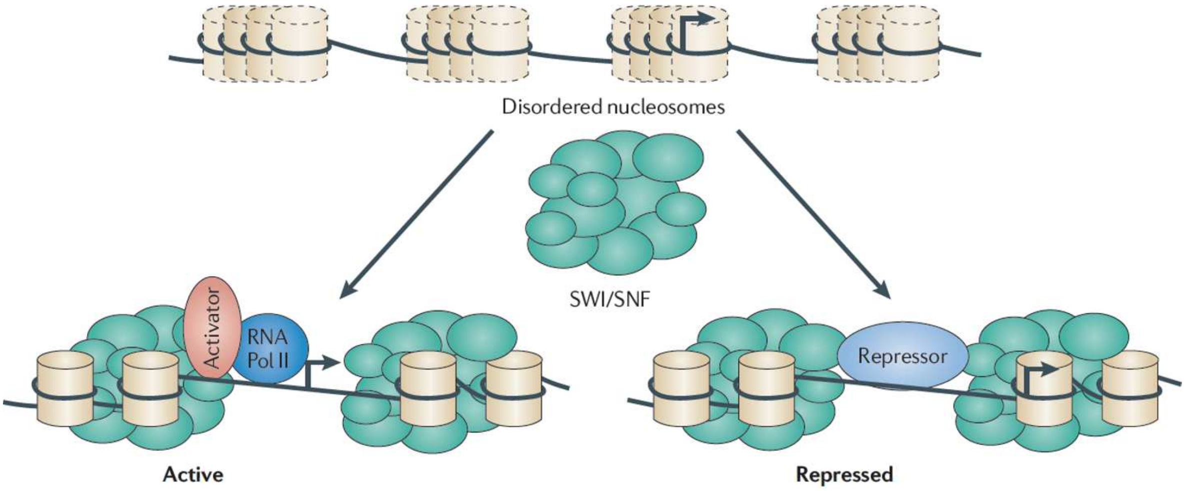


**Figure 2-5 - SWI/SNF induce TF binding**

At silent genes, where nucleosomes are less positioned (disordered) and TSS (arrow) are buried inside the chromatin, SWI/SNF enzymes are usually under-represented. Active, mostly lineage-specific genes however, show a strong SWI/SNF binding and their TSS is flanked by highly ordered nucleosomes allowing for the access of TF to their TSS in this nucleosome depleted region (bottom left). Moreover, these complexes can also be involved in the dynamic target silencing, mainly of genes involved in lineage-specific differentiation and proliferation by recruiting repressive cofactors (bottom right; adapted from Wilson and Roberts 2011).

Those complexes have key roles in the control of lineage-specific differentiation and were shown to play essential roles during neurogenesis, myogenesis, adipogenesis, osteogenesis as well as hematopoiesis. Studies analyzing the induction of cell type-specific gene regulation unveiled that many of the promotors of SWI/SNF-dependent genes were devoid of CGIs and more tightly packed into chromatin, indicating that SWI/SNF complexes are mostly involved in the remodeling of nucleosomes within repressive chromatin structures. This study is further supported by the antagonistic relationship between SWI/SNF complexes and the Polycomb group (PcG) proteins. PcG proteins are responsible for the tri-methylation of lysine 27 of histone 3, which is known to be a repressive chromatin mark and impairs SWI/SNF-mediated nucleosome remodeling (Wilson and Roberts 2011). Furthermore, nucleosome remodelers are able to cooperate with cell type-specific TFs to coordinate the activation of lineage-specific genes and the suppression of certain proliferation programs (see **Figure 2-5**). However, the mechanisms underlying these interactions are only poorly characterized so far. Pioneer factors though, help to solve the “chicken and egg” problem of how restrictive regions of nucleosomal DNA in chromatin become functional regulatory active regions, since they are able to target closed chromatin and thereby allow additional factors to bind. However, little is known about how these master regulators find their targets and coordinate additional cofactors (Zaret et al. 2016). The pioneering factors FoxA1 (forkhead box A1) and GATA4 (GATA binding factor 4) for instance were shown to target sites in isolated nucleosomes, suggesting that those factors are able to penetrate closed chromatin, thus creating open chromatin sites accessible for additional factors. Although this model suggests that long-term, static binding events are associated with pioneer factor binding, recent findings describe TF binding and nucleosome remodeling as a highly dynamic process, in which the residence times of most site-specific TFs on their cognate DNA motifs are rather short with a rapid exchange rate and a high mobility on those genomic sites (Swinstead et al. 2016). This leads to the assumption that additional mechanisms might exist, which allow pioneering factors to bind to restricted chromatin regions in a more dynamic manner. Those mechanisms strongly favor the idea that certain TFs can initiate chromatin remodeling events through the direct recruitment of chromatin remodeling enzymes. As part of this, it could be shown that BRG1 for example is correlated with the binding of the cell type-specific TF OCT4 in self-renewing ESCs. Moreover, it could be shown that the SWI/SNF complex itself is capable of displacing a TF from chromatin during nucleosome remodeling (Swinstead et al. 2016). Together these data suggest that ATP-dependent remodelers play a prominent role in regulating pioneering factor activity, since they either help to establish accessible chromatin regions allowing for TF binding or repress TF binding respectively. Thus, the functional analysis of mechanisms controlling the binding site selection of master regulators across different cell types with respect to the particular chromatin architecture is still of major interest for nowadays research and could unveil novel binding properties of these fascinating regulators of differential gene expression.

The hematopoietic system comprises a prime model to analyze TF functions and binding, and the myeloid and B cell-specific TF PU.1 (Spi1) is a well-studied example of a master regulator or pioneering factor involved in blood cell development. It is restricted to the hematopoietic compartment and required for the generation of common lymphoid and granulocyte-macrophage progenitor cells, as well as later stages of monocyte/macrophage (MO/MAC) and B cell development (Friedman 2007). Moreover, PU.1 is able to reprogram T cells or fibroblasts e.g. into macrophages (Arinobu et al. 2007; Feng et al. 2008; Laiosa et al. 2006). The next chapter will provide more insights into the biology of this important and fascinating regulator of the hematopoietic system.

## 2.2 The Master Regulator PU.1

The myeloid and B cell-specific TF PU.1 is a well-studied example of a master regulator or pioneering factor, respectively. This ETS-family member TF, encoded by the *Spi1* gene, is a key regulator of the hematopoietic system, controlling the expression of hundreds of genes including growth factor receptors, adhesion molecules, TFs and signaling components (Willis et al. 2017). However, its role at each hematopoietic stage is not fully understood yet. PU.1 is restricted to the hematopoietic compartment and is needed for the development of common myeloid progenitor cells (CMPs) as well as for later developmental stages of MO/MAC and B cells (Friedman 2007). The most important hematopoietic cell types regulated by PU.1 are depicted in **Figure 2-6**.

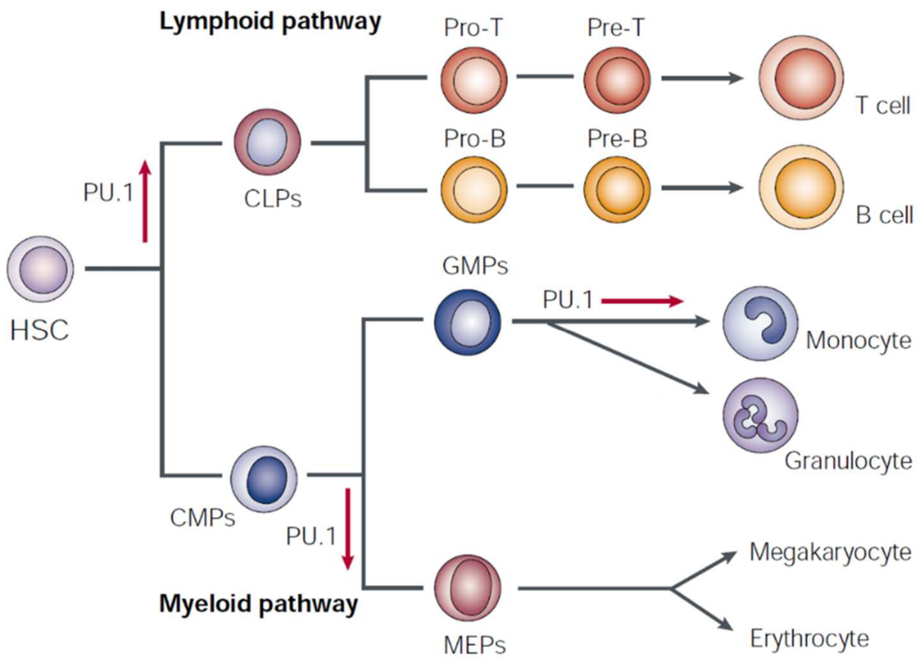


**Figure 2-6 - Hematopoiesis & PU.1**

PU.1 is involved in the differentiation of common myeloid progenitor cells as well as in the differentiation of granulocyte/monocyte progenitor cells (GMP) to mature granulocytes and monocytes/macrophages. Expression of PU.1 is also required for the transition from CMPs to megakaryocyte/erythroid progenitors (MEPs). PU.1-deficient progenitor cells are not able to differentiate and persist as CMPs (adapted from Tenen 2003).

PU.1 is expressed at low levels in CD34$^+$ HSCs and is upregulated during myeloid differentiation. PU.1-deficient mice show a strong impairment in hematopoiesis, die in utero and harbor deficiencies in the development of MAC, neutrophils, B cells and T cells, which was already described 1994 by Fisher and Scott. In terms of B cells, this requirement is limited to early lymphopoiesis however, as conditional deletion of PU.1 in CD19-expressing B cells is compatible with normal development and function (Willis et al. 2017). In addition, a study utilizing PU.1 disrupted bone marrow (BM) HSCs showed that those HSCs cannot maintain hematopoiesis and are overgrown by normal HSCs. They also fail to generate early myeloid and lymphoid progenitors and PU.1 disruption in GMPs blocks their maturation but not proliferation, resulting in myeloblast colony formation. In humans, loss-of-function mutations in the *Spi1* gene have been found in acute myeloid leukemia (AML; Iwasaki et al. 2005) and recurrent PU.1 fusions were also shown to be involved in high-risk pediatric T cell acute lymphoblastic leukemia (ALL; Seki et al. 2017).

PU.1 belongs to the ETS-family of TFs, which are characterized by an evolutionary well-conserved DNA-binding domain (DBD), namely the ETS domain. NMR-analysis of this domain unveiled a winged helix-turn-helix structure, which consists of three α-helices and four β-sheets, where the third α-helix is responsible for the contact and binding to the DNA (Donaldson et al. 1996). Besides the ETS-domain, PU.1 also harbors additional domains including an N-terminal acidic domain and a glutamine-rich domain, both mainly involved in transcriptional activation and therefore collectively named transactivation domain (TAD). Furthermore, PU.1 contains a so-called PEST domain (a peptide sequence rich in proline, glutamic acid, serine and threonine), which is involved in protein-protein interactions and likely responsible for its degradation (Burda et al. 2010; see **Figure 2-7**). For example, PU.1 is known to physically interact with general TFs like TFIID and TBP (TATA-box binding protein), with early hematopoietic TFs like GATA2 (GATA binding protein 2) and Runx-1 (runt-related transcription factor 1), and with other cell type-specific TFs like IRF4/8 (interferon regulatory factor 4/8) or c-Jun (Jun proto-oncogene, AP-1).

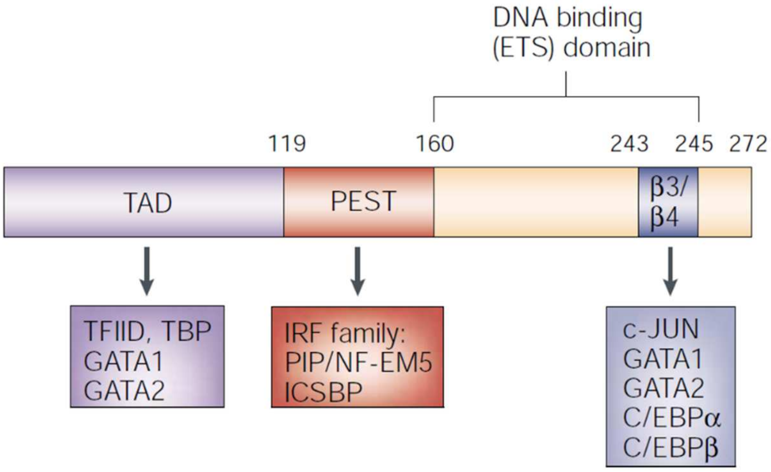


**Figure 2-7 - PU.1 & interacting proteins**

The N-terminal TAD of the 272 aa long protein PU.1 can interact with basal TFs like TFIID, TBP and with GATA1 and GATA2 for example. The PEST domain preferentially interacts with IRFs, including PIP/NF-EM5 (IRF4) and the interferon consensus sequence binding protein (ICSBP/IRF8). The C-terminal ETS domain interacts with c-Jun, GATA1, GATA2 and C/EBPα or C/EBPβ for example. These interactions result in either positive (c-Jun) or negative (GATA1, C/EBPα) effects on PU.1 activity (adapted from Tenen 2003).

In addition, the PU.1 protein can be post-translationally modified via phosphorylation of the serines 41, 142 and 148, which may enhance its activity. All ETS-family members share one major feature which is that they recognize a purine-rich 5'-GGAA/T-3' core sequence and that their specificity is often mediated by adjacent nucleotides (Wei et al. 2010). The ETS-family is one of the largest evolutionarily conserved TF families and comprises 28 factors in humans (Sizemore et al. 2017). They regulate the expression of multiple viral and cellular genes, often in a cooperative fashion with other sequence-specific TFs and additional cofactors. Adjacent to their DNA-binding sites, binding sites for ubiquitously expressed TFs like Sp1 (specificity protein 1), or for cell type-specific TFs like Runx-1 and C/EBP- factors have been found. The most well-known example however, is the interaction of the ETS-family with Jun-family proteins on DNA sequences called the Ras-responsive element (RRE), which consists of an ETS and an AP-1 binding site. ETS factors are preferentially expressed in specific cell lineages and regulate their development and differentiation by targeting specific genes, e.g. growth factors and integrines, which are unique for each cell lineage (Oikawa and Yamada 2003). Currently ETS proteins are divided into 9 classes according to their structural composition and their relative DNA binding site preferences, where PU.1 is grouped into the family of SPI factors together with SpiB and SpiC (Poon and Kim 2017).

Recent studies unveiled the cell type-specific distribution of PU.1 binding sites in mouse MAC or B cells (Ghisletti et al. 2010; Heinz et al. 2010), as well as in human MO and MAC (Pham et al. 2010/2013). PU.1 binding sites are often characterized by the co-occurrence of sequence motifs of other cell type-specific TFs. In murine MAC for example, PU.1 is preferentially associated with the nearby binding of the TFs C/EBPα/β and AP-1, whereas in murine B cells, PU.1 binding frequently occurs in combination with a distinct set of B cell specific TFs like E2A, EBF and OCT2 (Heinz et al. 2010). In human MAC PU.1 binding is mainly associated with nearby binding sites of EGR-2 and an unknown E-Box binding factor. Moreover, in primary human MO and MAC PU.1 shows a significant proportion of cell stage-specific binding events despite comparable PU.1 expression levels in both cell types (Pham et al. 2012). Besides binding to heterotypic clusters of TF binding sites, PU.1 is also an example for homotypic arrays of binding sites. Many murine MAC-specific promotors e.g. are characterized by the presence of multiple PU.1 binding sites (Ross et al. 1998). In addition to cell type-specific binding sites, which are mainly found at promoter distal sites, there is a core set of PU.1 occupied sites that is shared between hematopoietic cell types. These sites are frequently associated with motifs for promoter-located general transcription factors and likely regulate genes that are commonly involved in hematopoietic lineage selection.

Up to date, PU.1 consensus recognition sites, at least in human MO/MAC, can be grouped into three major classes (see **Figure 2-8**). Those are defined by the affinity of PU.1 to its specific binding sites, the cooperativeness between adjacent hetero- and homotypic binding sites, the nuclear concentration of PU.1, and the higher-order chromatin structure as well as the supposed capability of PU.1 to engage chromatin remodeling complexes or epigenetic modifiers. DNA methylation however, seems to play a minor role in restricting DNA binding of the master regulator PU.1, since cell type-specific recruitment of PU.1 to its binding sites is associated with local DNA demethylation (Pham et al. 2013).

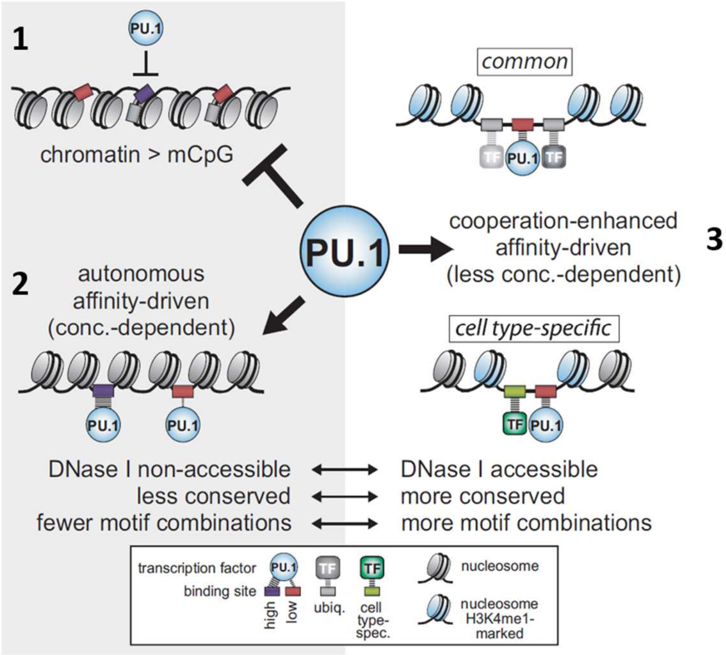


**Figure 2-8 - Three classes of PU.1 consensus motifs**

**1 |** Non-bound sites mainly show low binding affinity and reside in inactive chromatin. **2 |** PU.1-bound sites that are DNase I inaccessible mostly represent autonomous binding events preferentially at high affinity sites. **3 |** PU.1-bound, mostly intermediate- and low-affinity sites are DNase I accessible, and binding here is likely stabilized by cooperativeness with neighboring transcription factor binding sites. Increasing PU.1 concentration reduces the binding affinity threshold, leading to a marked increase in autonomous binding sites and to a lower extent in cell type-specific sites (adapted from Pham et al. 2013).

The dependency of PU.1 on motif binding affinity distinguishes this pioneering factor from many other TFs that are mainly guided by open chromatin and bind in a cooperative manner with other cell type-specific TFs. However, the contribution of each component and the mechanisms controlling and stabilizing TF access and binding to DNA on a functional level are not completely understood so far. Moreover, it is still not clear, why only a relatively small subset of TF binding sites distributed throughout the entire genome, are efficiently bound. Therefore, the myeloid and B cell-specific transcription factor PU.1 serves as a prime model to define specific molecular and functional features that enable it to overcome chromatin restriction and to stably bind to DNA.

## 2.3 Objectives

As already stated above, motif affinity, cooperativeness between neighboring binding sites and higher-order chromatin structure are the major determinants of PU.1 binding site selection. However, the impact of each of these parameters on PU.1 binding is not clarified yet and the rules for binding or not-binding of potential binding sites are only partially understood. This is why this thesis combines high-throughput-sequencing, computational, biochemical and functional analysis to address the question, how the binding of the myeloid and B cell-specific TF PU.1 to its binding sites on genomic DNA is controlled. To better understand the mechanisms that control cell type-specific binding of PU.1, its binding sites are mapped across different PU.1-expressing cell types and analyzed with regard to sequence-associated features of cell type-specific and common sites with respect to the distribution of distinct active and repressive histone marks as well as the co-occurrence of co-associated TF motifs. Moreover, the relationship between active and restrictive chromatin features - including DNA methylation and the chromatin structure - and PU.1 binding site selection is analyzed *in vivo*. Furthermore, the functional importance of PU.1 expression and its individual protein domains is studied to gain insights into the complex network of proteins, responsible for cell type-specific transcriptional regulation across different cell types. In summary, PU.1 binding properties are analyzed by high-throughput methods with regard to the accessibility of the chromatin, the local concentration of PU.1, its motif affinity and its functional impact as well as its dependence on additional co-associated factors.

# 3 MATERIAL AND EQUIPMENT

## 3.1 Equipment

| | |
|---|---|
| Autoclave | Walter, Geislingen, Germany |
| BioPhotometer | Eppendorf, Hamburg, Germany |
| Caliper LabChip XT | Perkin Elmer, Hamburg, Germany |
| Centrifuges | Heraeus, Hanau; Eppendorf, Hamburg, Germany |
| Covaris$^{TM}$ S220 | Covaris, Brighton, UK |
| Electrophoresis equipment | Bio-Rad, Munich, Germany |
| FLA-5000 | Fujifilm, Tokyo, Japan |
| Flow Cytometer (Aria, Fortessa, LSR II) | Becton Dickinson, Heidelberg, Germany |
| Fusion Pulse | Vilber Lourmat, Eberhardzell, Germany |
| Gene Pulser Xcell$^{TM}$ | Bio-Rad, Munich, Germany |
| Heat sealer | Eppendorf, Hamburg, Germany |
| Heatblock | Eppendorf, Hamburg, Germany |
| Incubators | Heraeus, Hanau, Germany |
| J6M-E centrifuge | Beckmann, Munich, Germany |
| Laminar air flow cabinet | Heraeus, Hanau, Germany |
| Mastercycler Nexus M2 | Eppendorf, Hamburg, Germany |
| Megafuge 3.0 R | Heraeus, Osterode, Germany |
| Microscopes | Zeiss, Jena, Germany |
| Multifuge 3S-R | Heraeus, Osterode, Germany |
| Multipipettor Multipette plus | Eppendorf, Hamburg, Germany |
| NanoDrop Spectrophotometer | PeqLab, Erlangen, Germany |
| Nanotemper Monolith NT.115 | Nanotemper, Munich, Germany |
| NextSeq 550 | Illumina, San Diego, USA |
| PCR-Thermocycler 4800 | Perkin Elmer, Überlingen, Germany |
| PCR-Thermocycler PTC-200 | MJ-Research/Biometra,Oldendorf, Germany |
| PCR-Thermocycler Veriti 384-well | Applied Biosystems, Foster City, USA |
| pH-Meter | Knick, Berlin, Germany |
| Picofuge | Heraeus, Hanau, Germany |
| Power supplies | Biometra, Göttingen, Germany |
| Qubit 2.0 Flurometer | Thermo Fisher Scientific, Waltham, USA |

| | |
|---|---|
| Realplex Mastercycler epGradientS | Eppendorf, Hamburg, Germany |
| Sonifier 250 | Branson, Danbury, USA |
| Sorvall RC 6 plus | Thermo Fisher Scientific, Waltham, USA |
| TapeStation 2200 | Agilent Technologies, Böblingen, Germany |
| Thermomixer | Eppendorf, Hamburg, Germany |
| Typhoon$^{TM}$ 9200 | Molecular Dynamics, Krefeld, Germany |
| Ultracentrifuge Optima L-70 | Beckman, Munich, Germany |
| Waterbath | Julabo, Seelstadt, Germany |
| Water purification system | Merck Millipore, Darmstadt, Germany |

## 3.2 Consumables

| | |
|---|---|
| 384-well PCR plates | Thermo Fisher Scientific, Waltham, USA |
| 8-channel pipettor tips Impact 384 | Thermo Fisher Scientific, Waltham, USA |
| Adhesive PCR sealing film | Thermo Fisher Scientific, Waltham, USA |
| Cell culture flasks and pipettes | Eppendorf, Hamburg, Germany |
| Cell culture plates (6-, 12-, 24-, 48-, 96-well) | Eppendorf, Hamburg, Germany |
| Centrifuge tubes (15, 50, 225 ml) | Falcon, Heidelberg, Germany |
| Cryo tubes | Nunc, Wiesbaden, Germany |
| Electroporation cuvettes (0.4 cm) | PeqLab, Erlangen, Germany |
| Heat sealing Film | Eppendorf, Hamburg, Germany |
| LabChip XT DNA 750 Assay Kit | Perkin Elmer, Hamburg, Germany |
| Micro test tubes (0.2 ml) | Biozym Scientific, Oldendorf, Germany |
| Micro test tubes (0.5, 1.5, 2, 5 ml) | Eppendorf, Hamburg; Sarstedt, Nümbrecht, Germany |
| Multiwell cell culture plates and tubes | Eppendorf, Hamburg, Germany |
| nProteinA Sepharose 4 FastFlow | GE Healthcare, Munich, Germany |
| nProteinG Sepharose 4 FastFlow | GE Healthcare, Munich, Germany |
| PCR plate Twin.tec 96 well | Eppendorf, Hamburg, Germany |
| Petri dishes | Falcon, Heidelberg, Germany |
| Pierce™ Streptavidin Magnetic Beads | Thermo Fisher Scientific, Waltham, USA |
| Sepharose Cl-4 beads | Sigma-Aldrich, Taufkirchen, Germany |
| Slide-A-Lyzer MINI Dialysis Devices | Thermo Fisher Scientific, Waltham, USA |
| Standard capillaries (NT.115) | Nanotemper, Munich, Germany |
| Sterile combitips | Eppendorf, Hamburg, Germany |
| Sterile micropore filters | Merck Millipore, Darmstadt, Germany |

| | |
|---|---|
| Sterile plastic pipettes | Costar, Cambridge, USA |
| Syringes and needles | Becton Dickinson, Heidelberg, Germany |
| Teflon foils | Heraeus, Hanau, Germany |
| twin.tec® real-time PCR-Platten | Eppendorf, Hamburg, Germany |

## 3.3 Chemicals

All reagents used were purchased from Sigma-Aldrich (Taufkirchen, Germany) or Merck Millipore (Darmstadt, Germany) unless otherwise noted. Oligonucleotides for PCR (polymerase chain reaction) and RT-qPCR (real time quantitative PCR) were synthesized and high-pressure liquid chromatography purified by Sigma-Aldrich (Taufkirchen, Germany). gBlocks gene fragments were synthesized and high-pressure liquid chromatography purified by Integrated DNA Technologies (IDT, San Jose, USA).

## 3.4 Enzymes and Kits

| | |
|---|---|
| Advantage 2 Polymerase Mix | Clontech, Mountain View, USA |
| Agencourt AMPure XP beads | Beckman Coulter, Krefeld, Germany |
| Alkaline Phosphatase | Thermo Fisher Scientific, Waltham, USA |
| ECL Prime Western Blotting System | Sigma-Aldrich, Taufkirchen, Germany |
| Blood and Tissue Culture Kit | Qiagen, Hilden, Germany |
| dNTP Mix | Agena Bioscience, San Diego, USA |
| dNTPs | Thermo Fisher Scientific, Waltham, USA |
| EZ DNA Methylation Kit | Zymo Research, Irvine, USA |
| Gibson Assembly® Master Mix | NEB, Frankfurt, Germany |
| Klenow Enzyme | Thermo Fisher Scientific, Waltham, USA |
| Klenow exo- (3’-5’ exo minus) | NEB, Frankfurt, Germany |
| Lipofectamine 2000 transfection reagent | Thermo Fisher Scientific, Waltham, USA |
| MinElute Gel Extraction Kit | Qiagen, Hilden, Germany |
| mMESSAGE mMACHINE® T7 Ultra Kit | Thermo Fisher Scientific, Waltham, USA |
| Monarch DNA Gel Extraction Kit | NEB, Frankfurt, Germany |
| Monarch PCR & DNA Cleanup Kit | NEB, Frankfurt, Germany |
| Monarch Plasmid Miniprep Kit | NEB, Frankfurt, Germany |
| MycoAlert® Mycoplasma Detection Kit | Lonza, Basel, Switzerland |
| Nextera XT DNA Library Preparation Kit | Illumina, San Diego, USA |
| Nextera XT Index Kit v2 | Illumina, San Diego, USA |
| NEXTflex® DNA Barcodes | Bioo Scientific Cooperation, Austin, USA |

| | |
|---|---|
| NextSeq 550 High Output v2 kit (75, 150, 300 cycles) | Illumina, San Diego, USA |
| NucleoSpin Plasmid Quick Pure | Macherey-Nagel, Düren, Germany |
| 25x Phosphataseinhibitor Cocktail | Thermo Fisher Scientific, Waltham, USA |
| 50x Proteaseinhibitor Cocktail | Thermo Fisher Scientific, Waltham, USA |
| Phusion High-Fidelity DNA Polymerase | Thermo Fisher Scientific, Waltham, USA |
| Plasmid Plus Midi Kit | Qiagen, Hilden, Germany |
| Proteaseinhibitors | Thermo Fisher Scientific, Waltham, USA |
| Proteinase K | Thermo Fisher Scientific, Waltham, USA |
| ScriptSeq™ Complete Kit (Human/Mouse/Rat) | Epicentre, Chicago, USA |
| ScriptSeq™ Index PCR Primers | Epicentre, Chicago, USA |
| QIAGEN Plasmid Plus Midi Kit | Qiagen, Hilden, Germany |
| QIAquick Gel Extraction Kit | Qiagen, Hilden, Germany |
| QIAquick PCR Purification Kit | Qiagen, Hilden, Germany |
| ReBlot Plus Mild Antibody Stripping Solution, 10x | Merck Millipore, Darmstadt, Germany |
| Restriction endonucleases | NEB, Frankfurt; Roche, Penzberg, Germany |
| Reverse Transcriptase SuperScript II | Promega, Madison, USA |
| RNeasy Midi and Mini Kit | Qiagen, Hilden, Germany |
| HS Taq DNA Polymerase | Thermo Fisher Scientific, Waltham, USA |
| T4 DNA Ligase | Promega, Madison, USA |

## 3.5 Antibodies

**Chromatin Immunoprecipitation (ChIP)**

| | |
|---|---|
| Anti-acetyl-Histone H3 (Lys27) | Abcam, Cambridge, UK |
| Anti-BRG1 | Abcam, Cambridge, UK |
| Anti-ETS1 | Santa Cruz, Heidelberg, Germany |
| Anti-FLAG M2 | Sigma-Aldrich, Taufkirchen, Germany |
| Anti-FLI-1 | Abcam, Cambridge, UK |
| Anti-IgG | Merck Millipore, Darmstadt, Germany |
| Anti-PU.1 (T-21) | Santa Cruz, Heidelberg, Germany |

**Western Blot (WB) Secondary Antibodies**

| | |
|---|---|
| Goat-anti-rabbit IgG, HRP-conjugated | Agilent Technologies, Böblingen, Germany |
| Goat-anti-mouse IgG, HRP-conjugated | Agilent Technologies, Böblingen, Germany |
| Rat-anti-mouse IgG, HRP-conjugated | Abcam, Cambridge, UK |

## 3.6 Molecular Weight Standards

| | |
|---|---|
| 1 kb Plus DNA Ladder | Thermo Fisher Scientific, Waltham, USA |
| GeneRuler$^{TM}$ 50 bp DNA Ladder | Thermo Fisher Scientific, Waltham, USA |
| PageRuler$^{TM}$ Plus Prestained Protein Ladder | Thermo Fisher Scientific, Waltham, USA |

## 3.7 Oligonucleotides

### 3.7.1 Primer for Cloning Experiments and PCR

| | |
|---|---|
| **hPU1_s** | 5’-GGGGACAAGTTTGTACAAAAAAGCAGGCTCGATGGAAGGGTTTCCCCTCGTC-3’ |
| **hPU1_as** | 5’-GGGGACCACTTTGTACAAGAAAGCTGGGTCGTGGGGCGGGTGGCGCCGCTC-3’ |
| **PU1_C_BamHI_s** | 5’-ATGAGGATCCGCCACCATGTTAC-3’ |
| **PU1_C_XbaI_as** | 5’-AGTATCTAGATCATTTGTCATCGTCAT-3’ |
| **PU1_N_BamHI_s** | 5’-ATTAGGATCCGCCACCATGGACT-3’ |
| **PU1_N_XbaI_as** | 5’-ATTATCTAGATCAGTGGGGCGGGTGG-3’ |
| **BS_CD14_01_s** | 5’-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTTTTAGGGATTAGGAAGGGATTTT-3’ |
| **BS_CD14_01_as** | 5’-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATACAACCCCACAACCTAAACCTCA-3’ |

### 3.7.2 Primer for Sequencing

| | |
|---|---|
| **pGENE fwd** | 5’-CCTGCTATTCTGCTCAACCT-3’ |
| **BGH rev** | 5’-TAGAAGGCACAGTCGAGG-3’ |
| **T7 Promotor** | 5’-TAATACGACTCACTATAGGG-3’ |
| **Linker seq** | 5’- ACTATGTCTGTGGACTCAAGC-3’ |

### 3.7.3 qPCR-Primer for ChIP

| | |
|---|---|
| **CHI3L1_s** | 5’-TGAGTCACATCTCCGGAAGC-3’ |
| **CHI3L1_as** | 5’-CTTTTATGGGAACTGAGCTATGTGTC-3’ |
| **CTCF_s** | 5’- CCCAGGCTCCTCTACACTTCAAGTC-3’ |
| **CTCF_as** | 5’- TAGGTGTGCCTCACACTGCCTG-3’ |
| **Empty 6.2_s** | 5’-GAAACCCTCACCCAGGAGATACAC-3’ |
| **Empty 6.2_as** | 5’-TGCAGTGGGACTTTATTCCATAGAAGAG-3’ |
| **GAPDH_1_s** | 5’-AGGCTGGATGGAATGAAAGGCAC-3’ |
| **GAPDH_1_as** | 5’-CTCCCACAAAGGCACTCCTG-3’ |
| **GAPDH_up_s** | 5’-GCTTCTCACAGGACTTCCCTTGTCTC-3’ |
| **GAPDH_up_as** | 5’-ACTGCCTATGGATCTGGAACTCCC-3’ |

| | |
|---|---|
| **GSX2_2_s** | 5’-GAAGGTCTATCTAATCCCTGCTGCGT-3’ |
| **GSX2_2_as** | 5’-CATTCCAGGGCAATCCTACAAACTCCA-3’ |
| | |
| **hSpi1_ups16kb_s** | 5’-CCAGTCACCACAGGAAGCATG-3’ |
| **hSpi1_ups16kb_as** | 5’-CGTTTCTCTGGGCCGCTGTG-3’ |
| | |
| **INS_IGF2_CTCF_s** | 5’-CCCAGAAATCTGCTACACCAAGCTTT-3’ |
| **INS_IGF2_CTCF_as** | 5’-AGACCTTTGGGTTATGCCACTGTAATTC-3’ |
| | |
| **SFRP2_2_s** | 5’- GGAGGGCGAAGTTCTTTCATATGTAAGG-3’ |
| **SFRP2_2_as** | 5’- TCTGAGCCTGTGAATGACTCTTAAGTGG-3’ |
| | |
| **TLR4_s** | 5’-CGCTATCACCGTCTGACCGAG-3’ |
| **TLR4_as** | 5’-CTTTCACTTCCTCTCACCCTTTAGCC-3’ |
| | |
| **PTPN6_1_s** | 5’-TCCGCCTTCCTTGTGACTTGAG-3’ |
| **PTPN6_1_as** | 5’-ACCAGAGGCAAAGAGAAACGCAG-3’ |

### 3.7.4 Primer for Library Preparation

| | |
|---|---|
| **BiooNEXTflex™ Primer 1** | 5’-AATGATACGGCGACCACCGAGATCTACAC-3’ |
| **BiooNEXTflex™ Primer 2** | 5’-CAAGCAGAAGACGGCATACGAGAT-3’ |

### 3.7.5 Oligonucleotides for Motif Affinity Measurements

All Cy-3-labeled and unlabeled oligonucleotides (high-affinity chromatography purified) were purchased from Sigma-Aldrich (Taufkirchen, Germany).

## 3.8 gBlocks® Gene Fragments

All gBlocks® gene fragments were synthesized and high-affinity chromatography purified from Integrated DNA Technologies (IDT, Coralville, USA).

**For Molecular Cloning**

PU.1-3xFLAG (C-terminal):

```
5’-ATTGCGCTAGCGCCACCATGTTACAGGCGTGCAAAATGGAAGGGTTTCCCCTCGTCCCCC
CTCCATCAGAAGACCTGGTGCCCTATGACACGGATCTATACCAACGCCAAACGCACGAGTATT
ACCCCTATCTCAGCAGTGATGGGGAGAGCCATAGCGACCATTACTGGGACTTCCACCCCCACC
ACGTGCACAGCGAGTTCGAGAGCTTCGCCGAGAACAACTTCACGGAGCTCCAGAGCGTGCAGC
CCCCGCAGCTGCAGCAGCTCTACCGCCACATGGAGCTGGAGCAGATGCACGTCCTCGATACCC
CCATGGTGCCACCCCATCCCAGTCTTGGCCACCAGGTCTCCTACCTGCCCCGGATGTGCCTCC
AGTACCCATCCCTGTCCCCAGCCCAGCCCAGCTCAGATGAGGAGGAGGGCGAGCGGCAGAGCC
CCCCACTGGAGGTGTCTGACGGCGAGGCGGATGGCCTGGAGCCCGGGCCTGGGCTCCTGCCTG
GGGAGACAGGCAGCAAGAAGAAGATCCGCCTGTACCAGTTCCTGTTGGACCTGCTCCGCAGCG
GCGACATGAAGGACAGCATCTGGTGGGTGGACAAGGACAAGGGCACCTTCCAGTTCTCGTCCA
AGCACAAGGAGGCGCTGGCGCACCGCTGGGGCATCCAGAAGGGCAACCGCAAGAAGATGACCT
ACCAGAAGATGGCGCGCGCGCTGCGCAACTACGGCAAGACGGGCGAGGTCAAGAAGGTGAAGA
AGAAGCTCACCTACCAGTTCAGCGGCGAAGTGCTGGGCCGCGGGGGCCTGGCCGAGCGGCGCC
ACCCGCCCCACGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGG
ATGACGATGACAAATGAACCGGTGTCAT-3’
```

pEF6-PU.1mut-3xFLAG:

```
5’-ATTGCGGATCCGCCACCTTGTTACAGGCGTGCAAATTGGAAGGGTTTCCCCTCGTCCCCC
CTCCATCAGAAGACCTGGTGCCCTTTGACACGGATCTATACCAACGCCAAACGCACGAGTATT
ACCCCTATCTCAGCAGTGTTGGGGAGAGCCATAGCGACCATTACTGGGACTTCCACCCCCACC
ACGTGCACAGCGAGTTCGAGAGCTTCGCCGAGAACAACTTCACGGAGCTCCAGAGCGTGCAGC
CCCCGCAGCTGCAGCAGCTCTACCGCCACTTGGAGCTGGAGCAGTTGCACGTCCTCGATACCC
CCTTGGTGCCACCCCATCCCAGTCTTGGCCACCAGGTCTCCTACCTGCCCCGGTTGTGCCTCC
AGTACCCATCCCTGTCCCCAGCCCAGCCCAGCTCAGTTGAGGAGGAGGGCGAGCGGCAGAGCC
CCCCACTGGAGGTGTCTGACGGCGAGGCGGTTGGCCTGGAGCCCGGGCCTGGGCTCCTGCCTG
GGGAGACAGGCAGCAAGAAGAAGATCCGCCTGTACCAGTTCCTGTTGGACCTGCTCCGCAGCG
GCGACTTGAAGGACAGCATCTGGTGGGTGGACAAGGACAAGGGCACCTTCCAGTTCTCGTCCA
AGCACAAGGAGGCGCTGGCGCACCGCTGGGGCATCCAGAAGGGCAACCGCAAGAAGTTGACCT
ACCAGAAGTTGGCGCGCGCGCTGCGCAACTACGGCAAGACGGGCGAGGTCAAGAAGGTGAAGA
AGAAGCTCACCTACCAGTTCAGCGGCGAAGTGCTGGGCCGCGGGGGCCTGGCCGAGCGGCGCC
ACCCGCCCCACGACTACAAAGACCTTGACGGTGATTATAAAGATCTTGACATCGATTACAAGG
TTGACGTTGACAATTGATCTAGAGTCAT-3’
```

3xFLAG-PU.1 (N-terminal):

```
5’-ATTGCGCTAGCGCCACCATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACA
TCGATTACAAGGATGACGATGACAAAATGTTACAGGCGTGCAAAATGGAAGGGTTTCCCCTCG
TCCCCCCTCCATCAGAAGACCTGGTGCCCTATGACACGGATCTATACCAACGCCAAACGCACG
AGTATTACCCCTATCTCAGCAGTGATGGGGAGAGCCATAGCGACCATTACTGGGACTTCCACC
CCCACCACGTGCACAGCGAGTTCGAGAGCTTCGCCGAGAACAACTTCACGGAGCTCCAGAGCG
TGCAGCCCCCGCAGCTGCAGCAGCTCTACCGCCACATGGAGCTGGAGCAGATGCACGTCCTCG
ATACCCCCATGGTGCCACCCCATCCCAGTCTTGGCCACCAGGTCTCCTACCTGCCCCGGATGT
GCCTCCAGTACCCATCCCTGTCCCCAGCCCAGCCCAGCTCAGATGAGGAGGAGGGCGAGCGGC
AGAGCCCCCCACTGGAGGTGTCTGACGGCGAGGCGGATGGCCTGGAGCCCGGGCCTGGGCTCC
TGCCTGGGGAGACAGGCAGCAAGAAGAAGATCCGCCTGTACCAGTTCCTGTTGGACCTGCTCC
GCAGCGGCGACATGAAGGACAGCATCTGGTGGGTGGACAAGGACAAGGGCACCTTCCAGTTCT
CGTCCAAGCACAAGGAGGCGCTGGCGCACCGCTGGGGCATCCAGAAGGGCAACCGCAAGAAGA
TGACCTACCAGAAGATGGCGCGCGCGCTGCGCAACTACGGCAAGACGGGCGAGGTCAAGAAGG
TGAAGAAGAAGCTCACCTACCAGTTCAGCGGCGAAGTGCTGGGCCGCGGGGGCCTGGCCGAGC
GGCGCCACCCGCCCCACTGAACCGGTGTCAT-3’
```

pEF6-3xFLAG-PU.1mut:

```
5’-ATTGCGGATCCGCCACCTTGGACTACAAAGACCTTGACGGTGATTATAAAGATCTTGACA
TCGATTACAAGGTTGACGTTGACAAATTGTTACAGGCGTGCAAATTGGAAGGGTTTCCCCTCG
TCCCCCCTCCATCAGAAGACCTGGTGCCCTTTGACACGGATCTATACCAACGCCAAACGCACG
AGTATTACCCCTATCTCAGCAGTGTTGGGGAGAGCCATAGCGACCATTACTGGGACTTCCACC
CCCACCACGTGCACAGCGAGTTCGAGAGCTTCGCCGAGAACAACTTCACGGAGCTCCAGAGCG
TGCAGCCCCCGCAGCTGCAGCAGCTCTACCGCCACTTGGAGCTGGAGCAGTTGCACGTCCTCG
ATACCCCCTTGGTGCCACCCCATCCCAGTCTTGGCCACCAGGTCTCCTACCTGCCCCGGTTGT
GCCTCCAGTACCCATCCCTGTCCCCAGCCCAGCCCAGCTCAGTTGAGGAGGAGGGCGAGCGGC
AGAGCCCCCCACTGGAGGTGTCTGACGGCGAGGCGGTTGGCCTGGAGCCCGGGCCTGGGCTCC
TGCCTGGGGAGACAGGCAGCAAGAAGAAGATCCGCCTGTACCAGTTCCTGTTGGACCTGCTCC
GCAGCGGCGACTTGAAGGACAGCATCTGGTGGGTGGACAAGGACAAGGGCACCTTCCAGTTCT
CGTCCAAGCACAAGGAGGCGCTGGCGCACCGCTGGGGCATCCAGAAGGGCAACCGCAAGAAGT
TGACCTACCAGAAGTTGGCGCGCGCGCTGCGCAACTACGGCAAGACGGGCGAGGTCAAGAAGG
TGAAGAAGAAGCTCACCTACCAGTTCAGCGGCGAAGTGCTGGGCCGCGGGGGCCTGGCCGAGC
GGCGCCACCCGCCCCACTGATCTAGAGTCAT-3’
```

**For Gibson Assembly**

pEF6-PU.1-BirA*:

```
5’-ACTATAGGGAGACCCAAGCTGGCTAGGTAAGCTTGGTACCGAGCTCGGATCCGCCACCAT
GTTACAGGCGTGCAAAATGGAAGGGTTTCCCCTCGTCCCCCCTCCATCAGAAGACCTGGTGCC
CTATGACACGGATCTATACCAACGCCAAACGCACGAGTATTACCCCTATCTCAGCAGTGATGG
GGAGAGCCATAGCGACCATTACTGGGACTTCCACCCCCACCACGTGCACAGCGAGTTCGAGAG
CTTCGCCGAGAACAACTTCACGGAGCTCCAGAGCGTGCAGCCCCCGCAGCTGCAGCAGCTCTA
CCGCCACATGGAGCTGGAGCAGATGCACGTCCTCGATACCCCCATGGTGCCACCCCATCCCAG
TCTTGGCCACCAGGTCTCCTACCTGCCCCGGATGTGCCTCCAGTACCCATCCCTGTCCCCAGC
CCAGCCCAGCTCAGATGAGGAGGAGGGCGAGCGGCAGAGCCCCCCACTGGAGGTGTCTGACGG
CGAGGCGGATGGCCTGGAGCCCGGGCCTGGGCTCCTGCCTGGGGAGACAGGCAGCAAGAAGAA
GATCCGCCTGTACCAGTTCCTGTTGGACCTGCTCCGCAGCGGCGACATGAAGGACAGCATCTG
GTGGGTGGACAAGGACAAGGGCACCTTCCAGTTCTCGTCCAAGCACAAGGAGGCGCTGGCGCA
CCGCTGGGGCATCCAGAAGGGCAACCGCAAGAAGATGACCTACCAGAAGATGGCGCGCGCGCT
GCGCAACTACGGCAAGACGGGCGAGGTCAAGAAGGTGAAGAAGAAGCTCACCTACCAGTTCAG
CGGCGAAGTGCTGGGCCGCGGGGGCCTGGCCGAGCGGCGCCACCCGCCCCACatgaaggacaa
cAccgtgcccctgaagctgatcgccctgctggccaacggcgagttccactctggcgagcagct
gggagagaccctgggaatGagcagagccgccatcaacaagcacatccagacactgagagactg
gggagtggacgtgttcaccgtgcctggcaagggct-3’
```

BirA*-3xFLAG-pEF6:

```
5’-catccagacactgagagactggggagtggacgtgttcaccgtgcctggcaagggctacag
cctgcctgagcctatccagctgctgaacgccaagcagatcctgggacagctggatggcggaag
cgtggccgtgctgcctgtgatcgactccaccaatcagtacctgctggacagaatcggagagct
gaagtccggcgacgcctgcatcgccgagtaccagcaggctggcagaggaggcagaggacggaa
gtggttcagcccattcggagccaacctgtacctgtccatgttctggagactggagcagggacc
tgctgctgccatcggactgagtctggtgatcggaatcgtgatggccgaggtgctgagaaagct
gggagccgacaaggtgagagtgaagtggcctaatgacctgtacctccaggaccgcaagctggc
tggcatcctggtggagctgacaggcaagacaggcgatgccgctcagatcgtgatcggagccgg
aatcaacatggccatgagaagagtggaggagagcgtggtgaaccagggctggatcaccctgca
ggaggctggcatcaacctggaccggaacaccctggccgccatgctgatcagagagctgagagc
cgctctggagctgttcgagcaggagggactggctccttacctgagcagatgggagaagctgga
caacttcatcaacagacctgtgaagctgatcatcggcgacaaggaaatcttcggcatctccag
aggaatcgacaagcagggagctctgctgctggagcaggacggaatcatcaagccctggatggg
```

```
cggagaaatctccctgagaagcgcagagaagctcgagGACTACAAAGACCATGACGGTGATTA
TAAAGATCATGACATCGATTACAAGGATGACGATGACAAATGATCTAGAGGGCCCGCGGTTCG
AAGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACGCGTACCGGT-3’
```

pEF6-PU.1-ΔA-BirA*:

```
5’-ACTATAGGGAGACCCAAGCTGGCTAGGTAAGCTTGGTACCGAGCTCGGATCCGCCACCAT
GTTACAGGCGTGCAAAATGGAAGGGCAGAGCGTGCAGCCCCCGCAGCTGCAGCAGCTCTACCG
CCACATGGAGCTGGAGCAGATGCACGTCCTCGATACCCCCATGGTGCCACCCCATCCCAGTCT
TGGCCACCAGGTCTCCTACCTGCCCCGGATGTGCCTCCAGTACCCATCCCTGTCCCCAGCCCA
GCCCAGCTCAGATGAGGAGGAGGGCGAGCGGCAGAGCCCCCCACTGGAGGTGTCTGACGGCGA
GGCGGATGGCCTGGAGCCCGGGCCTGGGCTCCTGCCTGGGGAGACAGGCAGCAAGAAGAAGAT
CCGCCTGTACCAGTTCCTGTTGGACCTGCTCCGCAGCGGCGACATGAAGGACAGCATCTGGTG
GGTGGACAAGGACAAGGGCACCTTCCAGTTCTCGTCCAAGCACAAGGAGGCGCTGGCGCACCG
CTGGGGCATCCAGAAGGGCAACCGCAAGAAGATGACCTACCAGAAGATGGCGCGCGCGCTGCG
CAACTACGGCAAGACGGGCGAGGTCAAGAAGGTGAAGAAGAAGCTCACCTACCAGTTCAGCGG
CGAAGTGCTGGGCCGCGGGGGCCTGGCCGAGCGGCGCCACCCGCCCCACactatgtctgtgga
ctcaagcctgcccagccccaaccagctgagcagccccagcctgggtttcgacggcctgcccgg
ccggatgaaggacaacaccgtgcccctgaagctgatcgccctgctggccaacggcgagttcca
ctctggcgagcagctgggagagaccctgggaatgagcagagccgccatcaacaagcacatcca
gacactgagagactggggagtggacgtgttcaccgtgcctggcaagggct-3’
```

pEF6-PU.1-ΔQ-BirA*:

```
5’-ACTATAGGGAGACCCAAGCTGGCTAGGTAAGCTTGGTACCGAGCTCGGATCCGCCACCAT
GCTGCAGGCTTGTAAAATGGAGGGCTTTCCGCTGGTGCCACCGCCAAGTGAAGACCTCGTCCC
TTATGATACTGATCTCTATCAGCGCCAGACACATGAATACTATCCATATCTTTCCAGCGACGG
TGAGTCCCACAGTGACCACTACTGGGATTTCCATCCACACCACGTCCATAGCGAGTTCGAGTC
CTTTGCTGAGAACAACTTCACCGAGTTGCCCCCCATCCTTCACTCGGGCACCAGGTCTCATA
TCTCCCTCGAATGTGCCTCCAGTATCCGAGCCTTAGCCCAGCTCAGCCGTCATCAGATGAGGA
AGAGGGCGAACGCCAATCTCCGCCCCTTGAAGTATCTGATGGCGAGGCTGATGGCCTTGAACC
TGGTCCCGGACTTTTGCCTGGAGAAACAGGGTCCAAAAAGAAGATCAGATTGTACCAGTTCCT
TCTTGACCTTTTGCGATCCGGTGATATGAAAGACTCCATCTGGTGGGTCGATAAGGACAAAGG
AACTTTCCAATTTTCAAGTAAACACAAGGAAGCGTTGGCACACAGATGGGGCATACAAAAGGG
CAATCGAAAGAAGATGACGTACCAAAAAATGGCGCGCGCTCTGCGCAACTATGGAAAGACCGG
AGAGGTTAAGAAAGTCAAGAAAAAACTTACATACCAGTTCAGCGGAGAGGTTTTGGGTCGGGG
AGGGCTTGCGGAAAGGCGACACCCTCCCCACactatgtctgtggactcaagcctgcccagccc
caaccagctgagcagccccagcctgggtttcgacggcctgcccggccggatgaaggacaacac
cgtgcccctgaagctgatcgccctgctggccaacggcgagttccactctggcgagcagctggg
agagaccctgggaatgagcagagccgccatcaacaagcacatccagacactgagagactgggg
agtggacgtgttcaccgtgcctggcaagggct-3’
```

pEF6-PU.1-ΔP-BirA*:

```
5’-ACTATAGGGAGACCCAAGCTGGCTAGGTAAGCTTGGTACCGAGCTCGGATCCGCCACCAT
GTTACAGGCGTGCAAAATGGAAGGGTTTCCCCTCGTCCCCCCTCCATCAGAAGACCTGGTGCC
CTATGACACGGATCTATACCAACGCCAAACGCACGAGTATTACCCCTATCTCAGCAGTGATGG
GGAGAGCCATAGCGACCATTACTGGGACTTCCACCCCCACCACGTGCACAGCGAGTTCGAGAG
CTTCGCCGAGAACAACTTCACGGAGCTCCAGAGCGTGCAGCCCCCGCAGCTGCAGCAGCTCTA
CCGCCACATGGAGCTGGAGCAGATGCACGTCCTCGATACCCCCATGGTGCCACCCCATCCCAG
TCTTGGCCACCAGGTCTCCTACCTGCCCCGGGGGCTCCTGCCTGGGGAGACAGGCAGCAAGAA
GAAGATCCGCCTGTACCAGTTCCTGTTGGACCTGCTCCGCAGCGGCGACATGAAGGACAGCAT
CTGGTGGGTGGACAAGGACAAGGGCACCTTCCAGTTCTCGTCCAAGCACAAGGAGGCGCTGGC
GCACCGCTGGGGCATCCAGAAGGGCAACCGCAAGAAGATGACCTACCAGAAGATGGCGCGCGC
GCTGCGCAACTACGGCAAGACGGGCGAGGTCAAGAAGGTGAAGAAGAAGCTCACCTACCAGTT
CAGCGGCGAAGTGCTGGGCCGCGGGGGCCTGGCCGAGCGGCGCCACCCGCCCCACactatgtc
tgtggactcaagcctgcCcagcccCaaccagctgagcagccccagcctgggtttcgacggcct
gcccggccggatgaaggacaacaccgTgcccctgaagctgatcgccctgctggccaacggcga
gttccactctggcgagcagctgggagagaccctgggaatgagcagagccgccatcaacaagca
catccagacactgagagactggggagtggacgtgttcaccgtgcctggcaagggct-3’
```

pEF6-PU.1-ΔAQP-BirA*:

```
5’-ACTATAGGGAGACCCAAGCTGGCTAGGTAAGCTTGGTACCGAGCTCGGATCCGCCACCAT
GTTACAGGCGTGCAAAATGGAAGGGGGGCTCCTGCCTGGGGAGACAGGCAGCAAGAAGAAGAT
CCGCCTGTACCAGTTCCTGTTGGACCTGCTCCGCAGCGGCGACATGAAGGACAGCATCTGGTG
GGTGGACAAGGACAAGGGCACCTTCCAGTTCTCGTCCAAGCACAAGGAGGCGCTGGCGCACCG
CTGGGGCATCCAGAAGGGCAACCGCAAGAAGATGACCTACCAGAAGATGGCGCGCGCGCTGCG
CAACTACGGCAAGACGGGCGAGGTCAAGAAGGTGAAGAAGAAGCTCACCTACCAGTTCAGCGG
CGAAGTGCTGGGCCGCGGGGGCCTGGCCGAGCGGCGCCACCCGCCCCACactatgtctgtgga
ctcaagcctgcccagccccaaccagctgagcagccccagcctgggtttcgacggcctgcccgg
ccggatgaaggacaacaccgtgcccctgaagctgatcgccctgctggccaacggcgagttcca
ctctggcgagcagctgggagagaccctgggaatgagcagagccgccatcaacaagcacatcca
gacactgagagactggggagtggacgtgttcaccgtgcctggcaagggct-3’
```

pEF6-3xFLAG-PU.1-ΔA:

```
5’-ACTATAGGGAGACCCAAGCTGGCTAGGTAAGCTTGGTACCGAGCTCGGATCCGCCACCAT
GGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGA
CAAAATGTTACAGGCGTGCAAAATGGAAGGGCAGAGCGTGCAGCCCCCGCAGCTGCAGCAGCT
CTACCGCCACATGGAGCTGGAGCAGATGCACGTCCTCGATACCCCCATGGTGCCACCCCATCC
CAGTCTTGGCCACCAGGTCTCCTACCTGCCCCGGATGTGCCTCCAGTACCCATCCCTGTCCCC
AGCCCAGCCCAGCTCAGATGAGGAGGAGGGCGAGCGGCAGAGCCCCCCACTGGAGGTGTCTGA
CGGCGAGGCGGATGGCCTGGAGCCCGGGCCTGGGCTCCTGCCTGGGGAGACAGGCAGCAAGAA
GAAGATCCGCCTGTACCAGTTCCTGTTGGACCTGCTCCGCAGCGGCGACATGAAGGACAGCAT
CTGGTGGGTGGACAAGGACAAGGGCACCTTCCAGTTCTCGTCCAAGCACAAGGAGGCGCTGGC
GCACCGCTGGGGCATCCAGAAGGGCAACCGCAAGAAGATGACCTACCAGAAGATGGCGCGCGC
GCTGCGCAACTACGGCAAGACGGGCGAGGTCAAGAAGGTGAAGAAGAAGCTCACCTACCAGTT
```

| | |
|---|---|
| | CAGCGGCGAAGTGCTGGGCCGCGGGGGCCTGGCCGAGCGGCGCCACCCGCCCCACTGATCTAG<br>AGGGCCCGCGGTTCGAAGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACGCGTA<br>CCGGT-3’ |
| pEF6-3xFLAG-PU.1-ΔQ: | 5’-ACTATAGGGAGACCCAAGCTGGCTAGGTAAGCTTGGTACCGAGCTCGGATCCGCCACCAT<br>GGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGA<br>CAAAATGTTACAGGCGTGCAAAATGGAAGGGTTTCCCCTCGTCCCCCCTCCATCAGAAGACCT<br>GGTGCCCTATGACACGGATCTATACCAACGCCAAACGCACGAGTATTACCCCTATCTCAGCAG<br>TGATGGGGAGAGCCATAGCGACCATTACTGGGACTTCCACCCCCACCACGTGCACAGCGAGTT<br>CGAGAGCTTCGCCGAGAACAACTTCACGGAGCTCCCACCCCATCCCAGTCTTGGCCACCAGGT<br>CTCCTACCTGCCCCGGATGTGCCTCCAGTACCCATCCCTGTCCCCAGCCCAGCCCAGCTCAGA<br>TGAGGAGGAGGGCGAGCGGCAGAGCCCCCCACTGGAGGTGTCTGACGGCGAGGCGGATGGCCT<br>GGAGCCCGGGCCTGGGCTCCTGCCTGGGGAGACAGGCAGCAAGAAGAAGATCCGCCTGTACCA<br>GTTCCTGTTGGACCTGCTCCGCAGCGGCGACATGAAGGACAGCATCTGGTGGGTGGACAAGGA<br>CAAGGGCACCTTCCAGTTCTCGTCCAAGCACAAGGAGGCGCTGGCGCACCGCTGGGGCATCCA<br>GAAGGGCAACCGCAAGAAGATGACCTACCAGAAGATGGCGCGCGCGCTGCGCAACTACGGCAA<br>GACGGGCGAGGTCAAGAAGGTGAAGAAGAAGCTCACCTACCAGTTCAGCGGCGAAGTGCTGGG<br>CCGCGGGGGCCTGGCCGAGCGGCGCCACCCGCCCCACTGATCTAGAGGGCCCGCGGTTCGAAG<br>GTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACGCGTACCGGT-3’ |
| pEF6-3xFLAG-PU.1-ΔP: | 5’-ACTATAGGGAGACCCAAGCTGGCTAGGTAAGCTTGGTACCGAGCTCGGATCCGCCACCAT<br>GGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGA<br>CAAAATGTTACAGGCGTGCAAAATGGAAGGGTTTCCCCTCGTCCCCCCTCCATCAGAAGACCT<br>GGTGCCCTATGACACGGATCTATACCAACGCCAAACGCACGAGTATTACCCCTATCTCAGCAG<br>TGATGGGGAGAGCCATAGCGACCATTACTGGGACTTCCACCCCCACCACGTGCACAGCGAGTT<br>CGAGAGCTTCGCCGAGAACAACTTCACGGAGCTCCAGAGCGTGCAGCCCCCGCAGCTGCAGCA<br>GCTCTCCGCCACATGGAGCTGGAGCAGATGCACGTCCTCGATACCCCCATGGTGCCACCCCAT<br>CCCAGTCTTGGCCACCAGGTCTCCTACCTGCCCCGGGGGCTCCTGCCTGGGGAGACAGGCAGC<br>AAGAAGAAGATCCGCCTGTACCAGTTCCTGTTGGACCTGCTCCGCAGCGGCGACATGAAGGAC<br>AGCATCTGGTGGGTGGACAAGGACAAGGGCACCTTCCAGTTCTCGTCCAAGCACAAGGAGGCG<br>CTGGCGCACCGCTGGGGCATCCAGAAGGGCAACCGCAAGAAGATGACCTACCAGAAGATGGCG<br>CGCGCGCTGCGCAACTACGGCAAGACGGGCGAGGTCAAGAAGGTGAAGAAGAAGCTCACCTAC<br>CAGTTCAGCGGCGAAGTGCTGGGCCGCGGGGGCCTGGCCGAGCGGCGCCACCCGCCCCACTGA<br>TCTAGAGGGCCCGCGGTTCGAAGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTAC<br>GCGTACCGGT-3’ |
| pEF6-3xFLAG-PU.1-ΔAQP: | 5’-ACTATAGGGAGACCCAAGCTGGCTAGGTAAGCTTGGTACCGAGCTCGGATCCGCCACCAT<br>GGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGA<br>CAAAATGTTACAGGCGTGCAAAATGGAAGGGGGGCTCCTGCCTGGGGAGACAGGCAGCAAGAA<br>GAAGATCCGCCTGTACCAGTTCCTGTTGGACCTGCTCCGCAGCGGCGACATGAAGGACAGCAT<br>CTGGTGGGTGGACAAGGACAAGGGCACCTTCCAGTTCTCGTCCAAGCACAAGGAGGCGCTGGC<br>GCACCGCTGGGGCATCCAGAAGGGCAACCGCAAGAAGATGACCTACCAGAAGATGGCGCGCGC<br>GCTGCGCAACTACGGCAAGACGGGCGAGGTCAAGAAGGTGAAGAAGAAGCTCACCTACCAGTT<br>CAGCGGCGAAGTGCTGGGCCGCGGGGGCCTGGCCGAGCGGCGCCACCCGCCCCACTGATCTAG<br>AGGGCCCGCGGTTCGAAGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACGCGTA<br>CCGGT-3’ |

## 3.9 Antibiotics

| | |
|---|---|
| Ampicillin | Roth, Karlsruhe, Germany |
| Blasticidin | InvivoGen, San Diego, USA |
| Chloramphenicol | Sigma-Aldrich, Taufkirchen, Germany |
| Hygromycin B | Clontech, Mountain View, USA |
| Kanamycin sulfate | Roth, Karlsruhe, Germany |

## 3.10 Plasmids

| | |
|---|---|
| pCR-Blunt II-TOPO | Thermo Fisher Scientific, Waltham, USA |
| pEF6/V5-His Topo | Thermo Fisher Scientific, Waltham, USA |
| pGENE/V5-His (mod.) | Australien |
| pORF9-hSPI1 | InvivoGen, San Diego, USA |

## 3.11 *E.coli* Strains

| | |
|---|---|
| Rosetta2(DE)pLysS | Merck Millipore, Darmstadt, Germany |
| DH10B | F- mcrA Δ(mrr-hsdRMS-mcrBC) ϕ80lacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK λ- rpsL nupG /pMON14272 / pMON7124 |

## 3.12 Cell Lines

**Human Cell Lines**

| | |
|---|---|
| THP-1 | Human acute monocytic leukemia (DSMZ ACC 16) |
| CTV-1 | Human acute myeloid leukemia (DSMZ ACC 40) |
| K-562 | Human chronic myeloid leukemia in blast crisis (DSMZ ACC 10) |

## 3.13 Databases and Software

| | |
|---|---|
| Adobe Illustrator v3/v5 | Adobe, San Jose, California, USA |
| Anaconda v5.0.1 | https://www.anaconda.com/download/ |
| ATOM | github.com/atom |
| Bcl2fastq v2.20.0.422 | Illumina, San Diego, USA |
| BEDtools2 v2.27.1 | Quinlan Laboratory, Virginia, USA |
| Bioconda packages | https://bioconda.github.io/ |
| BioGPS | http://biogps.org/#goto=welcome |
| BLAST | https://blast.ncbi.nlm.nih.gov/Blast.cgi |
| Bowtie2 v2.3.4 | www.bowtie-bio.sourceforge.net/index.shtml |
| Crispresso v1.0.10 | https://github.com/lucapinello/CRISPResso |
| Enrichr | http://amp.pharm.mssm.edu/Enrichr/ |
| ExPASy | http://web.expasy.org/ |
| FastQC v0.11.7 | https://www.bioinformatics.babraham.ac.uk/ projects/download.html |
| GENtle v1.9.4 | University of Cologne, Cologne, Germany |
| Graphia Professional v2.0 | KAJEKA, Edinburgh, UK |
| HOMER v4.10, 5-16-2018 | http://homer.ucsd.edu/homer/index.html |
| Illumina Experiment Manager v1.14.0 | Illumina Inc., San Diego, USA |
| Illumina Sequencing Analysis Viewer v2.1.18 | Illumina Inc., San Diego, USA |

| | |
|---|---|
| Integrative Genomics Viewer (IGV) v2.4.6 | Broad Institute, Cambridge, USA |
| IGVTools v2.3.98 | Broad Institute, Cambridge, USA |
| Jupyter v4.3.1 | https://jupyter.org/ |
| NCBI | https://www.ncbi.nlm.nih.gov/ |
| Metascape | http://metascape.org/ |
| Microsoft Office 2013 | Microsoft Corporation, Redmond, USA |
| NT-analysis acquisition software v1.2.229 | Nanotemper, Munich, Germany |
| Perlprimer v1.1.14 | http://perlprimer.sourceforge.net/ |
| Picard v2.17.3 | Broad Institute, Cambridge, USA |
| PubMed | www.ncbi.nlm.nih.gov/entrez |
| Python v3.6.3 | https://www.anaconda.com/download/ |
| Oracle VM VirtualBox v5.1.30 | Oracle Corporation, Redwood Shores, USA |
| RBioC v3.4.3 | http://www.bioconductor.org/ |
| SAMtools v1.6 | https://sourceforge.net/projects/samtools/files/ samtools |
| Sambamba v0.6.7 | https://github.com/biod/sambamba/releases |
| SnapGene Viewer v2.8.2 | GSL Biotech LLC, Chicago, USA |
| STAR v2.5.3a | https://github.com/alexdobin/STAR |
| STRING | https://string-db.org/ |
| TapeStation A.01.05 (SR1) | Agilent Technologies, Böblingen, Germany |
| UCSC Genome Browser | www.genome.ucsc.edu |
| UniProt | http://www.uniprot.org/ |

# 4 METHODS

## 4.1 General Cell Culture Methods

For washing and harvesting, mammalian cells were centrifuged using the general cell culture program: 8 min, 300×g, 4°C.

### 4.1.1 Cell Line Culture

#### 4.1.1.1 Cell Culture Conditions and Passaging

If not otherwise indicated, cells were cultured in RPMI-1640 (Gibco, Thermo Fisher Scientific, Waltham, USA, see **Table 4-1**) routinely supplemented with 10% heat inactivated FCS, L-glutamine (2 mM), sodium pyruvate (1 mM), antibiotics (50 U/ml penicillin and 50 µg/ml streptomycin), 2 ml vitamins, non-essential amino acids and 50 µM β-mercaptoethanol. Media supplements were purchased from Gibco (Waltham, USA) and Merck Millipore (Darmstadt, Germany; L-glutamine) respectively. FCS was heat inactivated for 20 min at 56°C before use. Each batch of FCS and culture medium was tested before use.

Cells were cultured at 37°C, 5% $CO_2$ and with 95% relative humidity in an incubator.

**Table 4-1 - Culture and passaging conditions**

| Cell Line | Culture Medium | Passaging |
|---|---|---|
| THP-1 | RPMI-1640 | Cells in suspension |
| CTV-1 | RPMI-1640 | Cells in suspension |
| K-562 | RPMI-1640 | Cells in suspension |
| GM-12878 | RPMI-1640 (including 15% FCS) | Cells in suspension |

Cell cultures were split 1:2 to 1:4 in fresh medium every 2-3 days. Adherent cells were washed once with PBS and scraped to detach the cells.

#### 4.1.1.2 Assessing Cell Number and Viability

**Required solutions and materials:**

Trypan blue solution: 0.2 % (w/v) trypan blue in 0.9% NaCl solution

Neubauer haemocytometer slide with coverslip

The number and viablity of cells was determined by Trypan blue exclusion. Cell suspensions were diluted with Trypan blue solution and immediately counted in a Neubauer haemocytometer.
The concentration of viable cells was then calculated using the following equiation:

Number of viable cells/ml (C): **$C = N \times D \times 10^4$**

N: average number of unstained cells per corner square (1 $mm^2$ containing 16 sub-squares)

D: dilution factor

#### 4.1.1.3 Freezing and Thawing Cells

| | | |
|---|---|---|
| **Freezing medium:** | 50% | RPMI 1640 |
| | 40% | FCS |
| | 10% | DMSO (Dimethylsulfoxid) |

Cells were harvested before reaching confluency and suspended at 5-10x$10^6$ cells/ml in 1 ml ice-cold freezing medium. Cells were then transferred into cryo tubes. To allow gradual freezing at a rate of 1°C/min, the cryo tubes were placed in isopropanol-filled cryo containers (Nalgene, Sigma-Aldrich, Taufkirchen, Germany) and frozen at -80°C for 24 h. For long-term storage, the tubes were transferred into liquid nitrogen (-196°C).

#### 4.1.1.4 Mycoplasma Assay

Cell lines were routinely checked for mycoplasma contamination using the MycoAlert® Mycoplasma Detection Kit according to the manufacturer’s instructions.

#### 4.1.1.5 Differentiation and Fixation of THP-1 Cells

**Required solutions and buffers:**

| | | |
|---|---|---|
| Cell Buffer Mix (CBM): | 1 ml (10 mM) | HEPES/KOH (1 M), pH 7.9 |
| | 4.25 ml (85 mM) | KCl (3 M) |
| | 200 µl (1 mM) | EDTA (500 mM, pH 8.0) |
| | Add dd$H_2O$ to 97 ml | |
| Nuclear Lysis Buffer (NL): | 5 ml (50 mM) | Tris/HCl (1M), pH 7.4 |
| | 5 ml (1%) | SDS (20%) |
| | 1.43 ml (0.5%) | Empigen BB (35%) |
| | 2 ml (10 mM) | EDTA (500 mM, pH 8.0) |
| | Add dd$H_2O$ to 97 ml | |

| | | |
|---|---|---|
| Cell Lysis Buffer L1: | 8.7 ml | CBM |
| | 100 µl (1 mM) | PMSF (100 mM) |
| | 200 µl (1x) | 50x Proteaseinhibitor Cocktail |
| | 1 ml (1%) | NP-40 (10%) |

| | | |
|---|---|---|
| Lysis buffer L2: | 9.7 ml | NL |
| | 100 µl (1 mM) | PMSF (100 mM) |
| | 200 µl (1x) | 50x Proteaseinhibitor Cocktail |

For differentiation 10-20x$10^6$ THP-1 cells in fresh culture medium were transferred into a petri dish and stimulated with $10^{-8}$ M PMA (phorbol-12-myristat-13-acetat; $10^{-4}$ M) and $10^{-7}$ M VD3 (Vitamin $D_3$; $10^{-3}$ M). Following stimulation cells were incubated for 3 d at 37°C in an incubator until cells became adherent upon differentiation.

For fixation cells were treated with 1% vol. formaldehyde (16%, Thermo Fisher Scientific, Waltham, USA) for 10 min at RT before the reaction was quenched with 1/20 volume 2.625 M glycine. After fixation cells were washed twice with ice-cold PBS (supplemented with 1 mM PMSF) directly on the plate. For lysis 1.5 ml L1 was added per plate, cells were scrapped using a cell scraper and transferred into a falcon. The cell number was determined before the cells were centrifuged using the general cell culture centrifugation program. The nuclei pellet was resuspended in 400 µl L2/10x$10^6$ cells and stored at -80°C.  Untreated THP-1 cells and other human cell lines, or primary cells respectively, were fixed in suspension and the nuclei pellets were harvested using centrifugation.

## 4.1.2 Primary Cells

### 4.1.2.1 Isolation of Monocytes

Peripheral blood mononuclear cells (PB-MNCs) were separated by leukapheresis of the blood of healthy donors (Graw et al. 1971), followed by density gradient centrifugation over Ficoll-Hypaque (Johnson et al. 1977). MO were then isolated from MNCs by counter current centrifugal elutriation (Sanderson et al., 1977). Elutriation was performed in a J6M-E centrifuge equipped with a JE 5.0 elutriation rotor and a 50 ml flow chamber (Beckman, Munich, Germany). After sterilizing the system with 6% $H_2O_2$ for 20 min, the system was washed with PBS. Following calibration at 2500 rpm and 4°C with Hanks BSS, MNCs were loaded at a flow rate of 52 ml/min. Fractions were collected and the flow through rate was sequentially increased according to **Table 4-2**.

**Table 4-2 - Elutriation parameters and cell types**

| Fraction | Volume (ml) | Flow rate (ml/min) | Main cell type contained |
|---|---|---|---|
| Ia | 1000 | 52 | platelets |
| Ib | 1000 | 57 | B and T lymphocytes, NK cells |
| IIa | 1000 | 64 | |
| IIb | 500 | 74 | |
| IIc | 400 | 82 | |
| IId | 400 | 92 | |
| III | 800 | 130 | monocytes |

MO represent the largest cells within the MNCs and are therefore mainly obtained in the last fraction. MO were >85% pure as determined by morphology and CD14 antigen expression. Low amounts of MO may also be detected in the IId fraction. MO from fraction III were centrifuged (8 min, 300xg, 4°C), resuspended in RPMI culture medium and counted. MO yields were donor dependent, typically between 10-20% of total MNCs. Supernatants of MO cultures were routinely collected and analyzed for the presence of interleukin-6 (IL-6), which was usually low, indicating that MO were not activated before or during elutriation.

### 4.1.2.2 Cultivation of Monocytes

#### 4.1.2.2.1 Dendritic Cells

Immature MO-derived dendritic cells (DCs) were generated by culturing 1x10$^6$/ml elutriated MO in RPMI containing 10% FCS, 20 U/ml recombinant human IL-4 (Promokine, Heidelberg, Germany) and 280 U/ml GM-CSF (Berlex, Seattle, USA) as described earlier (Meierhoff et al. 1998).

#### 4.1.2.2.2 Macrophages

In order to generate MAC *in vitro*, 1x10$^6$/ml MO were cultured in RPMI-1640 in the presence of 2% human pooled AB-group serum on teflon foils. For harvesting, macrophages were cooled to 4°C for 30 min and subsequently detached by carefully ‘grinding’ the teflon foils (Andreesen et al. 1983).

## 4.1.3 Transient Transfection of Mammalian Cells

### 4.1.3.1 Electroporation of Mammalian Cell Lines and Human Primary Cells

| | |
|---|---|
| **Required solutions and materials:** | RPMI (Gibco, Thermo Fisher Scientific, Waltham, USA; without phenol red)<br>Opti-MEM (Gibco, Thermo Fisher Scientific, Waltham, USA; without phenol red)<br>0.4 cm electroporation cuvettes (PeqLab, Erlangen, Germany) |
| **Setup:** | 3 square waves, 400 V, 5 ms, 1 pulse, interval |

Before transfection 1 ml of growth medium per 1x$10^6$ cells was prepared and incubated at 37°C in an incubator in the desired plate or flask. When MO were transfected, either 20 U/ml IL-4 and 280 U/ml GM-CSF was added to the culture medium to differentiate them into DCs, or 2% AB serum was added to the culture medium instead to obtain differentiation into MAC respectively. All centrifugation steps were carried out at 1000 rpm for 10 min at RT. Per transfection 1.5-15x$10^6$ cells were harvested depending on the specific experiment. The cell pellet was washed once with about 10 ml RPMI and once with about 10 ml Opti-MEM. The supernatant was removed completely and the cell pellet was resuspended in 200 µl Opti-MEM per transfection. Varying amounts of mRNA depending on the used cells and the according experiment were pipetted into the electroporation cuvette before 200 µl of the cell suspension was added. Cuvettes were checked for air bubbles and immediately placed into the electroporation chamber and pulsed. After electroporation cells were directly transferred into the pre-incubated growth medium and incubated for 40 min up to 48 h at 37°C in an incubator according to the particular downstream analysis protocol. Transfection efficiencies for each cell type were monitored at least once by transfection of GFP (green fluorescent protein) mRNA and subsequent FACS (fluorescence activated cell sorting) analysis carried out by the working group of PD Petra Hoffmann (Department of Internal Medicine III, University Hospital Regensburg).

# 4.2 General Molecular Biology

Unless otherwise noted, all protocols are based on the methods described in 'Current protocols of Molecular Biology' (Ausubel 2006) and 'Molecular cloning laboratory manual' (Sambrook 2001).

## 4.2.1 Bacterial Culture

### 4.2.1.1 Cultivation of *E.coli* strains

**Frequent used solutions:**

| | | |
|---|---|---|
| LB-medium: | 10 g | Bacto Tryptone |
| | 10 g | NaCl |
| | 5 g | Yeast extract |
| | Add $ddH_2O$ to 1000 ml, adjust pH to 7.5, autoclave | |

| | | |
|---|---|---|
| LB-agar plates: | 10 g | Bacto Tryptone |
| | 10 g | NaCl |
| | 5 g | Yeast extract |
| | 15 g | Agar |
| | Add $ddH_2O$ to 1000 ml, adjust pH to 7.5, autoclave, cool to 50°C and add the appropriate antibiotic,<br>pour the agar solution into 10 cm petri dishes and store inverted at 4°C | |

*E.coli* strains were streaked out on solid LB-agar with appropriate antibiotics (see section 3.9) and grown overnight (o/n) at 37°C. Single colonies were picked into liquid LB-medium containing the corresponding antibiotics and grown overnight at 37°C with shaking at about 200 rpm. For long-term storage, bacteria were stored at -80°C in 20% glycerol by adding 600 µl liquid culture to 200 µl of 80% glycerol.

## 4.2.2 Cloning Experiments

### 4.2.2.1 Transformation of Chemically Competent *E.coli*

**Required solutions:**

| | | |
|---|---|---|
| SOC-medium: | 20 g (2%) | Bacto Tryptone |
| | 5 g (0.5%) | Yeast extract |
| | 0.6 g (10 mM) | NaCl |
| | 0.2 g (3 mM) | KCl |
| | Add $ddH_2O$ to 1000 ml, adjust pH to 7.5, autoclave | |

| | | |
|---|---|---|
| Supplements: | 10 ml (10 mM) | $MgCl_2$ (1 M), sterile filtered |
| | 10 ml (10 mM) | $MgSO_4$ (1 M), sterile filtered |
| | 10 ml (20 mM) | Glucose (2 M), sterile filtered |

Chemically competent *E.coli* (50 µl/sample) were thawed on ice, 1-25 ng plasmid DNA was added, the suspension was mixed gently and incubated on ice for 30 min. Cells were heat-shocked in a heat block at 42°C for 90 s, immediately cooled on ice for 2 min before 250 µl pre-warmed SOC medium was added. Bacteria were incubated for at least 1 h at 37°C with shaking for recovery and 10-100 µl of the transformation were plated and incubated overnight at 37°C on LB-agar containing the appropriate antibiotic.

#### 4.2.2.2 Plasmid Isolation from *E.coli*

For plasmid isolation the NucleoSpin® Plasmid Quick Pure Kit (Macherey-Nagel, Düren, Germany) or the Monarch Plasmid Miniprep Kit (NEB, Frankfurt, Germany) were used according to manufacturer's instructions. To isolate larger amounts of ultrapure DNA (up to 100 µg) for transfection experiments e.g., plasmids were isolated using the endotoxin-free QIAGEN Plasmid Midi Kit (Hilden, Germany).

#### 4.2.2.3 Molecular Cloning

DNA fragments to be cloned were prepared by PCR from genomic DNA (gDNA) or cDNA (complementary DNA). For directional cloning, restriction sites were introduced by adding the appropriate recognition sequences to the primer sequences. Excised fragment and vector DNA was gel-purified and combined in a 10 µl ligation reaction at a 3- to 5-fold molar excess of insert to vector, using 25-50 ng of vector. Ligation was carried out overnight at 16°C with 1 U T4 DNA ligase and 1 µl 10xT4 DNA ligase buffer. 2 µl of the reaction were used to transform chemically competent *E.coli* (see section 4.2.2.1). Successful insertion of the fragment into the vector was controlled by preparing plasmid DNA from liquid cultures (see section 4.2.2.2). To control correct insertion and sequence integrity, plasmid constructs were sequenced by Thermo Fisher (Regensburg, Germany) using vector-specific primers.

#### 4.2.2.4 Gibson Assembly

The Gibson Assembly Method described 2009 by Gibson et al. is a rapid assembly method that provides directional cloning of multiple DNA fragments (up to 6) in a single reaction, without the need for specific restriction sequences. It relies on the use of an enzyme mixture consisting of a mesophilic exonuclease, a thermophilic ligase, and a high-fidelity polymerase. The system is very flexible and suitable for large DNA constructs. For the assembly reaction gBlocks® Gene Fragments (chemically synthesized, double-stranded DNA) ordered at IDT (Leuven, Belgium) were designed specific to the target of interest with overlapping sequences (20-80 bp) correlated to the desired, linearized vector sequence (see section 4.2.3.5), or if more fragments were joined, with the overlaps correlated to the vector and/or the other gBlocks® Gene Fragments used.

For one reaction 50-100 ng linearized plasmid DNA and a 2-3 fold molar excess of insert fragments was used. Reactions were set up as shown in the following table according to instructions of IDT:

**Table 4-3 - Gibson Assembly reaction setup and parameters**

| **Number of fragments including plasmid** | 2-3 Fragments | 4-6 Fragments |
|---|---|---|
| **Quantity** | 0.02-0.5 pmole ea. | 0.05-5.0 pmole ea. |
| **Gibson Assembly Master Mix (2x), NEB** | 10 µl | 10 µl |
| **ddH$_2$O** | Add to 20 µl | Add to 20 µl |

Reactions were incubated for 1 h at 50°C, resulting in the digestion of the 5’ ends of the double stranded DNA (dsDNA) by the exonuclease enzyme. The high temperature assures that the enzymes activity is rapidly degraded leaving complementary, 3’ single stranded DNA (ssDNA) ends. The resulting single stranded, complementary ends are then available to hybridize to each other, at which point the polymerase fills in missing nucleotides and the ligase finally covalently joins the fragments together.

## 4.2.3 Preparation and Analysis of DNA

### 4.2.3.1 Isolation and Quality Control of Genomic DNA

gDNA was isolated using the Qiagen Blood and Tissue Culture Kit (Hilden, Germany). Obtained gDNA concentration was determined with the NanoDrop spectrophotometer and its quality was assessed by agarose gel electrophoresis.

### 4.2.3.2 Precipitation of DNA using Polyethylene Glycol (PEG)

**Required solutions and buffers:**

| | | |
|---|---|---|
| **PEG-Mix:** | 26.2 g (26.2%) | PEG-8000 |
| | 20 ml (0.67 M) | NaOAc (3 M), pH 5.2 |
| | 660 µl (0.67 mM) | $MgCl_2$ |
| | Add ddH$_2$O to 250 ml | |

To precipitate DNA from small volumes, e.g. PCR reactions or endonuclease digestion, one volume of PEG-mix was added to the DNA-containing solution, vortexed and incubated for 15 min at RT. After centrifugation (15 min, 13000 rpm, RT), the precipitated DNA was washed by carefully adding 100 µl 100% EtOH to the tube wall opposite of the (often invisible) pellet. Following centrifugation (15 min, 13000 rpm, RT), the supernatant was carefully removed. The pellet was air-dried for 5 min and resuspended in ddH$_2$O in half to three-quarters of the initial volume.

### 4.2.3.3 Purification of DNA using Phenol-Chloroform Extraction and Ethanol Precipitation

DNA containing solutions and linearized DNA plasmids for *in vitro* synthesis of mRNA (see sections 4.2.3.5 and 4.2.5.2) were vigorously mixed with 1 Volume (V) Phenol-Chloroform-Isoamylalcohol (25:24:1, pH 8). After centrifugation (5 min, 13000 rpm, RT) the upper aqueous phase containing the DNA was carefully transferred to a new tube. If desired, this step was repeated until no protein was visible anymore between the organic and aqueous phase. The precipitation was repeated twice with 1V of Chloroform p.a. (Merck, Darmstadt, Germany) to remove residual phenol. After extraction the DNA-containing, aqueous phase was transferred into a new RNase-free Eppendorf cup and precipitated with 0.5V 5 M $NH_4OAc$ (pH 5.2) and 2.5V 100% EtOH for at least 1 h at -80°C. Precipitated DNA was then washed with 70% ice-cold EtOH and the air-dried pellet was dissolved in RNase-free water at a concentration of 1 µg/µl. Purified linearized DNA plasmids were stored at 4°C for later *in vitro* transcription.

### 4.2.3.4 Agarose Gel Electrophoresis

**Required solutions and buffers:**

| | | |
|---|---|---|
| TAE (50x): | 242.3 g (2 M) | Tris |
| | 20.5 g (250 mM) | NaOAc/HOAc, pH 7.8 |
| | 18.5 g (0.5 M) | EDTA, pH 8.0 |
| | Add $ddH_2O$ to 1000 ml | |
| DNA loading dye (DNA LD 5x): | 500 µl (50 mM) | Tris (1 M), pH 7.8 |
| | 500 µl (1%) | SDS (20%) |
| | 1 ml (50 mM) | EDTA (0.5 M), pH 8.0 |
| | 4 ml (40%) | Glycerol (100%) |
| | 10 g (1%) | Bromphenol blue |
| | Add $ddH_2O$ to 10 ml, store at 4°C | |
| 1% Agarose: | 1 g (1%) | Agarose (Biozym) |
| | Add 1x TAE to 100 ml and heat in a microwave until agarose is completely dissolved, cool to 50°C and add 2.5 µl Ethidium bromide (10 mg/ml) | |

The required amount of agarose as determined according to **Table 4-4** was added to the corresponding amount of 1xTAE. The slurry was heated in a microwave until the agarose was completely dissolved. Ethidium bromide was added after cooling the solution to 50-60°C. The gel was cast, mounted in the electrophoresis tank and covered with 1xTAE. DNA-containing samples were diluted 4:1 with DNA LD 5x, mixed and loaded into the slots of the gel. Depending on the size and the desired resolution, gels were run at 40-120 volt for 30 min to 3 h.

**Table 4-4 - Agarose concentration for different separation ranges**

| Efficient range of separation (kb) | % agarose in gel |
|---|---|
| 0.1 – 2 | 2.0 |
| 0.2 – 3 | 1.5 |
| 0.4 – 6 | 1.2 |
| 0.5 – 7 | 0.9 |
| 0.8 – 1.0 | 0.7 |
| gDNA | 0.5 |

#### 4.2.3.5 Restriction Endonuclease Digestion

To verify the presence and orientation of plasmid-inserts, to clone insert DNA of interest into plasmids, or to linearize plasmid DNA for *in vitro* transcription (see section 4.2.5.2), the plasmid and insert DNA respectively were digested with the appropriate restriction endonucleases. Used enzymes and their buffers were purchased from Sigma-Aldrich (Taufkirchen, Germany) or New England Biolabs (NEB, Frankfurt, Germany). 1 µg DNA was digested with 1 U enzyme in a 10 -50 µl reaction by incubation at 37°C for 1 h.

#### 4.2.3.6 Dephosphorylation of DNA with Alkaline Phosphatase

To prevent self-ligation, digested vectors were treated with 1 U of AP (calf intestinal alkaline phosphatase) at 37°C for 30 min prior to gel extraction.

#### 4.2.3.7 Purification of DNA Fragments by Gel Extraction

DNA fragments were purified by running on an ethidium bromide-containing agarose gel (see 4.2.3.4). The band containing the fragment of interest was excised under UV illumination. Fragments were then purified by gel extraction using the Monarch DNA Gel Extraction Kit (NEB, Frankfurt, Germany) following the manufacturer’s instructions.

#### 4.2.3.8 Ligation of DNA Fragments

For ligation of digested plasmid DNA and the particular PCR amplicon or cDNA, a 3-fold molar excess of insert was used in contrast to the vector DNA. 200 ng of vector DNA as well as 1 U T4 DNA ligase (Promega, Madison, USA) has been used in a 10-20 µl reaction. Ligation was carried out for 2h or o/n in a thermocycler at 16°C.

#### 4.2.3.9 DNA Sequencing and Sequence Analysis

DNA sequencing was done by Thermo Fisher (Regensburg, Germany) based on the Sanger sequencing method. Sequence files were analyzed and aligned with GENtle or with the BLAST function of the UCSC genome browser (see section 3.13).

### 4.2.4 Polymerase Chain Reaction (PCR)

#### 4.2.4.1 Analytical PCR

The polymerase chain reaction (PCR) allows *in vitro* synthesis of large amounts of DNA by primed, sequence-specific polymerization of nucleotide triphosphates, catalyzed by DNA polymerase (Mullis et al., 1986). PCR reactions were generally performed in thick 0.5 ml PCR tubes with a reaction volume of 20-100 µl in a MJ research PTC 200 thermocycler (Biozym, Hessisch Oldendorf, Germany) or a Mastercycler Nexus M2 (Eppendorf, Hamburg, Germany). The calculated temperature feature or the safe mode were used accordingly to decrease temperature hold times. Additionally the lid was heated to 105°C to prevent vaporization of the sample. The nucleotide sequences of the utilized primers are given in section 3.7. The primer annealing temperatures usually varied between 57°C and 65°C and the Phusion$^{TM}$ High Fidelity II DNA polymerase (Thermo Fisher Scientific, Waltham, USA) was used according to the following table:

**Table 4-5 - Reaction setup for analytical PCR**

| Component | Volume [µl] | Final Concentration |
|---|---|---|
| $H_2O$ | ad 50 µl | |
| Template DNA | x µl | |
| 5x Phusion HF buffer | 10 µl | 1 x |
| 10 mM dNTPs | 1 µl | 200 µM each |
| Primer forward (10 µM) | 1 µl | 0.2 µM |
| Primer reverse (10 µM) | 1 µl | 0.2 µM |
| Phusion HF Polymerase (2 U/µl) | 0.5 µl | 1 U |

General parameter settings for an analytical PCR are summarized in **Table 4-6**.

**Table 4-6 - Reaction parameter for analytical PCR**

| Stage | Temperature | Time | Cycles |
|---|---|---|---|
| Initial melting | 95°C | 2 min | 1 |
| Melting | 95°C | 15 s | 25 - 35 |
| Annealing | 65°C | 15 s | |
| Extension | 72°C | 60 s | |
| Final extension | 72°C | 5 min | 1 |
| Hold | 4°C | ∞ | |

### 4.2.4.2 Quantitative Real Time PCR (RT-qPCR)

Quantitative Real Time PCR (RT-qPCR) was used for quantification and analysis of DNA after ChIP (4.2.6). PCR reactions were performed using the QuantiFast SYBR Green Kit from Qiagen (Hilden, Germany) in 96-well plates adapted to the Eppendorf Realplex Mastercycler EpGradient S (Eppendorf, Hamburg, Germany). The relative amount of amplified DNA is measured through the emission of light by the SYBR green dye, when it is intercalated in dsDNA. Typical reaction setups and parameters are listed in **Table 4-7** and **Table 4-8**.

**Table 4-7 - Reaction setup for real time PCR**

| Component | Volume [µl] | Final Concentration |
|---|---|---|
| SYBR Green mix (2x) | 5 µl | 1x |
| $H_2O$ | 2 µl | |
| Primer forward (10 µM) | 0.5 µl | 0.5 µM |
| Primer reverse (10 µM) | 0.5 µl | 0.5 µM |
| DNA | 2 µl | |

**Table 4-8 - Reaction parameter for real time PCR**

| Stage | Temperature | Time | Cycles |
|---|---|---|---|
| Initial melting | 95°C | 5 min | 1 |
| Melting | 95°C | 8 s | 45 |
| Annealing & Extension | 60°C | 20 s | |
| Melting | 95°C | 15 s | 1 |
| Annealing & Extension | 60°C | 15 s | |
| Melting curve | | 10 - 20 min | |
| | 95°C | 15 s | |

To calculate amplification efficiency, a dilution series (1:10; 1:50; 1:100, 1:1000) of a suitable sample was additionally measured for each primer pair. The Realplex software automatically calculated DNA amounts based on the generated slope and intercept. Specific amplification was controlled by melting-curve analysis and data were imported and processed in Microsoft Excel 2010 or 2013. All samples were measured in duplicates and normalized to the input or a specific control region.

## 4.2.5 Preparation and Analysis of RNA

### 4.2.5.1 Isolation of Total RNA

Total cellular RNA was isolated using the Qiagen RNeasy Midi, Mini or Micro Kit (Hilden, Germany) depending on the available number of cells according to manufacturer's instructions and always including a DNase treatment. Concentration of purified total RNA was determined with the NanoDrop spectrophotometer and the quality was assessed by using the Agilent TapeStation (Böblingen, Germany) according to the manufacturer's instructions.

#### 4.2.5.2 *In Vitro* Synthesis of Capped mRNA

For *in vitro* synthesis of capped, poly-adenylated mRNA from purified, linearized DNA plasmids (see 4.2.3.3 and 4.2.3.5) under control of a T7 promotor, the mMESSAGE mMACHINE T7 Ultra Kit from Thermo Fisher Scientific (Waltham, USA) was used according to manufacturer's instructions. Briefly, the T7 2x NTP/ARCA and 10x T7 Reaction Buffer were thawed at RT to prevent precipitation. After thawing T7 2x NTP/ARCA was kept on ice. The following components were pipetted according to the listed order at RT:

**Table 4-9 - Reaction setup for mRNA synthesis**

| Component | Amount |
|---|---|
| $H_2O$ (Nuclease free) | x µl (ad to 20 µl) |
| T7 2x NTP/ARCA | 10 µl |
| 10x T7 Reaction Buffer | 2 µl |
| Linear DNA template (1 µg) | x µl |
| T7 Enzyme Mix | 2 µl |

The reaction was mixed carefully and depending on the size of the resulting transcript incubated for 1-3 h in a heat block at 37°C. After incubation 1 µl TURBO DNase was added and another 30 min of incubation at 37°C followed to digest remaining plasmid DNA. Before poly (A)-tailing (**Table 4-10**) 1 µl of the reaction was saved to control mRNA synthesis on the Agilent TapeStation (Böblingen, Germany). Following solutions were mixed with the above reaction for poly (A)-tailing:

**Table 4-10 - Reaction setup for poly (A)-tailing**

| Component | Amount |
|---|---|
| mMessage mMachine T7 Ultra reaction | 20 µl |
| $H_2O$ (Nuclease free) | 36 µl |
| 5x E-PAP Buffer | 20 µl |
| 25 mM $MnCl_2$ | 10 µl |
| ATP solution | 10 µl |

Additional 4 µl E-PAP enzyme was mixed with the reaction and the mixture was incubated for another 1 h at 37°C yielding a poly (A)-tail between 50-100 bases. Synthesized transcripts were then purified using the Qiagen RNeasy Mini Kit (Hilden, Germany) following manufacturer's instructions with described modifications. The samples were mixed with 350 µl RLT buffer before 250 µl 100% EtOH was added and the mixture was directly pipetted onto a spin column. Centrifugation was carried out at 11000 rpm for 30 sec at RT. Columns were washed once with 500 µl RW1 and twice with 500 µl RPE. Collection tubes were changed in between and the empty column was centrifuged 1 min at 13000 rpm to remove residual EtOH. mRNA transcripts were eluted in 50 µl RNase-free water (mMESSAGE mMACHINE T7 Ultra Kit). The concentration of the transcripts was determined using the NanoDrop spectrophotometer. Quality of transcribed mRNA was checked on the Agilent TapeStation (Böblingen,

Germany) according to manufacturer’s instructions. Good quality mRNA was then used for transfection of mammalian cells (see section 4.1.3).

### 4.2.6 Chromatin Immunoprecipitation (ChIP)

**Required solutions and buffers:**

| | | |
|---|---|---|
| Formaldehyde, methanol-free: | 16% (w/v) | Thermo Fisher Scientific |
| Glycine (20x): | 9.85 g (2,625 M) | Glycine |
| | Add ddH$_2$O to 50 ml | |
| Cell Buffer Mix (CBM): | 1 ml (10 mM) | HEPES/KOH (1 M), pH 7.9 |
| | 4.25 ml (85 mM) | KCl (3 M) |
| | 200 µl (1 mM) | EDTA (500 mM, pH 8.0) |
| | Add ddH$_2$O to 97 ml | |
| | Add just prior to use to 1 ml of CBM: | |
| | 10 µl (1 mM) | PMSF (100 mM) |
| | 20 µl (1x) | 50x Proteaseinhibitor cocktail |
| Nuclear Lysis Buffer (NL): | 5 ml (50 mM) | Tris/HCl (1M), pH 7.4 |
| | 5 ml (1%) | SDS (20%) |
| | 1.43 ml (0.5%) | Empigen BB (35%) |
| | 2 ml (10 mM) | EDTA (500 mM, pH 8.0) |
| | Add ddH$_2$O to 97 ml | |
| | Add just prior to use to 1 ml of NL: | |
| | 10 µl (1 mM) | PMSF (100 mM) |
| | 20 µl (1x) | 50x Proteaseinhibitor cocktail |
| Dilution Buffer (DB): | 2 ml (20 mM) | Tris/HCl (1 M), pH 7.4 |
| | 2 ml (100 mM) | NaCl (5 M) |
| | 400 µl (2 mM) | EDTA (500 mM, pH 8.0) |
| | 5 ml (0.5%) | TritonX-100 (10%) |
| | Add ddH$_2$O to 97 ml | |
| | Add just prior to use to 1 ml of DB: | |
| | 10 µl (1 mM) | PMSF (100 mM) |
| | 20 µl (1x) | 50x Proteaseinhibitor cocktail |
| Wash Buffer I (WBI): | 2 ml (20 mM) | Tris/HCl (1 M), pH 7.4 |
| | 3 ml (150 mM) | NaCl (5 M) |
| | 500 µl (0,1%) | SDS (20%) |
| | 10 ml (1%) | Triton X-100 (10%) |
| | 400 µl (2 mM) | EDTA (500 mM, pH 8.0) |
| | Add ddH$_2$O to 100 ml | |

| | | |
|---|---|---|
| Wash Buffer II (WBII): | 2 ml (20 mM) | Tris/HCl (1 M), pH 7.4 |
| | 10 ml (500 mM) | NaCl (5 M) |
| | 10 ml (1%) | TritonX-100 (10%) |
| | 400 µl (2 mM) | EDTA (500 mM, pH 8.0) |
| | Add ddH$_2$O to 100 ml | |
| Wash Buffer III (WBIII): | 1 ml (10 mM) | Tris/HCl (1 M), pH 7.4 |
| | 10 ml (250 mM) | LiCl (5 M) |
| | 10 ml (1%) | NP-40 (10%) |
| | 10 ml (1%) | Deoxycholat (10%) |
| | 200 µl (1 mM) | EDTA (500 mM, pH 8.0) |
| | Add ddH$_2$O to 100 ml | |
| TE Buffer: | 1 ml (10 mM) | Tris (1 M), pH 8.0 |
| | 0.2 ml (1 mM) | EDTA (500 mM, pH 8.0) |
| | Add ddH$_2$O to 100 ml | |
| Elution Buffer (EB): | 500 µl (0.1 M) | NaHCO$_3$ (1 M) |
| | 250 µl (1%) | SDS (20%) |
| | Add ddH$_2$O to 5 ml | |
| Sepharose Cl-4B: | 50 µl/IP | |
| nProtein A/G Sepharose: | 50 µl/IP | |
| | wash 3x with TE, pH 8.0, before use | |

Chromatin immunoprecipitation is routinely used to determine whether particular proteins are associated with a specific genomic region in living cells or tissues. The method is based on the principle that formaldehyde reacts with the primary amines of the amino acids of the associated proteins and the bases of gDNA, resulting in a covalent cross-link between adjacent proteins and DNA. Fixation and lysis of cells was carried out as described in section 4.1.1.5 for adherent and suspension cells respectively. For certain TF ChIPs however, a double-crosslink procedure was applied. In brief, cells were harvested, washed once with cold PBS and crosslinked with 2 mM of DSG in PBS (disuccinimidyl glutarate; Thermo Fisher Scientific, Waltham, USA) for 30 min at RT. This induced a protein-protein crosslink needed for TFs, which bind to the DNA as dimers. After the cells were directly fixed with 1% formaldehyde as already described above and further processed as described next. Cross-linked chromatin was sheared to an average DNA fragment size of around 200 – 500 bp using a Branson Sonifier 250 (Danbury, USA). After centrifugation, 5% of the lysate was saved as input control and 20 µl were kept for agarose gel analysis of the sonicated chromatin. After preclearing with 50 µl Sepharose CL-4B beads (GE Healthcare, Munich, Germany) (blocked with 0.5% BSA and 20 µg glycogen) for 2 h, chromatin samples were immunoprecipitated overnight with 2.5 µg of the appropriate antibody. Before precipitation, nProtein A Sepharose beads (GE Healthcare, Munich, Germany) or nProtein G Sepharose beads (GE Healthcare, Munich, Germany), depending on the source of the antibody epitope, were blocked with 0.5% BSA and 20 µg glycogen o/n at 4°C. Complexes were then recovered by a 2 h incubation with the pre-blocked beads at 4°C, where the Fc-tail of the used antibody binds to the nProtein Sepharose beads. For washing beads were treated twice with WB I, WB II and WB II, and three times with TE buffer (400 µl each). DNA was then eluted in 200 µl EB and crosslinks were reverted by adding 10 µl 5 M NaCl and incubation at 65°C o/n. Obtained DNA was then treated with 7 µl RNase (10 µg/µl) and 5 µl Proteinase K (20 µg/µl) for 1 h each and purified using the Monarch PCR & DNA Cleanup Kit (NEB, Frankfurt, Germany) according to the manufacturer's instructions, except that the samples were incubated for 30 min with binding buffer and eluted in 43 µl EB. Enrichment of specific DNA fragments was determined by quantitative RT-qPCR on the Realplex Mastercycler as described above (see section 4.2.4.2) or directly used for next generation sequencing (NGS) analysis (see section 4.2.7).

### 4.2.7 Generation of DNA Libraries for Next Generation Sequencing (NGS)

For sequence-based analysis of ChIP experiments, adaptors with specific barcodes were ligated to enriched DNA fragments. Library preparation was either carried out as described below or using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB, Frankfurt, Germany) according to manufacturer's instructions. Homemade library preparation was carried out in several steps. All enzymes, except Phusion High-Fidelity Polymerase (Thermo Fisher Scientific, Waltham, USA), were purchased from Enzymatics (Beverly, USA). Initially ds-DNA ends were polished for 30 min at 20°C using T4 DNA-Polymerase, T4 Polynucleotide Kinase and Klenow fragment resulting in blunt ended DNA. To purify blunt ended DNA fragments either the MinElute PCR purification Kit (Qiagen, Hilden, Germany) or the Monarch PCR & DNA Cleanup Kit (NEB, Frankfurt, Germany) was used. Purified 3'-DNA ends were poly-adenylated for 30 min at 37°C using the Klenow fragment harboring a 3'-5'-exonuclease activity. Polyadenylation of DNA in turn facilitates adaptor ligation, because they harbor complementary T-overhangs. DNA was then purified twice with magnetic beads in a ratio between sample and beads of 1:1.1 (Agencourt AMPure XP; Beckman Coulter, Krefeld, Germany). After purification the adaptors were ligated for 10 min at 30°C. DNA was purified again for two times using magnetic beads (ratio 1:1.1), before an analytical four cycle PCR with subsequent bead purification was carried out (ratio 1:1.8). For size selection of DNA fragments the Caliper LabChip XT together with the LabChip XT DNA 750 Assay Kit (Perkin Elmer, Hamburg, Germany) was used, yielding an average DNA fragment size of about 275 bp (+/- 15%). Size selected DNA was again amplified using a 12 cycle PCR or a 16 cycle PCR was directly carried out before size selection on the Caliper device. In a final step, DNA was purified two times with magnetic beads (ratio 1:1.1) and eluted in 16 µl of EB (Qiagen, Hilden, Germany or NEB, Frankfurt, Germany). Quality of DNA libraries was analyzed using the Agilent TapeStation (Böblingen, Germany) using a high sensitivity DNA screen tape. Library concentrations were assessed with the Qubit 2.0 flurometer (Thermo Fisher Scientific, Waltham, USA). Sequencing was carried out at the Biomedical Sequencing Facility (BSF) of the Research Center for Molecular Medicine of the Austrian Academy of Sciences (Ce-M-M, Vienna, Austria) using an Illumina HiSeq 3000/4000 sequencer.

### 4.2.8 Assay for Transposase-Accessible Chromatin Sequencing (ATACseq)

**Required solutions and buffers:**

| | | |
|---|---|---|
| Digitonin: | 1% (w/v) | Promega |
| DNase: | 35.5 U/µl | Sigma-Aldrich |
| NP-40: | 10% (w/v) | Sigma-Aldrich |
| Tween-20: | 10% (w/v) | Sigma-Aldrich |
| Resuspension Buffer (RSB): | 0.5 ml (10 mM) | Tris/HCl (1 M), pH 7.4 |
| | 0.1 ml (10 mM) | NaCl (5 M) |
| | 0.15 ml (3 mM) | $MgCl_2$ (1 M) |
| | Add $ddH_2O$ to 50 ml | |
| 3x RSB: | 970 µl | RSB |
| | 10 µl (0.1%) | NP-40 (10%) |
| | 10 µl (0.1%) | Tween-20 (10%) |
| | 10 µl (0.01%) | Digitonin (1%) |
| RSB-T: | 990 µl | RSB |
| | 10 µl (0.1%) | Tween-20 (10%) |
| Transposition Mix (for 1 rxn): | 25 µl (1x) | 2x TD buffer (Nextera, Illumina) |
| | 2.5 µl (100 nM) | transposase (Nextera, Illumina) |
| | 16.5 µl | PBS |
| | 0.5 µl (0.01%) | Digitonin (1%) |
| | 0.5 µl (0.1%) | Tween-20 (10%) |
| | Add sterile $ddH_2O$ to 50 µl | |

ATACseq is used to map regulatory landscapes that control gene expression. It was essentially carried out as described 2017 by Corces et al. in Nature Methods. Briefly, prior to transposition the viability of the cells was assessed. For samples with 5-15% dead cells, cells were treated in culture medium with DNase at a final concentration of 200 U/ml for 30 min at 37°C. After the treatment, cells were washed twice with ice-cold PBS and cell viability as well as the cell count was determined. For each ATAC reaction 50.000 cells were aliquoted into a new tube, spun down at 500×g for 5 min at 4°C, before the supernatant was discarded completely. The cell pellet was resuspended in 50 µl of 3x RSB and incubated on ice for 3 min to lyse the cells. Lysis was washed out with 1 ml of RSB-T and the tube was inverted 3 times. Nuclei were pelleted at 500×g for 10 min at 4°C in a fixed angle centrifuge. The supernatant was discarded carefully and the cell pellet was resuspended in 50 µl of transposition mixture by pipetting up and down 6 times. This reaction was incubated at 37°C for 30 min with mixing (1000 rpm), before DNA was purified using the Monarch PCR & DNA Cleanup Kit (NEB, Frankfurt, Germany) according to manufacturer's instructions. Purified DNA was eluted in 20 µl EB and 10 µl purified sample was objected to PCR amplification for Illumina platforms using Nextera XT i7- and i5-index primers (Illumina, San Diego, USA). Setups for a typical PCR Master Mix and the according PCR program are shown below:

**Table 4-11 - Reaction setup for PCR amplification of transposed DNA fragments**

| Component | Volume [µl] | Final Concentration |
|---|---|---|
| 5x HF buffer | 10 µl | 1x |
| $H_2O$ | 10 µl | |
| dNTPs (10 mM) | 1.5 µl | 0.3 mM |
| Betaine (5 M, Sigma-Aldrich) | 13 µl | 1.3 M |
| Nextera XT i7 index (25 µM, Illumina) | 2.5 µl | 1.25 µM |
| Nextera XT i5 index (25 µM, Illumina) | 2.5 µl | 1.25 µM |
| Phusion HF Polymerase (2 U/µl, Thermo Fisher) | 0.5 µl | 1 U |
| DNA | 10 µl | |

**Table 4-12 - Reaction parameters for ATACseq-PCR**

| Stage | Temperature | Time | Cycles |
|---|---|---|---|
| Initial melting | 72°C | 5 min | 1 |
| Initial denaturation | 98°C | 30 s | 1 |
| Denaturation | 98°C | 10 s | 11 |
| Annealing | 63°C | 30 s | |
| Extension | 63°C | 1 min | |
| Final Extension | 72°C | 1 min | 1 |
| Hold | 4°C | forever | - |

Purification and size selection of the amplified DNA was done with magnetic beads (Agencourt AMPure XP; Beckman Coulter, Krefeld, Germany). For purification ratio of sample to beads was 1:1.8, whereas for size selection ratio was set to 1:0.55. Purified samples were eluted in 15 µl of EB and quality of DNA libraries was analyzed using the Agilent TapeStation (Böblingen, Germany) using a high sensitivity DNA screen tape. Library concentrations were assessed with the Qubit 2.0 flurometer (Thermo Fisher Scientific, Waltham, USA). Sequencing was carried out at the BSF (Ce-M-M, Vienna, Austria) using an Illumina HiSeq 3000/4000 sequencer or onsite using an Illumina NextSeq 550 sequencer according to Illumina's instructions.

### 4.2.9 RNA sequencing (RNAseq)

To analyze gene expression profiles of various cell types under several conditions, for example transfected vs. un-transfected cells, RNAseq was utilized. Generation of dsDNA libraries for Illumina sequencing out of total cellular RNA (see section 4.2.5.1) was carried out using the ScriptSeq™ Complete Kit (Human/Mouse/Rat) – Low Input from Epicentre (Chicago, USA). Typically, 1 µg of DNA-free RNA was used for each reaction. In a first step of this protocol rRNA (ribosomal RNA) is depleted using the Ribo-Zero™ rRNA removal reagents included in the kit. rRNA-free RNA is then converted into cDNA, which is 3'-terminal tagged and used for PCR amplification and library purification. The quality of DNA libraries was analyzed using the Agilent TapeStation (Böblingen, Germany) using a high sensitivity DNA screen tape. Library concentrations were assessed with the Qubit 2.0 flurometer (Thermo Fisher Scientific, Waltham, USA). Sequencing was carried out at the BSF (Ce-M-M, Vienna, Austria) using an Illumina HiSeq 3000/4000 sequencer or onsite using an Illumina NextSeq 550 sequencer according to Illumina's instructions.

### 4.2.10 Targeted Bisulfite Amplicon sequencing (TBSAseq)

For demethylation analysis of CTV-1 cells with decitabine (DAC) TBSAseq was used. DAC is an epigenetic modifier that inhibits DNA methyltransferase activity, which results in DNA demethylation (Mund et al. 2011). 0.2 µg – 1 µg gDNA was used as starting material for bisulfite conversion. Therefore, the EZ DNA Methylation Kit from Zymo Research (Irvine, USA) was used according to manufacturer's instructions. The kit is based on a three-step reaction that takes place between cytosine and sodium bisulfite, where unmethylated cytosine is converted into uracil. Purified, converted DNA was then used for PCR amplification for downstream analysis of the demethylation status. In a first PCR the modified region of interest (300 – 400 bp) was amplified using a specific primer (see section 3.7.1), which was designed using bisulfite converted gDNA as a template. This primer also inserted the binding sites for the Illumina indices, to be able to amplify the targets for downstream NGS analysis. Reaction setup and parameters for the first PCR of the TBSAseq procedure are shown in **Table 4-13** and **4-14**.

**Table 4-13 - Reaction setup for PCR amplification of bisulfite converted DNA**

| Component | Volume [µl] | Final Concentration |
|---|---|---|
| 10x PCR buffer | 0.5 µl | 1x |
| $H_2O$ | 1.42 µl | |
| dNTP Mix (25mM, Agena) | 0.04 µl | 200 µM |
| Primer Mix (0.5 µM each) | 2 µl | 0.2 µM each |
| HS Taq Polymerase (5 U/µl, Roche) | 0.04 µl | 0.2 U |
| DNA (10 ng/µl) | 1 µl | 10 ng |

**Table 4-14 - Reaction parameters for first TBSAseq-PCR**

| Stage | Temperature | Time | Cycles |
|---|---|---|---|
| HS Activation | 94°C | 4 min | 1 |
| Denaturation | 94°C | 20 s | 23 |
| Annealing | 56°C | 30 s | |
| Extension | 72°C | 1 min | |
| Final Extension | 72°C | 3 min | 1 |
| Hold | 4°C | forever | - |

Amplified DNA was purified using the Monarch PCR & DNA Cleanup Kit (NEB, Frankfurt, Germany) according to manufacturer's instructions. Purified DNA was eluted in 10 µl EB and the first PCR product was objected to a second PCR amplification for Illumina platforms using Nextera XT i7- and i5-index primers (Illumina, San Diego, USA). Setups for the second PCR Master Mix and the used PCR program are shown below:

**Table 4-15 - Reaction setup for PCR amplification of amplified bisulfite converted DNA**

| Component | Volume [µl] | Final Concentration |
|---|---|---|
| 10x Advantage 2 PCR buffer | 5 µl | 1x |
| $H_2O$ | 29 µl | |
| dNTPs (10 mM) | 1 µl | 0.2 mM |
| Nextera XT i7 index (25 µM, Illumina) | 2 µl | 1 µM |
| Nextera XT i5 index (25 µM, Illumina) | 2 µl | 1 µM |
| 50x Advantage 2 Polymerase Mix (Clontech) | 1 µl | 1x |
| DNA | 10 µl | |

**Table 4-16 - Reaction parameters for second TBSAseq-PCR**

| Stage | Temperature | Time | Cycles |
|---|---|---|---|
| HS Activation | 95°C | 1 min | 1 |
| Denaturation | 95°C | 30 s | 22 |
| Annealing & Extension | 68°C | 70 s | |
| Final Extension | 72°C | 7 min | 1 |
| Hold | 8°C | forever | - |

Amplicons were purified with the Monarch PCR & DNA Cleanup Kit (NEB, Frankfurt, Germany) according to manufacturer's instructions and eluted in 18 µl of EB. The quality of the amplicons was analyzed using the Agilent TapeStation (Böblingen, Germany) using a high sensitivity DNA screen tape. Library concentrations were assessed with the Qubit 2.0 flurometer (Thermo Fisher Scientific, Waltham, USA). Sequencing was carried out at the Faculty of Biochemistry of the University of Regensburg (Department of Prof. Dr. Meister) using an Illumina MiSeq sequencer.

## 4.3 Protein Biochemical Methods

### 4.3.1 Preparation of Whole Cell Lysates

**Required solutions and buffers:**

| | | |
|---|---|---|
| SDS Sample Buffer (2x): | 10 ml (150 mM) | Tris (1.5 M), pH 6.8 |
| | 6 ml (1.2%) | SDS (20%) |
| | 30 ml | Glycerol |
| | 15 ml | β-mercaptoethanol |
| | 1.8 mg | Bromophenol blue |
| | Add ddH$_2$O to 100 ml; aliquot in 10 ml stock solution and store at -20°C, store working solution at 4°C | |

1.5x10$^6$ (can be scaled up and down as needed) suspension cells were harvested by centrifugation. Cells were washed once with cold PBS. The supernatant was removed completely after washing and the cell pellet was resuspended in 100 µl 2x SDS sample buffer. After the samples were immediately incubated for 10 min at 95°C in a heating block and vortexed directly after incubation. Lysates were stored at -20°C for subsequent SDS-PAGE (see 4.3.3) or Western Blot (see 4.3.5) analysis. Adherent cells were washed with cold PBS directly on the plate. Cells were then scrapped in 2x SDS sample buffer from the plate and collected into Eppis. After samples were treated as stated above.

### 4.3.2 Preparation of Nuclear Extracts

**Required solutions and buffers:**

| | | |
|---|---|---|
| Triton X-100 (10%): | 1 ml (10%) | Triton X-100 |
| | Add ddH$_2$O to 10 ml | |
| Hypotonic Buffer: | 5 ml (1%) | Triton X-100 (10%) |
| | 5.48 g (320 mM) | Sucrose |
| | 500 µl (10 mM) | HEPES (1 M), pH 7.5 |
| | 250 µl (5 mM) | MgCl$_2$ (1 M) |
| | Add ddH$_2$O 50ml | |
| | Add just prior to use to | 1 ml of Sucrose/1%Triton: |
| | 10 µl (1 mM) | PMSF (100 mM) |
| | 20 µl (1x) | 50x Proteaseinhibitor cocktail |
| | 40 µl (1x) | 25x Phosphataseinhibitor cocktail |
| Sucrose Wash Buffer: | 5.48 g (320 mM) | Sucrose |
| | 500 µl (10 mM) | HEPES (1 M), pH 7.5 |
| | 250 µl (5 mM) | MgCl$_2$ (1 M) |
| | Add ddH$_2$O 50ml | |

| | | |
|---|---|---|
| | Add just prior to use to 1 ml of Sucrose Wash Buffer: | |
| | 10 µl (1 mM) | PMSF (100 mM) |
| | 20 µl (1x) | 50x Proteaseinhibitor cocktail |
| | 40 µl (1x) | 25x Phosphataseinhibitor cocktail |

| | | |
|---|---|---|
| Nuclear Sonication Buffer: | 2.5 ml (50 mM) | Tris/HCl (1 M), pH 8.0 |
| | 1.5 ml (150 mM) | NaCl (5 M) |
| | 100 µl (1 mM) | EDTA (0.5 M), pH 8.0 |
| | 5 ml (10%) | Glycerol (100%) |
| | Add ddH$_2$O 50ml | |
| | | |
| | Add just prior to use to 1 ml of Nuclear Sonication Buffer: | |
| | 10 µl (1 mM) | PMSF (100 mM) |
| | 20 µl (1x) | 50x Proteaseinhibitor cocktail |
| | 40 µl (1x) | 25x Phosphataseinhibitor cocktail |

| | | |
|---|---|---|
| SDS Sample Buffer (2x): | 10 ml (150 mM) | Tris (1.5 M), pH 6.8 |
| | 6 ml (1.2%) | SDS (20%) |
| | 30 ml | Glycerol |
| | 15 ml | β-mercaptoethanol |
| | 1.8 mg | Bromophenol blue |
| | Add ddH$_2$O to 100 ml; aliquot in 10 ml stock solution and store at -20°C, store working solution at 4°C | |

5x10$^6$ - 10x10$^6$ (can be scaled up and down as needed) suspension cells were harvested by centrifugation. Cells were washed once with cold PBS (including 1mM PMSF). The supernatant was removed completely after washing and the cell pellet was resuspended in 1 ml hypotonic buffer. Cells were homogenized on ice with 10 to 15 strokes in a Dounce Tissue Grinder (Thermo Fisher Scientific, Waltham, USA) and the suspension was transferred to a new tube. Nuclei were pelleted at 500×g for 5 min at 4°C in a fixed angle centrifuge. The supernatant, containing the cytosolic fraction, was transferred to  a new tube and stored at -80°C for subsequent SDS-PAGE (see 4.3.3) and Western Blot (see 4.3.5) analysis. Pellets were washed twice with 500 µl of sucrose wash buffer and centrifuged at 500×g for 5 min at 4°C, before 450 µl of nuclear sonication buffer was added to the cell pellet and the cells were sonicated three times for 10 s. All sonication steps were carried out with a constant duty cycle and output control 2 using a Branson Sonifier 250 (Danbury, USA). After sonication 250 U Benzonase Nuclease (Sigma-Aldrich, Taufkirchen, Germany; 250 U/µl) was added and the cell suspension was incubated for 30 min at 4°C. Suspensions were centrifuged at 11.000×g for 15 min at 4°C and the nuclear extract was transferred to a new tube and stored at -80°C for subsequent SDS-PAGE (see 4.3.3) and Western Blot (see 4.3.5) analysis. The protein concentration of cytosolic and nuclear lysates was assessed with the Qubit 2.0 flurometer (Thermo Fisher Scientific, Waltham, USA). In general 20 µg of cytosolic and nuclear extracts in 2x SDS sample buffer boiled for 10 min at 95°C were loaded for Western Blot analysis.

### 4.3.3 Discontinuous SDS-PAGE

**Required solutions and buffers:**

| | | |
|---|---|---|
| Separating Gel Buffer: | 90.83 g (1.5 M) | Tris/HCl, pH 8.8 |
| | Add $ddH_2O$ to 500 ml | |
| Stacking Gel Buffer: | 30 g (0.5 M) | Tris/HCl, pH 6.8 |
| | Add $ddH_2O$ to 500 ml | |
| SDS (10%): | 100 g (10%) | SDS |
| | Add $ddH_2O$ to 1000 ml, adjust pH to 7.2 | |
| Ammonium Persulfate (APS): | 1 g (10%) | APS |
| | Add $ddH_2O$ to 10 ml | |
| Laemmli Buffer (5x): | 15 g (40 mM) | Tris |
| | 216 g (0.95 M) | Glycine |
| | 15 g (0.5%) | SDS |
| | Add $ddH_2O$ to 3000 ml | |

Protein samples were separated by using a discontinuous gel system, which is composed of stacking and separating gel layers that differ in salt and acrylamide (AA) concentration.

**Table 4-17 - SDS-PAGE stock solutions**

| Stock solutions | Separating gel stock solution | Stacking gel stock solution |
|---|---|---|
| Final AA concentration | 12% | 5% |
| Stacking gel buffer | - | 25 ml |
| Separating gel buffer | 25 ml | - |
| SDS (10 %) | 1 ml | 1 ml |
| Rotiphorese Gel 30 (30%) | 40 ml | 16.65 ml |
| $H_2O$ | Adjust to 100 ml | |

**Table 4-18 - SDS-PAGE gel mixture**

| Stock solution | Separating gel | Stacking gel |
|---|---|---|
| Separating gel stock solution | 10 ml | - |
| Stacking gel stock solution | - | 5 ml |
| TEMED | 10 µl | 5 µl |
| APS (10%) | 100 µl | 50 µl |

The separating gel was prepared the day before electrophoresis and overlaid with water-saturated isobutanol until it was polymerized. Isobutanol was exchanged by separating gel buffer diluted 1:5 with water and the gel was stored overnight at 4°C. The following day, the stacking gel was poured on top of the separating gel, and a comb was inserted immediately. After polymerization, the gel was mounted in the electrophoresis tank, which was filled with 1x Laemmli buffer. Protein samples were loaded and the gel was run with 100 V until the sample buffer bands reached the surface of the stacking gel. Next, the voltage was increased to 120 V and the gel was run for 1-3 h. Proteins were then resolved through the separating gel according to their size.

### 4.3.4 Coomassie Staining of SDS-Gels

**Required solutions:**

| | |
|---|---|
| Bio-Safe Coomassie Stain: | BioRad, Munich, Germany |

SDS-gels were tossed in ddH$_2$O (three times, 5 min each) and subsequently incubated in the Coomassie solution for about 10 – 60 min. After washing overnight in ddH$_2$O, proteins appeared as blue bands on a transparent background.

### 4.3.5 Western Blot Analysis and Immunostaining (WB, Semi-dry technique)

**Required solutions and buffers:**

| | | |
|---|---|---|
| Anode Buffer A: | 36.3 g (0.3 M) | Tris |
| | 200 ml (20%) | Methanol |
| | Add ddH$_2$O to 1000 ml | |
| Anode Buffer B: | 3.03 g (25 mM) | Tris |
| | 200 ml (20%) | Methanol |
| | Add ddH$_2$O to 1000 ml | |
| Cathode Buffer C: | 5.20 g (4 mM) | ε-Amino-n-capron acid |
| | 200 ml (20%) | Methanol |
| | Add ddH$_2$O to 1000 ml | |
| TBS (10x): | 45.8 g (100 mM) | Tris/HCl, pH 8 |
| | 175.5 g (1.5 M) | NaCl |
| | Add ddH$_2$O to 2000 ml | |
| Washing Buffer (1x TBST): | 100 ml | TBS (10x) |
| | 1 ml (0.05%) | Tween-20 |
| | Add ddH$_2$O to 1000 ml | |
| Blocking Buffer: | 3.0 g (3%) | nonfat dried milk |
| | 100 ml | TBS |

After separation by SDS-PAGE (see 4.3.3), proteins were blotted electrophoretically onto a PVDF membrane (Immobilon-P, Millipore) using a three-buffer semi-dry system (Towbin *et. al.*, 1979) and visualized by immunostaining using specific antibodies and the ECL Prime Western Blotting System (Sigma-Aldrich, Taufkirchen, Germany). The membrane was cut to gel size, moistened first with isopropanol followed with buffer B and placed on top of three Whatman3MM filter paper soaked with buffer A (bottom, on the anode), followed by three Whatman3MM filter paper soaked with buffer B. The SDS-PAGE gel was then removed from the glass plates, immersed in buffer B and placed on top of the membrane. Another three Whatman 3MM filter papers soaked with buffer C were placed on top of the gel followed by the cathode. Air bubbles in-between the layers had to be avoided. Protein transfer was conducted for 30-45 min at 0.8 mA/cm$^2$ gel surface area. The typical composition of a semi-dry WB is shown in **Figure 4-1**.

| |
|---|
| **Cathode (-)** |
| 3 Whatman 3MM filter papers soaked with buffer C |
| SDS-PAGE gel |
| PVDF membrane |
| 3 Whatman 3MM filter papers soaked with buffer B |
| 3 Whatman 3MM filter papers soaked with buffer A |
| **Anode (+)** |

**Figure 4-1 - Typical WB composition**

Blotted membranes were washed once for 2-3 min with water, before they were blocked with 3% nonfat milk in TBS for 30 min at RT. Afterwards blots were washed once for 10 min with TBS before incubation with the primary antibody for 30 min at RT followed. After washing one time for 10 min with TBS, the membranes were incubated for 30 min at RT with a horseradish-peroxidase (HRP)-coupled secondary antibody, detecting the isotype of the first antibody. Eight washing steps with TBST in a total of 20 min preceded the visualization of bound antibody using the ECL Prime Western Blotting System (Sigma-Aldrich, Taufkirchen, Germany). Blots were developed using the Fusion Pulse imaging system from Vilber Lourmat (Eberhardzell, Germany) for 5 seconds to 30 min depending on the signal intensity. For re-blotting membranes were washed once with TBS before ReBlot Plus Mild Antibody Stripping Solution (Merck Millipore, Darmstadt, Germany) was added and incubated with the blots for 15 min at RT. Stripped blots were then washed once with TBS and blots were again blocked and stained as described above.

### 4.3.6 Microscale Thermophoresis (MST)

**Required solutions and buffers:**

| | | |
|---|---|---|
| Annealing buffer (10x): | 20 mM | Tris/HCl, pH 7.4 |
| | 2 mM | $MgCl_2$ |
| | 50 mM | NaCl |
| | ad 10 ml mit $ddH_2O$ | |
| MST buffer: | 20 mM | Tris/HCl, pH 7.6 |
| | 1.5 mM | $MgCl_2$ |
| | 0.5 mM | EGTA |
| | 300 mM | KCl |
| | 10% | glycerol |
| | 10 mM | DTT (1 M) |

Microscale thermophoresis is used to analyze interaction-dependent changes regarding the hydration shell of molecules and their thermophoretic mobility in solution. The difference between the diffusion of the substrate compared to the protein-bound substrate due to a thermophoretic gradient is detected via changes in the fluorescence of the molecules (Zillner et al. 2012). The sequence of the full-length human PU.1 protein was amplified by PCR from pORF9-hSPI1 (InvivoGen, San Diego, USA) and recombined into a modified pDM8 vector, encoding an N-terminal His-tag, using the Gateway technology (Thermo Fisher Scientific, Hudson, USA). The protein was expressed in Rosetta2(DE)pLysS (Merck Millipore, Darmstadt, Germany) and purified by Nickel affinity chromatography (Qiagen, Hilden, Germany). Double-stranded DNA molecules were annealed from single-stranded, HPLC-purified oligonucleotides (Sigma-Aldrich, Taufkirchen, Germany). The annealing reaction (10 µl) was performed in 1x annealing buffer and comprised 20 µM of the Cy3-labeled oligonucleotide (upper strand) and 20.8 µM of the unlabeled oligonucleotide (lower strand). The annealing reaction was incubated for 15 min at 95°C in a thermoblock (peQLab, Erlangen, Germany) and afterwards allowed to slowly cool down to room temperature o/n. The annealing reaction was checked on an 8% native polyacrylamide gel which was analyzed on a fluorescence imager (Fujifilm, Tokyo, Japan). The binding assay was carried out using the Nanotemper Monolith NT.115 (initial settings: LED power: 90%, IR-laser power: 80%, 25°C). For each motif affinity measurement 16 reactions were prepared on ice in MST buffer. The Cy3-labeled dsDNA oligo was always kept at a constant concentration of 50 nM. The unlabeled protein was titrated in a 1:1-dilution series with a starting concentration of 23 µM. Every binding assay comprised one control reaction without any protein. After loading the binding reactions into standard capillaries (NT.115), the mixture was incubated for 15 min at 25 °C in the Nanotemper device before starting the measurement. The data was analyzed using the NT-analysis acquisition software (1.2.229), which plots a binding curve using the normalized fluorescence of the labeled dsDNA at different concentrations of the unlabeled protein. Each binding assay was performed twice and the mean value was calculated. For every single binding assay a maximum of three outlier values were eliminated.

### 4.3.7 Proximity-dependent Biotin Identification (BioID)

**Required solutions and buffers:**

| | | |
|---|---|---|
| Cell buffer Mix (CBM): | 0.5 ml (10 mM) | HEPES/KOH (1 M), pH 7.9 |
| | 4.25ml (85 mM) | KCl (1 M) |
| | 0.1 ml (1 mM) | EDTA (500 mM), pH 8.0 |
| | Add ddH$_2$O to 50 ml | |
| | Add just prior to use to | 1 ml of CBM: |
| | 10 µl (1 mM) | PMSF (100 mM) |
| | 5 µl (1 mM) | Sodium-o-vanadate (200 mM) |
| | 20 µl (1x) | 50x Proteaseinhibitor cocktail |
| Lysis buffer 1A (L1A): | 0.9 ml | CBM (including inhibitors) |
| | 0.1 ml (0.4%) | ddH$_2$O |
| Lysis buffer 1B (L1B): | 0.9 ml | CBM (including inhibitors) |
| | 0.1 ml (1%) | NP-40 (10%) |
| Lysis buffer 2 (L2): | 2.5 ml (50 mM) | Tris/HCl, pH 7.4 (1 M) |
| | 2 ml (0.4%) | SDS (10%) |
| | 0.5 ml (5 mM) | EDTA pH 8.0 (0.5 M) |
| | 5 ml (500 mM) | NaCl (5M) |
| | Add ddH$_2$O to 50 ml | |
| | Add just prior to use to | 1 ml of L2: |
| | 10 µl (1 mM) | PMSF (100 mM) |
| | 5 µl (1 mM) | Sodium-o-vanadate (200 mM) |
| | 20 µl (1x) | 50x Proteaseinhibitor cocktail |
| Dilution buffer (DB): | 2.5 ml (50 mM) | Tris/HCl, pH 7.4 (1 M) |
| | Add ddH$_2$O to 50 ml | |
| | Add just prior to use to | 1 ml of DB: |
| | 10 µl (1 mM) | PMSF (100 mM) |
| | 5 µl (1 mM) | Sodium-o-vanadate (200 mM) |
| | 20 µl (1x) | 50x Proteaseinhibitor cocktail |
| Triton X-100 (10%): | 1 ml (10%) | Triton X-100 |
| | Add ddH$_2$O to 10 ml | |

| | | |
|---|---|---|
| Wash buffer 1 (WB1): | 5 ml (100 mM) | Tris/HCl, pH 8.0 (1 M) |
| | 12.01 g (4 M) | Urea (60.06 g/mol) |
| | Add ddH$_2$O to 50 ml | |
| Wash buffer 2 (WB2): | 0.395 g (100 mM) | NH$_4$HCO$_3$ (79.06 g/mol) |
| | Add ddH$_2$O to 50 ml | |

BioID was established by Roux et al. in 2013 and is used to analyze proximity-dependent interactions of proteins in living cells or tissues. The method is based on the use of a promiscuous prokaryotic biotin protein ligase. The biotin ligase is fused to a protein of interest and then introduced into mammalian (or other) cells where it will biotinylate vicinal proteins upon supplementation of the culture medium with biotin. Biotinylated proteins can then be selectively isolated and identified by mass spectrometry. For this approach fusion proteins, containing the protein of interest and a flexible linker region fused to *E.coli* Biotin Ligase (BirA*, harboring a R118G point mutation), were first cloned using Gibson Assembly (see section 4.2.2.4), before linearized DNA templates were used for *in vitro* transcription (see section 4.2.5.2). Transcripts (50 µg of PU.1-BirA*-mRNA, others according to their size) were then introduced into the desired cells (25x10$^6$ per construct) using electroporation (see 4.1.3.2). 3 h after transfection the cell culture medium was supplemented with 50 µM biotin to achieve biotinylation of vicinal proteins. 5 h after biotin supplementation cells were harvested for subsequent lysis and purification of biotinylated proteins. The optimal time points for addition of biotin and harvesting the cells were determined by WB (see section 4.3.5 and **Figure 5-30**). Cell pellets were washed twice with ice-cold PBS to eliminate residual biotin. For lysis cell pellets were first swelled in 625 µl of L1A, before an equal amount of L1B was added. Suspensions were incubated on ice for 10 min before nuclei were spun down at 700×g for 5 min at 4°C. In a next step nuclei were resuspended in 500 µl L2 buffer and sonicated to fragment the gDNA of the cells. All sonication steps were carried out with a constant duty cycle, output control 2 for 10 s using a Branson Sonifier 250 (Danbury, USA). After sonication in L2, Triton X-100 (10%) was added to a final concentration of 2%, before a second sonication step took place. Then an equal amount of DB was added and suspensions were sonified one last time. Lysates were centrifuged at 11.000×g for 15 min at 4°C and supernatants were transferred into a new tube. Pierce™ Streptavidin Magnetic Beads (Thermo Fisher Scientific, Waltham, USA) were prewashed according to manufacturer’s instructions and 75 µl prewashed beads were added to each lysate, therefore recovering biotinylated proteins on the beads via the biotin-streptavidin bond. Lysates were incubated on a rotating wheel at 4°C o/n. The next day beads were washed on a magnet for three times with WB1 and WB2 (400 µl each) at RT and transferred to a new tube with the last washing step. Protein-conjugated beads were finally resuspended in 25 µl of WB2 and sent for subsequent on-bead digestion and mass spectrometry analysis to the Zentrallabor für Proteinanalytik (ZfP) of the LMU (Prof. Dr. Axel Imhof). However, the protocol was slightly modified when using CTV-1 cells. In brief, the amount of used cells, mRNA and all buffers accordingly was doubled. Furthermore a dialysis step using Slide-A-Lyzer MINI Dialysis Devices (Thermo Fisher Scientific, Waltham, USA) according to manufacturer's instructions was included after lysis of the cells. Dialysis was performed o/n to completely remove residual, unbound biotin, before lysates were incubated with the Pierce™ Streptavidin Magnetic Beads (Thermo Fisher Scientific, Waltham, USA) for 1-2 h. Washing of the beads and all following steps were essentially carried out as already described above.

# 5 RESULTS

In the human system, each cell comprises 2 m of DNA, which is ordered into 23 chromosomes that need to be tightly compacted into the nucleus. As already described above, most of the cells of such a multicellular organisms comprise the same genetic information, however, their gene expression profiles differ from cell to cell, which is mainly due to the activity of sequence-specific TFs. Most of those TFs recognize specific DNA motifs that frequently range from 6-12 bp in length, which indicates that each TF harbors 1-2 million potential binding sites throughout the genome. However, the access to their potential binding sites is highly restricted and only a relatively small proportion of binding sites is effectively bound. In primary human MO and MAC for instance, the cell type-specific TF PU.1 shows a significant proportion of cell stage-specific binding events despite comparable PU.1 expression levels in both cell types (Pham et al. 2012). This raises the general question, which parameters distinguish bound vs. non-bound sites. To analyze this in detail, the binding properties of the pioneer TF PU.1 were studied with regard to the higher-order chromatin structure, its sequence affinity and motif cooperativeness with co-associated factors.

## 5.1 Cell Type-Specific PU.1 Binding Site Selection

### 5.1.1 Cooperativeness between Sequence-Specific Transcription Factors

Recent studies have reported the cell type-specific distribution of PU.1- binding sites in mouse lymphoid and myeloid progenitor cells, mouse MAC or B cells, as well as in human MO and MAC. In MO, MAC and B cells, cell type-specific PU.1 binding is often characterized by the co-occurrence of sequence motifs for additional cell type-specific TFs (Pham et al. 2012). To further analyze the impact of motif cooperativeness on PU.1 binding site selection, a comprehensive analysis of PU.1 binding across 23 different blood cell types and cell lines was carried out using ChIPseq (see sections 4.2.6 & 4.2.7). For this purpose own, as well as public available PU.1 ChIPseq data sets were analyzed. NCBI GEO accession numbers of public available data sets are listed in supplementary **Table 10-1**. Obtained raw read data was mapped to the human reference genome version 19 (hg19; see section 10.1.1.1). Computational analyses of the corresponding ChIPseq data is described in more detail in section 10.1.2. In brief, to include copy-number variation (CNV) calculations, obtained raw read data from cell lines was CNV-normalized. Moreover, scaffolds were removed to reduce the analysis to the reference chromosomes only. Data of corresponding replicates was merged, resulting in 23 distinct data sets, which were used for further analysis.

To visualize the ChIPseq signals, bigWigs were generated and the read counts were averaged across replicates. The distribution of those PU.1 reads in six representative cell types is shown in **Figure 5-1**.

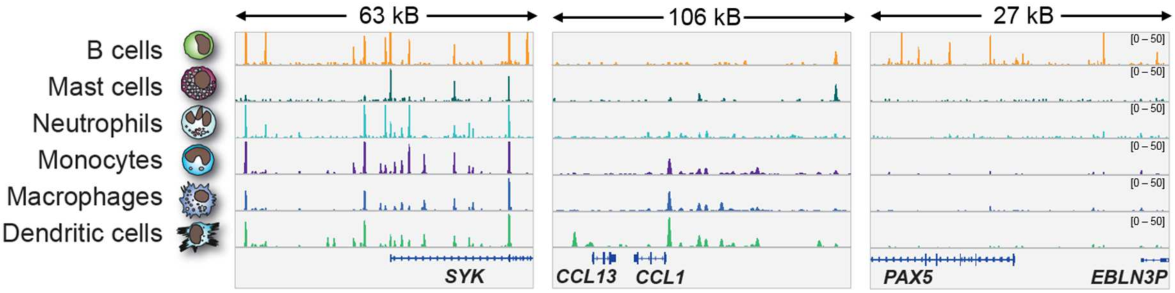


**Figure 5-1 - Selected IGV genome browser tracks of PU.1 reads across three representative loci**

ChIPseq PU.1 reads of B cells, mast cells, neutrophils, MO, MAC and DC across three different genomic loci are shown. All reads are CNV-normalized and averaged across replicates. The exact localization of the PU.1 binding sites is depicted for each locus (SYK, chromosome 13; CCL13/CCL1, chromosome 17; PAX5/EBLN3P, chromosome 9).

As seen in the above IGV genome browser section, the SYK-locus contains PU.1 binding sites, which are bound by PU.1 in all different cell types, whereas binding sites in the CCL1-region are only bound by PU.1 in the myeloid lineage (MO, MAC and DC), and binding sites in the PAX5-region are only bound by B cells respectively. This indicates that there exist common, as well as cell type-specific PU.1 binding sites in the different PU.1-expressing cell types. To further analyze the cell type-specific binding preferences of PU.1, a merged peak set containing all PU.1 binding sites across the different cell types was generated. For this purpose, an individual peak was called, when at least 15 PU.1 reads with a false discovery rate (FDR) below 0.05 were present. The resulting peak set was further filtered to include the mapability across the reference genome. For normalization, read counts of the peaks where regularized log-transformed (rlog) and no further background subtraction was applied prior to annotation. This method is useful when working with data sets varying strongly in size, because it minimizes differences between the samples by normalization to the corresponding library size. This analysis resulted in roughly 155.000 annotated PU.1 peaks. The cell type-specific distribution of those peaks is depicted in **Figure 5-2** with the t-Distributed Stochastic Neighbor Embedding (t-SNE) technique, which allows for the visualization of high-dimensional data sets by reducing the dimensions to a two-dimensional map.

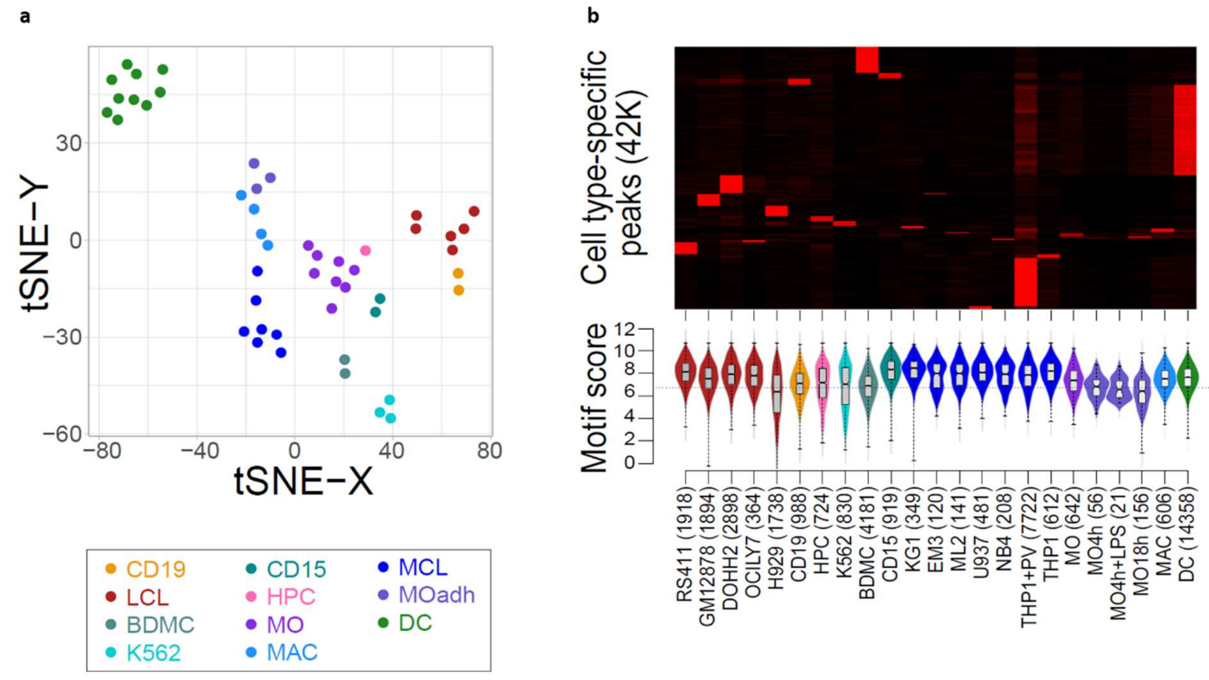


**Figure 5-2 - Distribution of PU.1 peaks across different cell types**

**a |** 2-dimensional visualization of the distribution of annotated PU.1 peaks across analyzed cell types using t-SNE embedding. Correlated cell types are depicted with identical colors (DC, green; stimulated MO (MOadh), light purple; MAC, light blue; MO, purple; HPC, pink; myeloid cell lines (MCL), dark blue; BDMC, dark gray; CD15, dark cyan; K562, dark turquoise; lymphoid cell lines (LCL), red; CD19, orange). **b |** Cell type-specific peaks and the frequency of their occurrence in the depicted cell type are shown in the heat map. Red coloring correlates with a high enrichment of those PU.1 peaks in a certain cell type. The quantity of peaks in each cell type is given for each sample. Below the heat map, the motif score distribution of the consensus MAC PU.1 motif, generated earlier, is shown for each cell type. The median of the specific distribution across all clusters is depicted inside the bean plot with a conventional boxplot.

As seen in **Figure 5-2a** public available and own ChIPseq replicates cluster together. Myeloid cell lines (MCL, dark blue) for instance show a good correlation with each other, whereas lymphoid cell lines (LCL, red) cluster together in close proximity to CD19-positive B cells (orange). This indicates that each cell type comprises individual PU.1 peaks, which are either specific or shared between related cell types. To visualize the cell type-specific distribution of PU.1 binding sites and their correlated motif affinity, the annotated rlog-transformed peak file was joined with the peak calling info and sorted according to the individual cell types resulting in roughly 42.000 cell type-specific PU.1 peaks. As seen in **Figure 5-2b**, many of those peaks are highly abundant in a certain cell type, while they are not necessarily bound by PU.1 in any of the other cell types. Moreover, the quantity of PU.1 peaks highly differs across the analyzed cell types. THP-1 cells differentiated with PMA and VD3 for example, comprise 7722 specific PU.1 peaks, whereas ML2 cells, also belonging to the myeloid lineage, roughly contain 141 specific PU.1 peaks. As the affinity between a TF and its recognition sequence is also a proximate determinant of DNA binding, we further studied the relationship of motif scores and PU.1 binding, which is shown below the heat map in **Figure 5-2b**. The motif used represents a consensus sequence derived from human MAC and covers up to 80% of all PU.1 ChIPseq peaks identified in HPC, MO or MAC (Pham et al. 2012). As illustrated, the motif score distribution also varies between the cell types as CD15-positive neutrophils for example are enriched for PU.1 binding sites with high motif scores, whereas CD19-positive B cells comprise lower motif scores within their PU.1 peaks. Since the cell type-specific binding site-selection of PU.1 depends on cooperativeness with additional sequence-specific TFs, we further analyzed the co-enrichment of additional TF motifs across our data set. Therefore, *de novo* motifs in each of the cell type-specific peak sets were determined using the HOMER suite (Hypergeometric Optimization of motif enrichment; Heinz et al. 2010). HOMER is a *de novo* motif-finding algorithm based on a Fisher exact or so-called hypergeometric test. Enriched motifs are obtained by screening all found oligomer sequences of the target set in comparison to a background set, which is based on a cumulative, hypergeometric distribution resulting in probability matrices for the motifs with the lowest p-value. To analyze PU.1's cell type-specific dependency on cooperativeness a merged file containing all motif enrichments across the different cell types was generated and further filtered to obtain only the top motif hits with a p-value of at least 1e$^{-12}$ and a motif correlation above 0.85. Solely those motifs were conducted to the downstream analyses. To illustrate the motif enrichment across the different cell types a balloon plot was generated, where the correlation between the cell type-specific peaks with their distinct motif signatures is exemplified **(Figure 5-3a)**.

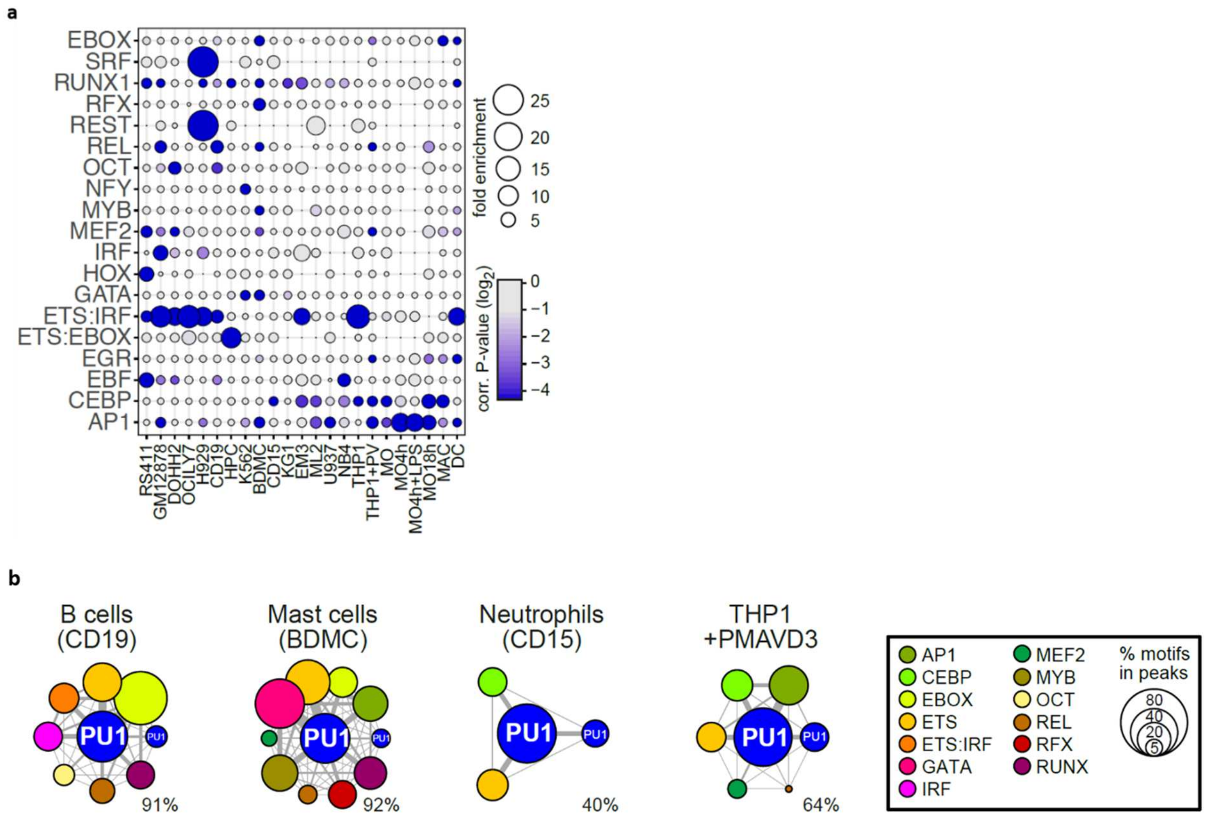


**Figure 5-3 - Motif enrichment across analyzed cell types**

**a |** The balloon plot depicts the motif enrichment of the selected motifs (y-axis) in each cell type (x-axis). The node thickness represents the enrichment and the coloring correlates with the p-value of the motif co-occurrence in PU.1 peaks of each cell type. **b |** Motif co-enrichment networks for four selected cell types are displayed. The fraction of PU.1 peaks overlapping with co-associated TF motifs is shown below each network. The size of each node represents the according motif enrichment in percent and the coloring indicates the co-associated TF motif. PU.1-PU.1-associations are drawn in blue.

The composite binding site of ETS and IRF for instance (ETS:IRF) is highly enriched in PU.1 peaks of lymphoid cells (RS411, GM12878, DOHH2, OCILY7, H929 & CD19) and myeloid cells like EM3 and THP-1, but is not significantly enriched in erythrocyte cells like K-562 or neutrophils (CD15-postive cells) respectively. The motif for the cell type-specific TF AP-1 on the other hand, is mainly found in MO, but is not enriched in most lymphoid cells, whereas C/EPB TF motifs mainly co-occur within PU.1 peaks of the myeloid lineage (EM3, ML2, THP-1, MO & MAC). This indicates that cell type-specific PU.1 peaks harbor distinct co-associated motif signatures, which might be involved in PU.1 binding site selection. To analyze this in more detail, we also generated networks of motif co-associations for four selected samples to investigate the fraction of PU.1 peaks overlapping with cooperative TF motifs. As seen in **Figure 5-3b,** B cells and mast cells (BDMC) comprise around 91% or 92% of PU.1 peaks, respectively, containing motifs for additional co-associated TFs and homotypic PU.1 motifs. In terms of B cells, those motifs are mainly consensus sites for EBOX and ETS TFs, whereas PU.1 peaks of mast cells are mainly co-enriched for GATA TF motifs. These signatures correlate well with TFs known to be expressed in those cell types. Neutrophils and differentiated THP-1 cells on the other hand, have a lower fraction of PU.1 peaks co-associated with additional cooperative motifs (40% and 64% respectively). The main motifs enriched in the PU.1 peaks in those sets are mostly PU.1 binding motifs itself. This implies that motif cooperativeness of PU.1 binding sites is highly cell type-dependent, since the analyzed cell types show a diverse dependency on co-association with additional TFs. Since the static analysis of PU.1 binding sites using ChIPseq does not take into account the existing chromatin structure or epigenetic features, the next chapter will focus on DNA methylation and its impact on PU.1 binding site selection.

## 5.2 Epigenetic Determinants of PU.1 Binding Site Selection

### 5.2.1 *In vitro* Binding Affinity to Methylated DNA

Since cooperativeness alone is not sufficient to explain all differences in PU.1 binding site selection, we also analyzed the impact of pre-existing epigenetic modifications on PU.1 binding preferences. To gain insights into the general question, if DNA methylation inhibits PU.1 binding or if PU.1 binding leads to the subsequent demethylation of the corresponding binding site, we first measured the TF's ability to bind to methylated or hemi-methylated DNA oligomers. Therefore, we assayed the relationship between motif log-odds scores of selected 12mers out of the human MAC-derived position weight matrix (PWM) and their affinity to full-length PU.1 in solution. The corresponding PWM was already analyzed by our group in 2013 (Pham et al.) and we further selected certain oligomers, which harbor a CpG-motif either next to the PU.1 GGAA-core binding sequence or distal from the core sequence with high motif log-odds score between 8 and 9.3. The analyzed oligomers were synthesized Cy-3-labled, with either a methylated or hydroxyl-methylated CpG-site (5mC/5hmC), or as unmodified DNA oligomers (section 3.7.5; Sigma-Aldrich, Taufkirchen, Germany). We first determined dissociation constants ($K_D$ values) for two individual sequence motifs and bacterially expressed full-length PU.1 in solution using microscale thermophoresis (MST; see section 4.3.6), which detects interaction-dependent changes in the hydration shell of molecules, as well as their changes in thermophoretic mobility in solution. The modifications were established on the sense and antisense strand respectively. Each of the 12mer was measured in triplicates and the resulting mean was calculated to analyze the binding affinity of PU.1 to the methylated DNA. **Table 5-1** shows the result for one oligomer, which harbors the modification distal from the PU.1 core binding sequence, and **Table 5-2** shows the corresponding results for one oligomer, which was modified within the core recognition sequence.

**Table 5-1 - $K_D$ values analyzed by MST for the oligomer modified distal from the core sequence**

| Cy-3 labeled sense strand | complementary antisense strand | $K_D$ exp. 1 | $K_D$ exp. 2 | $K_D$ exp. 3 | $K_D$ mean [nM] |
|---|---|---|---|---|---|
| acgtCGCTTCCTCTTTacgt | acgtAAAGAGGAAGCGacgt | 541 | 547 | 575 | 554 |
| | acgtAAAGAGGAAG**(5mC)**Gacgt | 628 | 617 | 608 | 618 |
| | acgtAAAGAGGAAG**(5hmC)**Gacgt | 516 | 462 | 543 | 507 |
| | | | | | |
| acgt**(5mC)**CGCTTCCTCTTTacgt | acgtAAAGAGGAAGCGacgt | 705 | 659 | 720 | 695 |
| | acgtAAAGAGGAAG**(5mC)**Gacgt | 651 | 649 | 582 | 627 |
| | acgtAAAGAGGAAG**(5hmC)**Gacgt | 886 | 1000 | 870 | 919 |
| | | | | | |
| acgt**(5hmC)**CGCTTCCTCTTTacgt | acgtAAAGAGGAAGCGacgt | 542 | 507 | 482 | 510 |
| | acgtAAAGAGGAAG**(5mC)**Gacgt | 677 | 718 | 662 | 686 |
| | acgtAAAGAGGAAG**(5hmC)**Gacgt | 665 | 645 | 738 | 683 |

**Table 5-2 - $K_D$ values analyzed by MST for an individual oligomer modified within the core sequence**

| Cy-3 labeled sense strand | complementary antisense strand | $K_D$ exp. 1 | $K_D$ exp. 2 | $K_D$ exp. 3 | $K_D$ mean [nM] |
|---|---|---|---|---|---|
| acgtCACTTCCGCTTCacgt | acgtGAAGCGGAAGTGacgt | 458 | 436 | 391 | 428 |
| | acgtGAAG**(5mC)**GGAAGTGacgt | 1990 | 2020 | 1860 | 1957 |
| | acgtGAAG**(5hmC)**GGAAGTGacgt | 2210 | 2150 | 2210 | 2190 |
| | | | | | |
| acgtCACTTC**(5mC)**GCTTCacgt | acgtGAAGCGGAAGTGacgt | 464 | 513 | 532 | 503 |
| | acgtGAAG**(5mC)**GGAAGTGacgt | 1730 | 1720 | 1970 | 1807 |
| | acgtGAAG**(5hmC)**GGAAGTGacgt | 1590 | 1590 | 1750 | 1643 |
| | | | | | |
| acgtCACTTC**(5hmC)**GCTTCacgt | acgtGAAGCGGAAGTGacgt | 271 | 215 | 260 | 249 |
| | acgtGAAG**(5mC)**GGAAGTGacgt | 1550 | 1500 | 1290 | 1447 |
| | acgtGAAG**(5hmC)**GGAAGTGacgt | 1420 | 1320 | 1260 | 1333 |

Considering the oligomer modified distal from the PU.1 core binding sequence **(Table 5-1)**, no significant difference between the individual motif affinities is detected. The corresponding $K_D$ values show only slight variations, which are independent of the methylation status of the DNA oligomer. Regarding the oligomer modified within the PU.1 core motif however **(Table 5-2)**, the methylation of the antisense, but not the sense strand, leads to an increase of the corresponding $K_D$ value. This effect is also detectable for the hydroxyl-methylated oligomer and the findings are summarized in the following figure **(Figure 5-4)**.

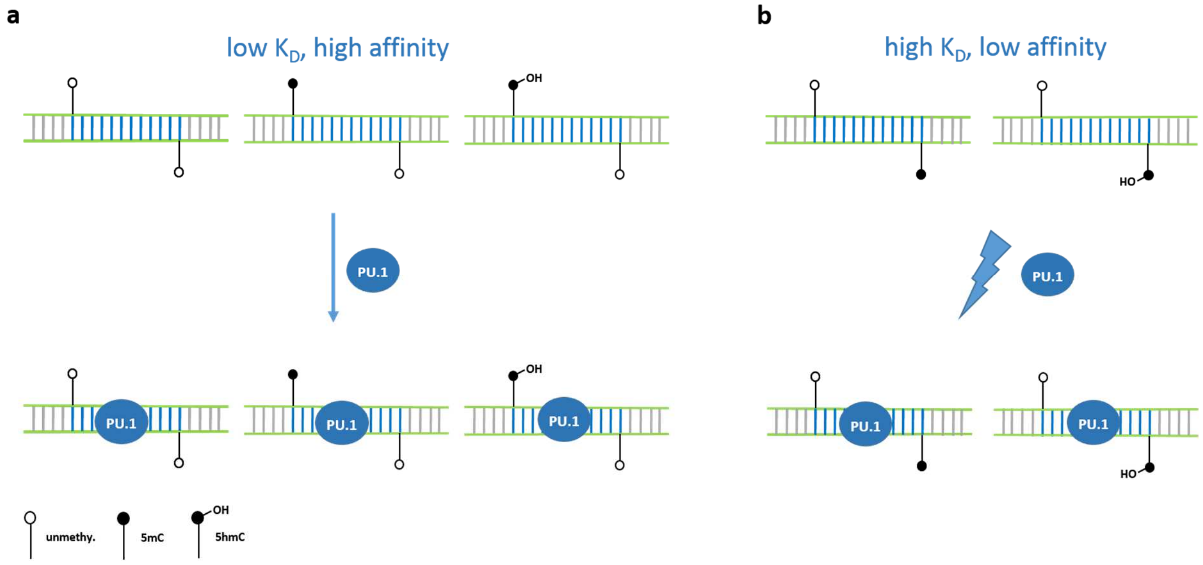


**Figure 5-4 - PU.1 binding affinity to methylated DNA oligomers**

**a |** PU.1’s binding affinity to unmodified or hemi-methylated DNA (5mC/5hmC on the **sense strand**) is illustrated. Lollipops indicate the methylation status of the CG-site located within the 12mer. **b |** PU.1’s binding affinity to hemi-methylated DNA (5mC/5hmC on the **antisense strand**) is illustrated. Lollipops indicate the methylation status of the CG-site located within the 12mer.

To strengthen these results, two additional oligomers, modified within the core recognition sequence, were measured. Here, only the 5mC-modification was introduced, since no significant differences in PU.1’s binding affinity to 5mC or 5hmC binding sites were detected before (see **Table 5-2**). The results of this analysis are shown in **Table 5-3** and further indicate that PU.1’s motif affinity is decreased in hemi-methylated DNA, with the modification on the antisense strand resulting in an asymmetric inhibition of PU.1 binding *in vitro*.

**Table 5-3 - $K_D$ values analyzed by MST for two additional oligomers modified within the core sequence**

| Cy-3 labeled sense strand | complementary antisense strand | $K_D$ exp. 1 | $K_D$ exp. 2 | $K_D$ exp. 3 | $K_D$ mean [nM] |
|---|---|---|---|---|---|
| acgtCACTTCCGCCTCacgt | acgtGAGGCGGAAGTGacgt | 608 | 693 | 765 | 689 |
| | acgtGAGG**(5mC)**GGAAGTGacgt | 1490 | 3620 | 2990 | 2700 |
| | | | | | |
| acgtCACTTC**(5mC)**GCCTCacgt | acgtGAGGCGGAAGTGacgt | 659 | 684 | 673 | 672 |
| | acgtGAGG**(5mC)**GGAAGTGacgt | 1040 | 2930 | 1890 | 1953 |
| | | | | | |
| **Cy-3 labeled sense strand** | **complementary antisense strand** | **$K_D$ exp. 1** | **$K_D$ exp. 2** | **$K_D$ exp. 3** | **$K_D$ mean [nM]** |
| acgtCACTTCCGTTTTacgt | acgtAAAACGGAAGTGacgt | 543 | 679 | 649 | 624 |
| | acgtAAAA**(5mC)**GGAAGTGacgt | 2450 | 3170 | 3320 | 2980 |
| | | | | | |
| acgtCACTTC**(5mC)**GTTTTacgt | acgtAAAACGGAAGTGacgt | 590 | 705 | 744 | 680 |
| | acgtAAAA**(5mC)**GGAAGTGacgt | 2050 | 2880 | 2850 | 2593 |

## 5.2.2 Comparison of Methylated vs. Unmethylated Binding Sites *in vivo*

To further study the impact of DNA methylation on PU.1 binding site selection *in vivo*, we analyzed the expression profile of PU.1 and its binding occupancy on methylated vs. partial unmethylated gDNA. As a model system, we chose the human lymphoid leukemia cell line CTV-1, which lacks endogenous PU.1 expression and which will be further elucidated in the following chapter (section 5.3). To express PU.1 in this cell line, we made use of *in vitro* transcribed mRNA, which was transfected into the cells using electroporation (see sections 4.2.5.2 and 4.1.3.1). To determine the transfection efficiency, *in vitro* transcribed GFP mRNA was used and the cells were analyzed using the FACS technique (in cooperation with the working group of PD. Petra Hoffmann; see **Figure 5-5**). Cells were either mock-transfected, or a mutated PU.1 transcript (hereafter PU.1mut), where all transcription start sites (ATG) were mutated, was used as control. This transcript served as the ideal control, having the same size then the wildtype (WT) PU.1 transcript, but lacking the ability to be translated into protein. The human PU.1-ORF (open reading frame) and its corresponding mutated version were synthesized as gBlock® gene fragments (IDT, Coralville, USA; see section 3.8), also harboring a N-terminal triple FLAG-tag (3xFLAG) for detection. Those fragments were directly used as templates for *in vitro* transcription. Successful protein expression and its nuclear localization was assessed first using WB analysis as depicted in **Figure 5-5**. For this purpose, CTV-1 cells were transfected with varying amounts of 5’-capped, poly-adenylated PU.1 mRNA, with 100% referring to 1 µg per 1x10$^6$ cells. Nuclear extracts, as well as cytosolic fractions were harvested 8 h after transfection (see 4.3.2) and further analyzed using WB analysis (see 4.3.5). For protein expression analysis an anti-PU.1 antibody was used, whereas the nuclear localization of the protein was verified using an anti-histone H3 antibody. The anti-actin antibody used in addition served as a general loading control (for a list of used antibodies see section 3.5).

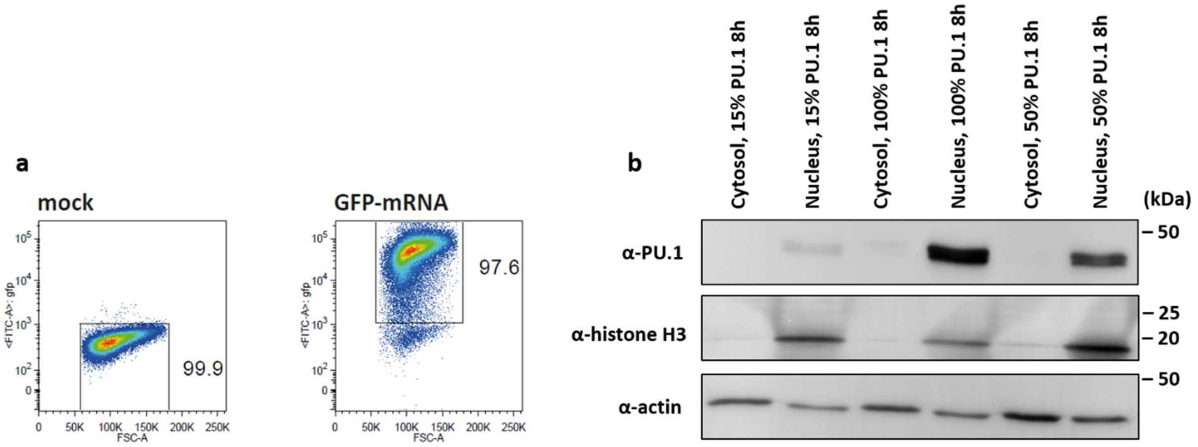


**Figure 5-5 - FACS and WB analysis of transfected CTV-1 cells**

**a |** Representative FACS analysis of mock-transfected and GFP-transfected CTV-1 cells. 1 µg GFP mRNA per 1x10$^6$ cells was used. The percentage of GFP-positive cells is indicated. Cells were stained with DAPI prior to the analysis, to obtain the fraction of viable cells. **b |** CTV-1 cells were transfected with varying amounts of PU.1 mRNA (15%, 50% and 100%). Nuclear extracts were prepared 8 h after transfection and analyzed using WB. Proteins were detected with an anti-PU.1, anti-histone H3 and an anti-actin antibody. 15 µg of the corresponding cytosolic and nuclear fractions were loaded per lane.

As illustrated in the above FACS plots, electroporation of *in vitro* transcribed mRNA into the CTV-1 cells led to a transfection efficiency of nearly 100% with a good survival of the cells. Moreover, as seen in the WB image, *in vitro* transcribed PU.1 mRNA is translated into the corresponding full-length PU.1 protein in the lymphoid CTV-1 cell line (approx. 45 kDa) and the protein is expressed in a concentration-dependent manner, since lower amounts of PU.1 mRNA (15% and 50% respectively) led to decreased PU.1 expression (upper blot). Moreover, the TF is successfully located to the nucleus, since its expression correlates with the nuclear expression of histone H3 (approx. 18 kDa; middle blot). As the ratio of 1 µg mRNA per 1x10$^6$ cells led to a prominent PU.1 expression in the non-PU.1-expressing CTV-1 cells, this ratio was used for all further experiments unless otherwise noted.

As we aimed to analyze the effect of DNA methylation on PU.1 binding *in vivo*, CTV-1 cells were either left untreated, or treated with DAC. This epigenetic modifier inhibits DNA methyltransferase activity resulting in global DNA demethylation (Mund et al. 2011). To obtain the ideal working concentration for our model system, we first treated the cells for eight consecutive days with varying amounts of the drug. Afterwards the gDNA was isolated and used as a template for TBSAseq (see sections 4.2.3.1 & 4.2.10). The analyzed amplicon was selected from preliminary data obtained in our working group and is illustrated in the following IGV genome browser track **(Figure 5-6)**. The amplicon is located within a region of the CD14 gene, which is methylated in the lymphoid cell line (CTV-1), as measured by MCIp-seq (methyl-CpG immunoprecipitation coupled to sequencing), whereas it remains unmethylated in myeloid cells like MO (lower MCIp-track). To verify the demethylation of the 15 CpG-sites within the amplicon and, as consequence the global demethylation of all methylated CpG-sites in the CTV-1 cell line, generated TBSAseq raw read data was further processed with the public available CRISPResso software (Pinello et al. 2016). The detailed computational analysis is highlighted in chapter 10.1.3 and the corresponding heat map is shown in **Figure 5-6** below the IGV genome browser track.

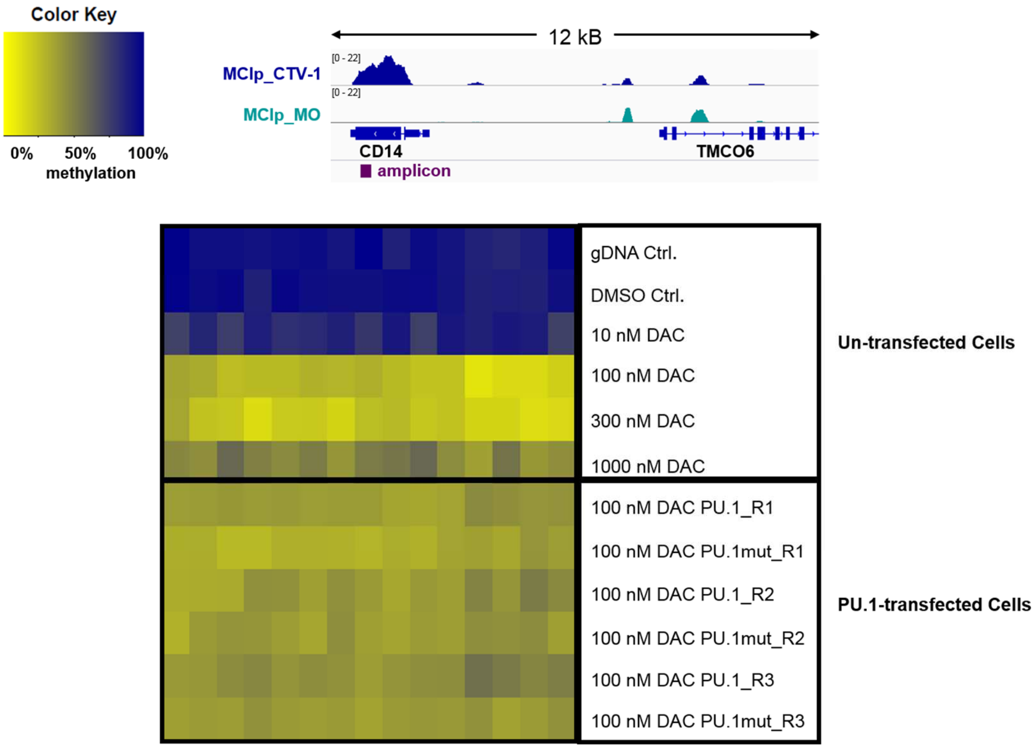


**Figure 5-6 - Methylation (MCIp-seq) & Targeted Bisulfite Amplicon Sequencing data (TBSAseq) of untreated and treated CTV-1 cells**

The IGV genome browser section shows the MCIp-seq signal of CTV-1 cells and MO across the CD14 locus (chr5:140,011,247-140,023,629). The exact localization of the analyzed amplicon within the CD14 region is depicted in purple. The heat map below shows the methylation degree of CTV-1 gDNA under the corresponding conditions. gDNA of untreated cells, as well as gDNA of cells treated with DAC and with DMSO used as a control is depicted. The upper black box depicts the results of DAC-treated, un-transfected CTV-1 cells, while the lower box corresponds to DAC-treated CTV-1 cells transfected with PU.1wt and PU.1mut mRNA respectively (each in triplicates). Each vertical line depicts a single CpG site of the analyzed CD14-amplicon (15 CpGs in total). The methylation degree is indicated with coloring (yellow, low methylation; blue, highly methylated).

As the heat map illustrates, gDNA of CTV-1 cells in the selected amplicon is efficiently demethylated using 100 and 300 nM of the drug. In terms of lower concentrations, the demethylation is not complete (10 nM), whereas with higher concentration (1000 nM) the survival of the cells, as mirrored by cell number and viability assessing (see section 4.1.1.2) is very low (data not shown). Since 100 nM DAC was sufficient to demethylate the gDNA within the selected region with a good survival of the cells, this concentration was used for all further experiments. Therefore, CTV-1 cells were again treated for eight consecutive days with 100 nM DAC, before the PU.1 mRNA (and PU.1mut mRNA respectively) was transfected into the cells. gDNA of those cells was also isolated after 8 h and the demethylation was verified by TBSAseq as also depicted in the above heat map (lower black box; obtained raw data for each individual CpG is listed in **supplementary Table 10-2**).

To analyze the gene expression profile of PU.1-transfected DAC-treated vs. untreated cells RNAseq (see chapter 4.2.9) was utilized. Moreover, ChIPseq and ATACseq were used to mirror PU.1 binding sites and chromatin accessibility with regard to the DNA methylation status of the cells (see sections 4.2.6-4.2.8). Total RNA of transfected CTV-1 cells was isolated after 24 h (see sections 4.2.5), whereas chromatin and nuclei for ATACseq were already harvested 8 h after transfection. To study changes in the PU.1-induced transcriptome of treated vs. untreated transfected CTV-1 cells a combined read count table of all replicates was generated and the corresponding expression profile was analyzed (detailed computational analysis is provided in section 10.1.4, 'Basic analysis of RNAseq data'). To visualize the level of similarity of the individual conditions a multidimensional scaling (MDS) plot was generated at first **(Figure 5-7)** out of the batch-corrected data.

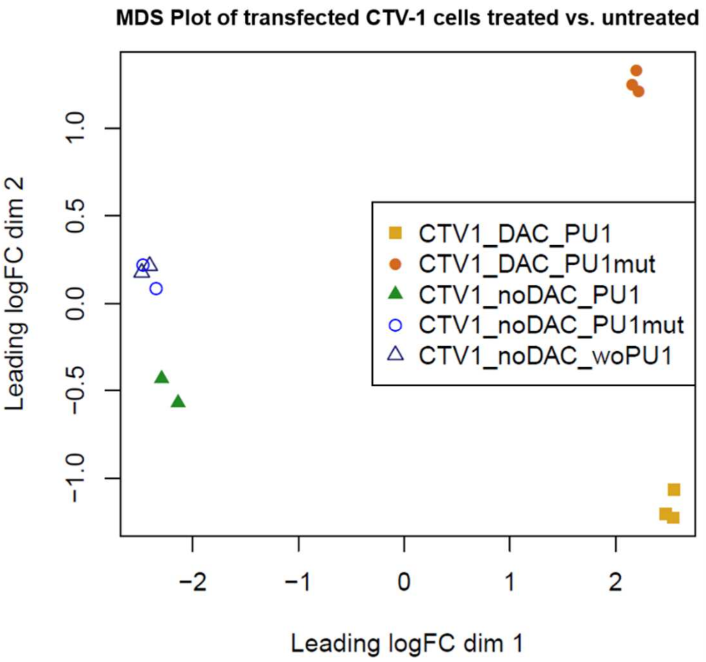


**Figure 5-7 - MDS Plot of transfected CTV-1 cells treated with DAC vs. untreated**

MDS plot showing the correlation of the log fold-change (logFC) of transfected CTV-1 cells, which were treated with DAC (DAC) or left untreated (noDAC). Cells were either mock-transfected (wo_PU1), transfected with mutated PU.1 mRNA (PU1mut) or with wildtype PU.1 mRNA (PU1). Experiments were done in duplicates for untreated cells and performed in triplicates for the DAC-treatment. Replicates share the same color code (noDAC_woPU1, dark blue; noDAC_PU1mut, blue; noDAC_PU1, green; DAC_PU1mut, brown; DAC_PU1, goldenrod).

As illustrated in the MDS plot, dimension 1 separates DAC-treated from untreated CTV-1 cells, while dimension 2 separates PU.1-transfected from PU.1mut- and accordingly mock-transfected cells with regard to the log-fold change (logFC) of expressed genes between the samples. To further analyze the relationship between DNA methylation and PU.1 expression, RNAseq data of PU.1-transfected cells of both conditions was further processed for differential gene expression (DGE) analysis. Therefore, a conservative quasi-likelihood F (QLF)-test was applied. This test is preferred when the number of replicates is small as it reflects the uncertainty in estimating the dispersion for each gene and provides a more robust and reliable error rate control. Consequently, the data set was conducted to q-statistics testing, which is a statistical test used for multiple significance testing across a number of means (the corresponding multi-variance analysis is pictured in **supplementary Figure 10-1**). To visualize the differential expressed genes (DEGs) between the two populations, we made use of a hierarchical Z-score clustering, with a Z-score being the number of standard deviations (SDs) from the mean a data point is. This means it is a measure of how many SDs below or above the analyzed population mean a raw data score is and it represents the distance between the mean and the measured observation. The corresponding heat map is shown in **Figure 5-8** and it exemplifies that the expressed genes between PU.1- and PU.1mut-transfected CTV-1 cells untreated or treated with DAC split into three distinct clusters. Cluster 1 (green bar) represents the major group and contains genes upregulated by PU.1-transfection in DAC-treated cells. Cluster 2 (blue bar) distinguishes DAC-treated vs. untreated CTV-1 cells and cluster 3 (red bar) PU.1-transfected vs. PU.1mut-transfected cells, respectively. Moreover, a volcano plot was generated. This type of a scatter plot is useful to quickly identify changes in large data sets composed of replicate data. It is constructed by plotting the negative log10 of the p-value on the y-axis, which results in data points with low p-values (highly significant) appearing toward the top of the plot. The x-axis comprises the log of the fold-change between the two conditions, which is used so that changes in both directions appear equidistant from the center. Regions of interest in the plot are those points that are found toward the top of the plot that are far too either the left- or right-hand side. As seen in the volcano plot of **Figure 5-8b**, genes differentially expressed between PU.1-transfected cells in DAC vs. mock conditions, are mainly genes upregulated by PU.1 in the DAC-treatment (234 genes), whereas a smaller fraction of genes (47 genes) is specific for mock-treated PU.1-transfected cells.

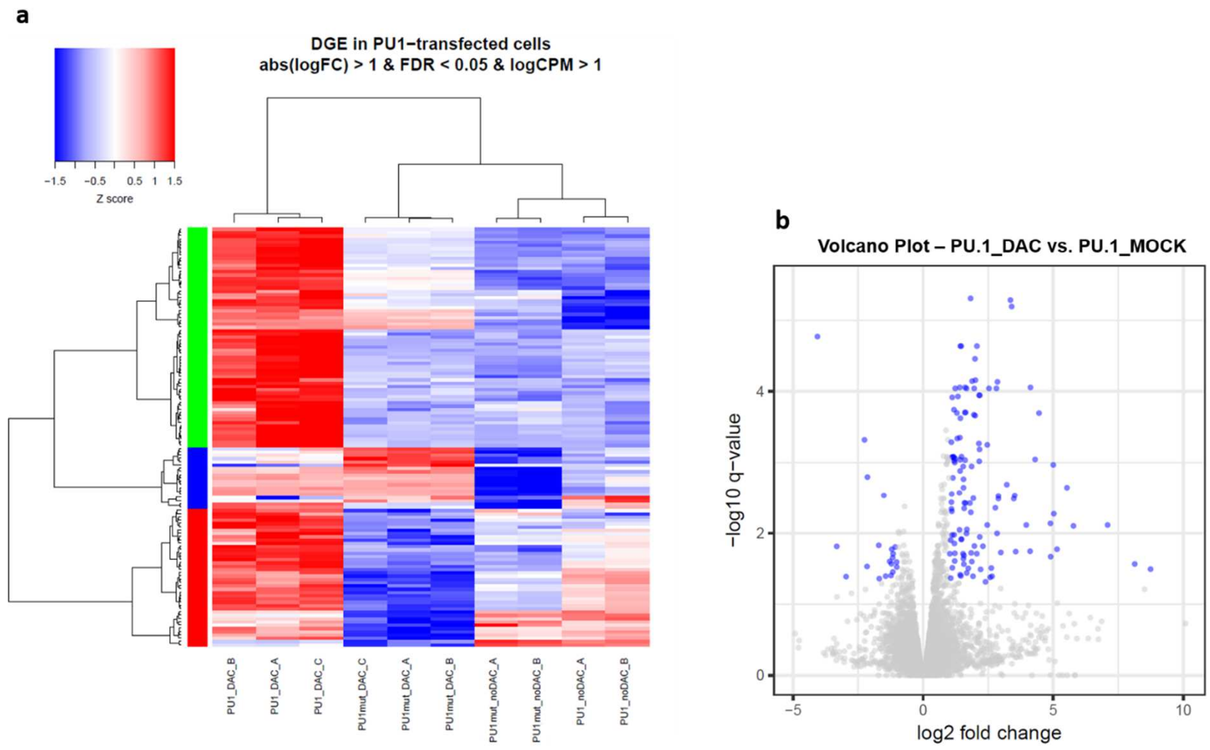


**Figure 5-8 - DGE in PU.1-transfected CTV-1 cells treated with DAC vs. untreated**

**a |** Hierarchical Z-score clustering of differential expressed genes of PU.1- and PU.1mut-transfected cells treated with DAC or left untreated. PU.1mut-transfected (mutPU1_noDAC_A/B; mutPU1_DAC_A/B/C ) and PU.1-transfected (PU1_noDAC_A/B; PU1_DAC_A/B/C ) CTV-1 cells are depicted. Only genes with an absolute logFC > 1, logCPM > 1 and a FDR < 0.05 were selected. Dendrogram shows clustering of the data into three groups. Each horizontal line represents a single differential expressed gene while the corresponding Z-score is indicated with the cluster color. The colored bars on the left side show classification of DEGs into the three different clusters. **b |** Volcano plot showing genes upregulated in PU.1-transfected DAC-treated CTV-1 cells compared to PU.1-transfected untreated cells (both normalized to the corresponding PU.1mut transfection). The significance is plotted against the fold-change. Blue dots represent genes with a FDR < 0.05 and a logFC > 1.

To analyze the genes upregulated in PU.1-transfected, treated CTV-1 cells in more detail, a heat map (done with Metascape) of the enriched GO-terms was generated and a scatter plot exemplifying the correlation of these genes with the PU.1 expression across divers blood cell types was conducted (see **Figure 5-9**).

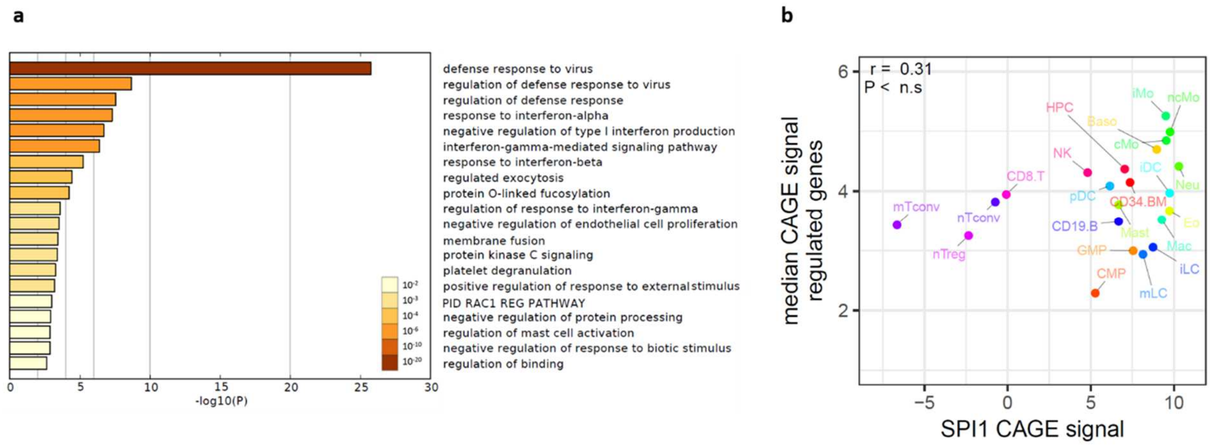


**Figure 5-9 - GO-terms & blood expression of 2-fold upregulated genes in PU.1-transfected, DAC-treated CTV-1 cells**

**a |**Gene ontology (GO) analysis of PU.1-induced upregulated genes in DAC-treated CTV-1 cells ran with Metascape. The significance of the enrichment of a particular term is depicted with the log10 of the p-value and indicated with coloring. **b |** Comparison of PU.1-induced upregulated genes with the SPI1 CAGE (Cap Analysis of Gene Expression) expression data from the FANTOM Consortium across various human blood cell types. The median CAGE signal of the PU.1-regulated genes is plotted on the y-axis, whereas the SPI1 CAGE signal of different blood cells is plotted on the x-axis. Correlated cell types share similar colors. Cell types with high PU.1 expression levels are located in the upper right-hand side of the plot. The Pearson correlation is given.

The gene ontology (GO) analysis suggests, that genes which are upregulated by PU.1 overexpression in the DAC-treated cells, are mainly genes involved in immune response (e.g. defense response to virus and response to interferon-gamma). As depicted in the scatter plot in **Figure 5-9b**, genes induced at least 2-fold correlate partly with the PU.1 expression in diverse blood cell subsets. Expression data was obtained from CAGE data of the FANTOM Consortium. CAGE is used to comprehensively map the vast majority of human TSSs and their promotors, thus correlating with active gene expression. On the y-axis, the median CAGE signal of the upregulated genes is plotted across distinct human blood cells, whereas the SPI1 (PU.1) CAGE signal is plotted on the x-axis. Myeloid cells like MO or neutrophils (green), as well as DC (turquoise) show the highest PU.1 expression, which is correlated with the expression of the genes upregulated by PU.1 in lymphoid CTV-1 cells treated with DAC (r=0.31). However, this correlation is not significant. Taken together, PU.1 overexpression in CTV-1 cells treated with DAC induces a distinct transcriptome when compared to untreated and un-transfected cells.

Since PU.1 overexpression in DAC-treated CTV-1 cells induced aforementioned changes in their gene expression profile, the impact of DNA methylation on PU.1 binding site selection and the chromatin structure was analyzed as well. Therefore, ChIPseq and ATACseq were performed in addition (see sections 4.2.6 – 4.2.8). Detailed computational analysis is listed in section 10.1.4 (‘Basic analysis of ATAC & ChIPseq data’). Since ChIPseq data of CTV-1 cells done with an anti-PU.1 or anti-FLAG antibody respectively showed a good Pearson correlation ($r^2$=0.705), all subsequent ChIPseq data used was conducted with the monoclonal anti-FLAG antibody unless otherwise noted (see **supplementary Figure 10-2**). An overview of all sequencing-based methods used for the DAC-treated CTV-1 cells is shown in the following IGV genome browser track **(Figure 5-10)**.

To analyze the ATAC signals and PU.1 binding sites of DAC-treated cells compared to conventional CTV-1 cells, obtained raw read data was mapped to the human reference genome (hg19; see sections 10.1.1.1-10.1.1.2), CNV-normalized and data sets of corresponding replicates were merged (three replicates each; besides PU.1 ChIPseq of conventional cells, n=2) . For visualization, bigWigs were generated and the signals were averaged across replicates.

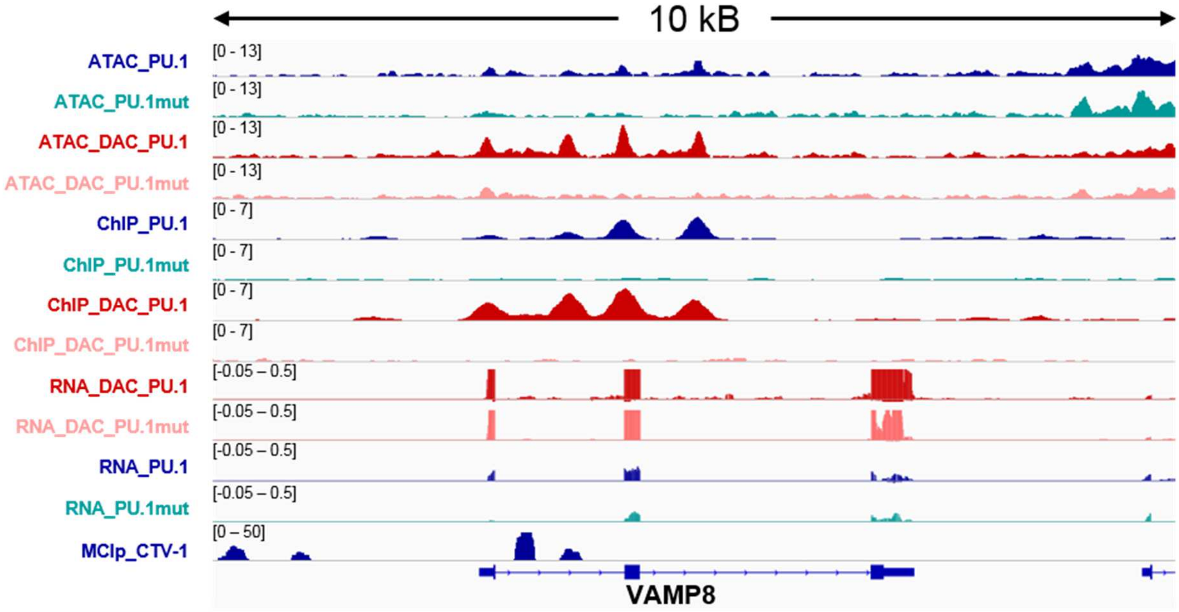


**Figure 5-10 - IGV genome browser track of transfected, DAC-treated CTV-1 cells across the VAMP8 locus**

The ATACseq signal of untreated CTV-1 cells transfected with PU.1 (blue) and PU.1mut (turquoise) mRNA and the corresponding reads of DAC-treated PU.1 (red) and PU.1mut (light red) transfected cells are shown in the first four rows. ChIPseq PU.1 reads, as well as the RNAseq signals for all conditions are shown across the VAMP8 region. The last row depicts the MCIp-seq signal in the cell line. The exact location of the gene is depicted. All ATACseq and ChIPseq reads are CNV-normalized and averaged across replicates (chr2:85,801,796-85,811,973).

As seen in **Figure 5-10** DAC-treated cells show a PU.1-induced increase in the corresponding ATAC signal when compared to untreated CTV-1 cells in the selected region. This is further accompanied by the gain of PU.1 binding in this region. Moreover, the VAMP8 gene is almost exclusively transcribed in CTV-1 cells treated with DAC (RNAseq tracks), which is inversely correlated with the methylation status of the gene as depicted by MCIp-seq. To analyze the chromatin accessibility and the binding preferences of PU.1 in relation to epigenetic mechanisms, ATACseq as well as PU.1 ChIPseq peaks of conventional and DAC-treated CTV-1 cells were determined first. ChIPseq peaks of merged data sets were obtained by normalization to the corresponding input and an individual peak was called, when at least 15 PU.1 reads with a FDR below 0.05 were present. The resulting peak sets were then further filtered to include the mapability across the reference genome (hg19). In terms of ATACseq peaks, the regions were stitched to nucleosome size regions (150 bp) with a stringent FDR cut-off below 0.05. Furthermore, the peaks were also centered on 250 bp regions with a stringent FDR cut-off below 0.05 as well. In a final step, the peak files were joined, to obtain the final nucleosome-centered ATAC peak set for each sample. To restrict the analysis to regions depending on the DNA methylation status and thus on epigenetic regulation, differential ATACseq peaks were generated next. Therefore, the corresponding ATACseq peak sets of PU.1-transfected cells were intersected to obtain differential peaks between DAC-treated vs. conventional CTV-1 cells only. Moreover, a second differential peak set, containing only DAC-specific ATACseq peaks from PU.1-overexpressing cells was generated by comparison with the ATACseq peak set of untreated PU.1-overexpressing cells. This analysis resulted in roughly 33.000 ATACseq peaks made accessible during the DAC-treatment induced by the PU.1 expression, and in nearly 14.200 new peaks, which were specifically gained upon DAC-treatment. The overlap of these peaks is shown in the following Venn diagram **(Figure 5-11a)**. The vast majority of newly induced open chromatin regions is due to the PU.1 overexpression in the cell line (PU.1 induced, 25406 peaks), whereas a smaller proportion of peaks is specifically induced by the DAC-treatment itself in PU.1-expressing CTV-1 cells (DAC specific, 6572 peaks).

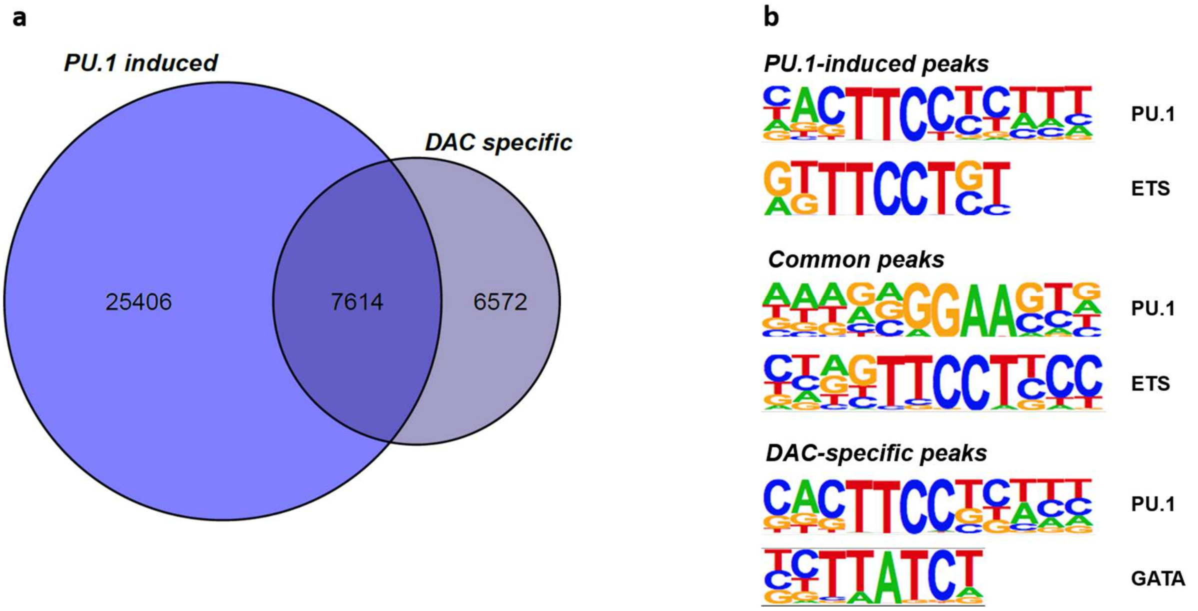


**Figure 5-11 - Differential ATACseq peak distribution and motif signature of DAC-treated CTV-1 cells**

**a |** Venn diagram displaying the correlation of differential ATACseq peaks between PU.1-induced (in comparison to PU.1mut) and DAC-specific (in comparison to noDAC) ATACseq peaks in PU.1-transfected CTV-1 cells. **b |** *De novo* found motif signatures across the differential peak sets analyzed with HOMER (Heinz et al. 2010). The top two motifs hits for each condition are shown. All motifs were significantly enriched above the background with a corresponding p-value of at least 1e$^{-30}$.

Moreover, PU.1-induced, DAC-specific and shared ATACseq regions between both sets (common peaks) were combined with PU.1 ChIPseq reads of PU.1-transfected DAC-treated cells to focus on accessible regions bound by PU.1. The percentage of bound vs. non-bound peaks in common and PU.1-induced regions is nearly 50%, whereas DAC-specific ATACseq regions are only bound by PU.1 in about 14% of the peaks.  In addition, the different regions show distinct co-associated motif signatures in their ATACseq regions bound by PU.1 as depicted in **Figure 5-11b**. PU.1-induced and common peaks are mainly enriched for PU.1 (74% of targets compared to the background) and ETS (37% and 22% of targets respectively) consensus motifs, whereas DAC-specific peaks are enriched for PU.1 (50% of targets) and GATA (22% of targets) TF motifs. Besides, we studied the CpG- and GC-distribution of PU.1-bound regions in the PU.1-induced and DAC-specific ATACseq peak set respectively, to gain insights into epigenetically induced changes in PU.1 binding site selection. Therefore, our peak sets were annotated and the corresponding CpG- and GC-distribution of the PU.1 binding sites across both sets was calculated. The results of this analysis are shown in the bean plots in **Figure 5-12**.

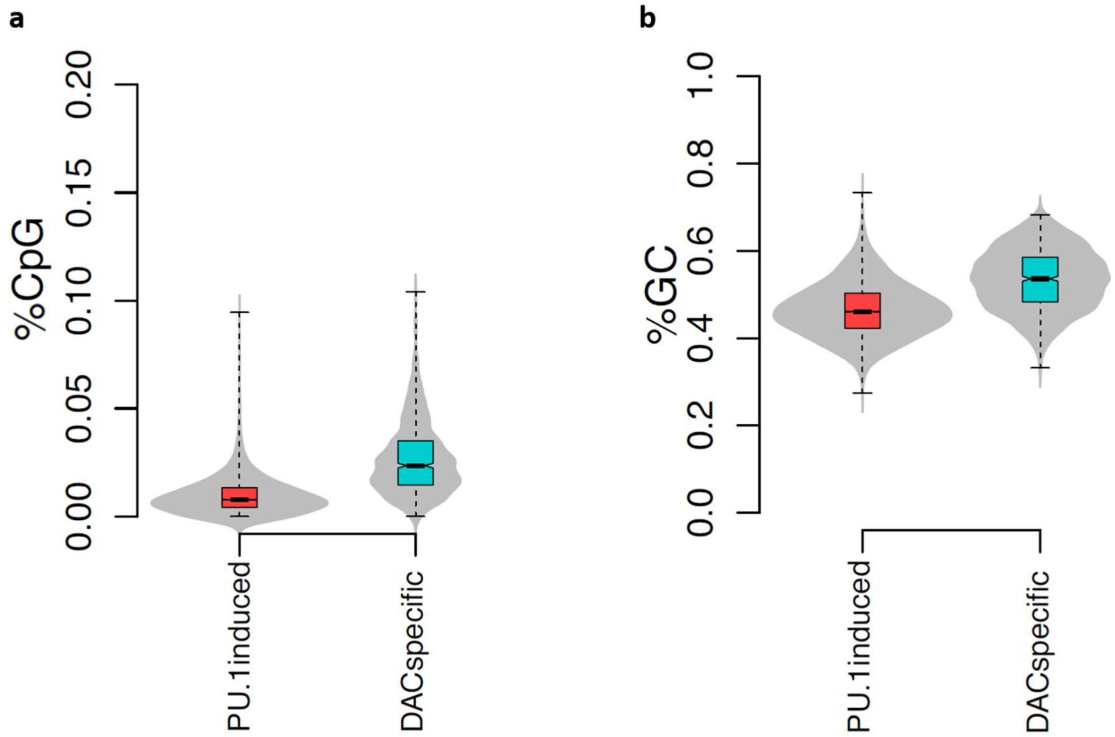


**Figure 5-12 - CpG- and GC-content of PU.1 binding sites in DAC-treated CTV-1 cells**

**a |** The bean plot shows the CpG-content (%) of PU.1 binding sites in PU.1-induced (red boxplot), PU.1-bound ATACseq regions compared to DAC-specific (turquoise boxplot), PU.1-bound regions. The median of the specific distribution across all regions is depicted inside the bean with a conventional boxplot. **b |** The bean plot shows the GC-content (%) of PU.1 binding sites in PU.1-induced (red boxplot), PU.1-bound ATACseq regions compared to DAC-specific (turquoise boxplot), PU.1-bound regions. The median of the specific distribution across all regions is depicted inside the bean with a conventional boxplot.

The bean plot in **Figure 5-12a** illustrates, that mainly binding sites of DAC-specific regions show a high CpG-content. Moreover, the overall GC-content is also enriched in the DAC-specific peak set **(Figure 5-12b)**. This trend is also seen in the motif signature of the PU.1 consensus site in DAC-specific peaks in **Figure 5-11**. Taken together, this indicates that the DAC-treatment, which induces a genome-wide demethylation in the CTV-1 cells, might be able to open up non-accessible PU.1 binding sites. Likewise, the distinct peak sets were also compared in terms of DNA methylation as measured by MCIp-seq. Here, the specific regions were first combined with a set containing all MCIp-detectable regions (eight independent replicates of SsI-fully methylated CTV-1 gDNA), before the aforementioned MCIp-seq set of the CTV-1 cells (see **Figure 5-10**) was annotated. Then a histogram plot, comparing PU.1-induced and DAC-specific regions was generated **(Figure 5-13)**. In addition, the association of the PU.1 ChIPseq coverage across the different peak sets is also shown in this figure.

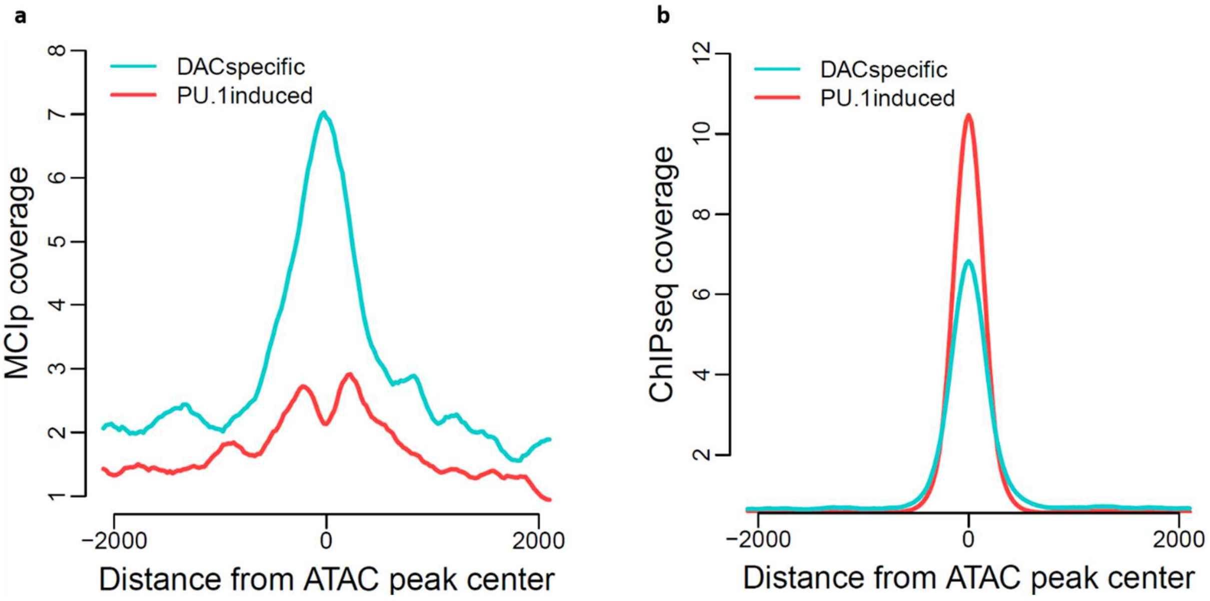


**Figure 5-13 – MCIp-seq and PU.1 ChIPseq coverage of DAC-specific & PU.1-induced CTV-1 ATACseq regions bound by PU.1**

**a |** The histogram shows the comparison of the MCIp-seq coverage of PU.1-induced (red) and DAC-specific (turquoise) ATACseq regions of treated CTV-1 cells bound by PU.1. **b |** The histogram shows the comparison of the PU.1 ChIPseq coverage of PU.1-induced (red) and DAC-specific (turquoise) ATACseq regions of treated CTV-1 cells bound by PU.1

The histograms depicted show that the DAC-specific regions are highly associated with regions methylated in conventional CTV-1 cells, whereas these regions are rarely associated with differential, PU.1-induced ATACseq peaks. An inverse trend however, is seen for the correlation of the differential ATACseq peak sets and their corresponding PU.1 ChIPseq coverage. Here PU.1-induced ATACseq regions are mainly the ones, which overlap with the overall PU.1 ChIPseq coverage in DAC-treated CTV-1 cells, nonetheless DAC-specific ATACseq regions of those cells are also remarkably associated with PU.1 binding.

In summary, the analysis of chromatin accessibility and PU.1 binding in PU.1-expressing CTV-1 cells in terms of epigenetics implies that the DNA methylation landscape can – at least in part - explain a fraction of PU.1 binding events in the analyzed cells.

## 5.3 PU.1 Binding Site Selection in Lymphoid CTV-1 Cells

PU.1 binding analyses in PU.1-expressing cell types only provides a static system to study its binding site selection in terms of chromatin accessibility. Using such model systems, one is not able to distinguish bound vs. non-bound sites with regard to the particular chromatin structure of the analyzed cell type. The general question, if PU.1 induces local changes in the chromatin structure to gain access to its binding sites, or if PU.1 binds to the pre-established chromatin structure correlated with its own endogenous expression still remains. To analyze PU.1 binding dynamics in an unbiased way, we made use of the aforementioned lymphoid leukemia cell line CTV-1. This cell line does neither express PU.1, nor its ETS-family members SpiB and SpiC, which share the same consensus binding site. In addition, its known hetero-dimerization partners IRF4 and IRF8 are not expressed **(Figure 5-14)** as analyzed by RNAseq, which will be explained in detail below. For this reason CTV-1 cells serve as the ideal model to study PU.1 binding and hence its dependency and impact on the existing chromatin structure.

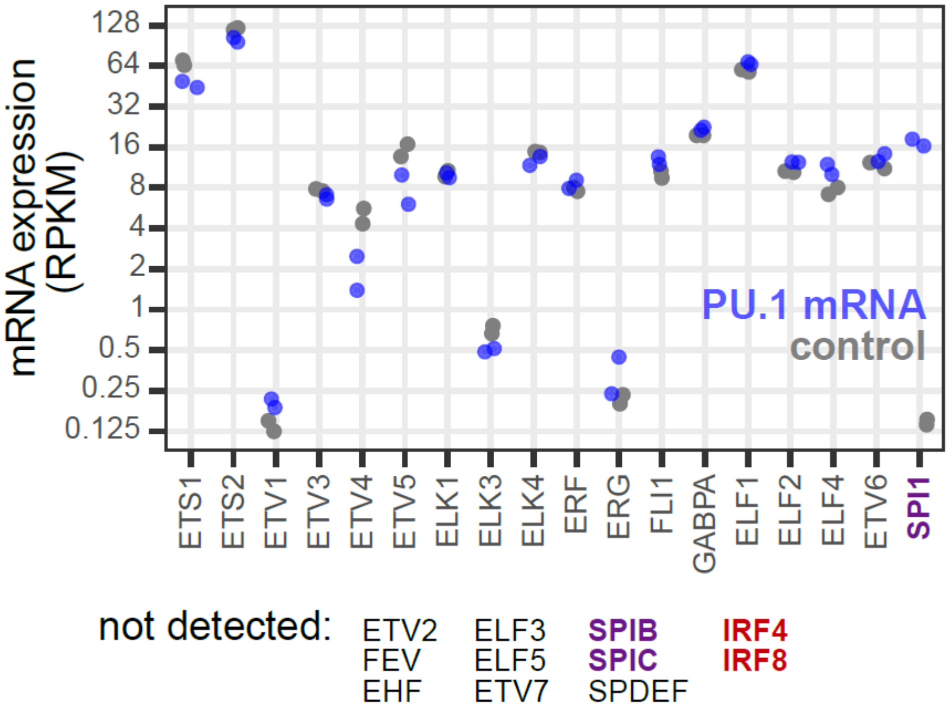


**Figure 5-14 - mRNA expression profile of all ETS-family TFs & PU.1-heterodimerization partners IRF4 & IRF8 in CTV-1 cells**

RNAseq profile of lymphoid CTV-1 cells transfected with PU.1 (blue) and PU.1mut mRNA (control; grey) respectively. mRNA expression of ETS-family TFs and overexpressed PU.1 (SPI1; purple) is shown in the scatter plot in reads per kilobase million (RPKM). ETS-family TFs and hetero-dimerization partners (IRF4/IRF8; red) not detected are listed below. PU.1 ETS-family members SPIB & SPIC, which recognize the same consensus motif, are highlighted in purple.

Overexpression of PU.1 was accomplished as already described in chapter 5.2.2 and is summarized in the following figure **(Figure 5-15)** next to the corresponding IGV genome browser track, which shows an overview of all sequencing-based utilized methods to study PU.1 binding preferences in CTV-1 cells.

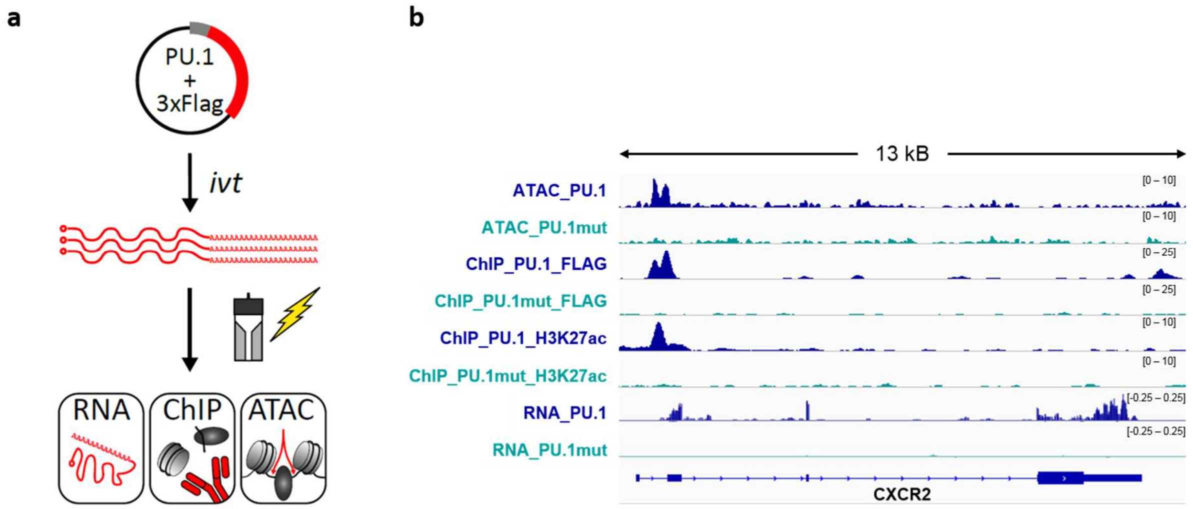


**Figure 5-15 - Methodology & IGV genome browser track of transfected CTV-1 cells across the CXCR2 locus**

**a |** Schematic overview of PU.1 overexpression procedure in CTV-1 cells. *In vitro* transcribed (ivt) mRNA was electroporated into the cell line and the transcriptome (RNA) as well as the chromatin landscape (ATAC) and the PU.1 binding site selection (ChIP) was analyzed. **b |** The ATACseq signal of CTV-1 cells transfected with PU.1 (blue) and PU.1mut (turquoise) mRNA, the corresponding ChIPseq PU.1- and H3K27ac reads, as well as the RNAseq signals for both conditions are shown across the CXCR2 region located on chromosome 2. The exact location of the gene is depicted. All ATACseq and ChIPseq signals are CNV-normalized and averaged across replicates (chr2:218,989,323-219,002,950).

As already implied in the above figure, PU.1 binding correlates with changes in the local chromatin structure. PU.1-transfected, but not PU.1mut-transfected, CTV-1 cells gain an ATACseq signal in the selected region upon PU.1 binding, which is a measure of open chromatin. This is alongside with the deposition of the active histone mark H3K27ac and the expression of the CXCR2 mRNA as analyzed by RNAseq.

### 5.3.1 PU.1-induced Transcriptome in Lymphoid CTV-1 Cells

To analyze gene expression changes induced by PU.1 in lymphoid CTV-1 cells, PU.1wt and PU.1mut mRNA was transfected into the cells. A mock control was further included. Total RNA was harvested after 24 h and further processed using RNAseq (see sections 4.2.5.1 & 4.2.9). Experiments were repeated once, resulting in two replicates each. Moreover, to study long-term effects of PU.1 overexpression, CTV-1 cells were transfected once for five consecutive days with PU.1wt and PU.1mut mRNA prior to total RNA isolation. The detailed computational analysis is depicted in section 10.1.5 (‘Basic analysis of RNAseq data’). After mapping of the raw reads (see section 10.1.1.3) and generating a combined read count table including the raw data of all samples, the data was further analyzed using edgeR (Robinson et al. 2010). In a first step, the particular treatment was defined and batch effects were removed, before the corresponding counts per million (cpm) were calculated and log-transformed.

Scaled, batch-corrected logCPM data was subjected to a principal component analysis (PCA) to address the initial question how related the different transfection conditions are (see **Figure 5-16**). This unsupervised technique reduces high dimensionality data sets to fewer dimensions allowing for easier interpretation of the data.

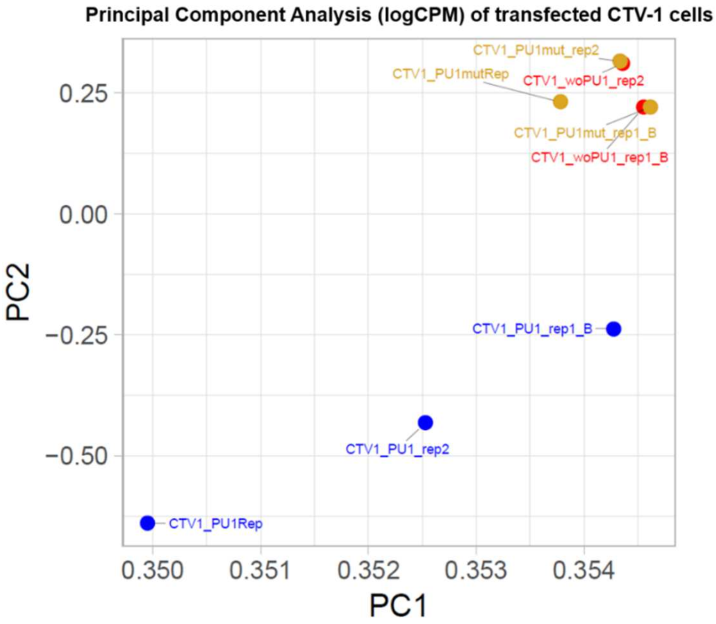


**Figure 5-16 - PCA of the logCPM data of transfected CTV-1 cells**

PCA plot showing the correlation of the logCPM values of PU.1 (blue), PU.1mut (goldenrod) and mock transfected (red) CTV-1 cells. Samples are distributed between principal component 1 and 2 (PC1/PC2).

Principal component analysis included RNAseq data from all mentioned conditions and controls. In general, the RNAseq data of CTV-1 cells transfected with PU.1wt mRNA was found to be in close proximity to each other illustrating the similarity between the replicates and the repetitive PU.1-transfection. In contrast, control cells transfected with PU.1mut mRNA or mock-transfected were localized further away, but in close proximity to each other, with mock and mutant cells even overlapping. In summary, principal component 2 was able to separate PU.1-transfected and PU.1mut- or rather un-transfected cells. To study this further, differential expressed genes between PU.1wt- and PU.1mut-transfected cells were extracted and conducted to q-statistics testing using a conservative QLF-test already described in chapter 5.2.2 (the corresponding multi-variance analysis plot is pictured in **supplementary Figure 10-3**). After, hierarchical Z-score clustering was applied to visualize the distribution of genes differentially expressed in PU.1-transfected cells between all conditions. The according heat map, as well as the corresponding volcano plot, illustrating myeloid genes upregulated in PU.1-transfected vs. PU.1mut-transfected cells, is shown in **Figure 5-17**.

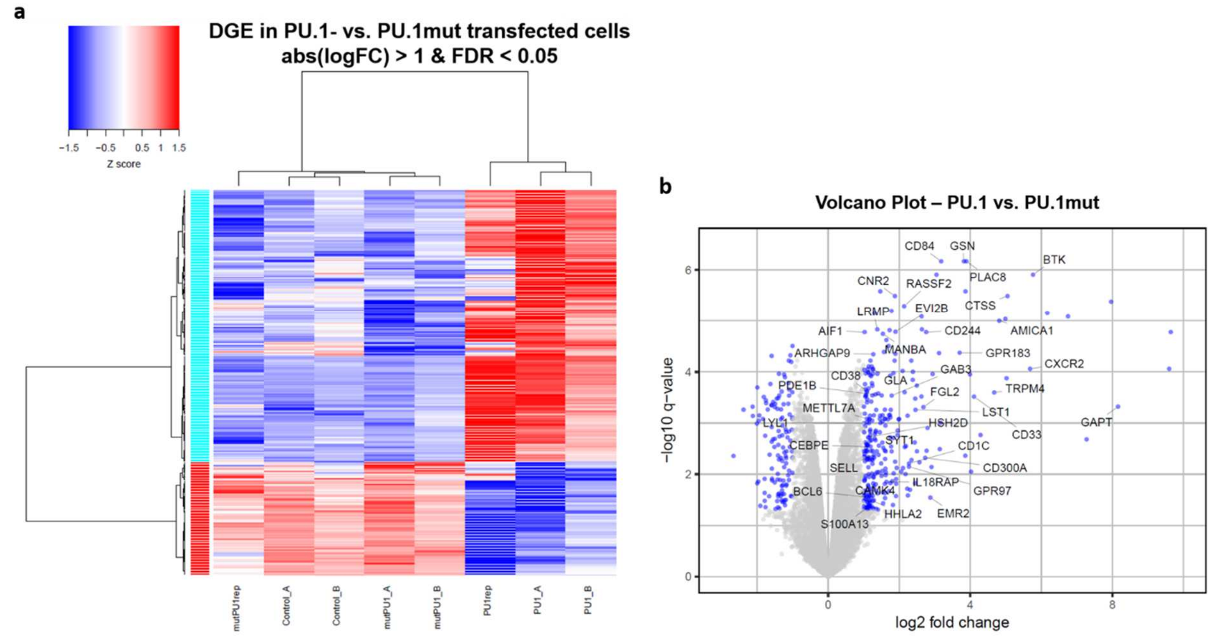


**Figure 5-17 - DGE in PU.1- vs. PU.1mut-transfected CTV-1 cells**

**a |** Hierarchical Z-score clustering of differential expressed genes of PU.1- and PU.1mut-transfected cells. Mock-transfected (Control_A/B), PU.1mut-transfected (mutPU1_A/B & mutPU1rep) and PU.1-transfected (PU1_A/B & PU1rep) CTV-1 cells are depicted. Only genes with an absolute logFC > 1 and a FDR < 0.05 were selected. Dendrogram shows clustering of the data into two groups. Each horizontal line represents a single differential expressed gene while the corresponding Z-score is indicated with the cluster color. The colored bars on the left side show classification of DEGs into the two different clusters. **b |** Volcano plot showing genes upregulated in PU.1-transfected CTV-1 cells, when compared to PU.1mut-transfected cells. The significance is plotted against the fold-change. Blue dots represent genes with a FDR < 0.05 and a logFC > 2. Only myeloid genes upregulated in PU.1-transfected cells are depicted.

The heat map exemplifies that differential expressed genes of PU.1- vs. PU.1mut-transfected CTV-1 cells split into two distinct clusters. In detail, PU.1-transfected cells cluster together with CTV-1 cells repetitive transfected with PU.1, whereas control cells (mock-transfected) and PU.1mut-transfected cells, short- and long-term respectively, fall into a separate cluster. Moreover, a closer look at the genes induced by PU.1 expression in CTV-1 cells implies that upregulated genes are mainly regulators of the myeloid lineage. CD84 and CXCR2 (upper right-hand side of the volcano plot) for example, are both induced during myelopoiesis and are usually not expressed in lymphoid cells like CTV-1. In addition, these genes are correlated with the PU.1 expression pattern in various blood cell subsets as shown in **Figure 5-18a**.

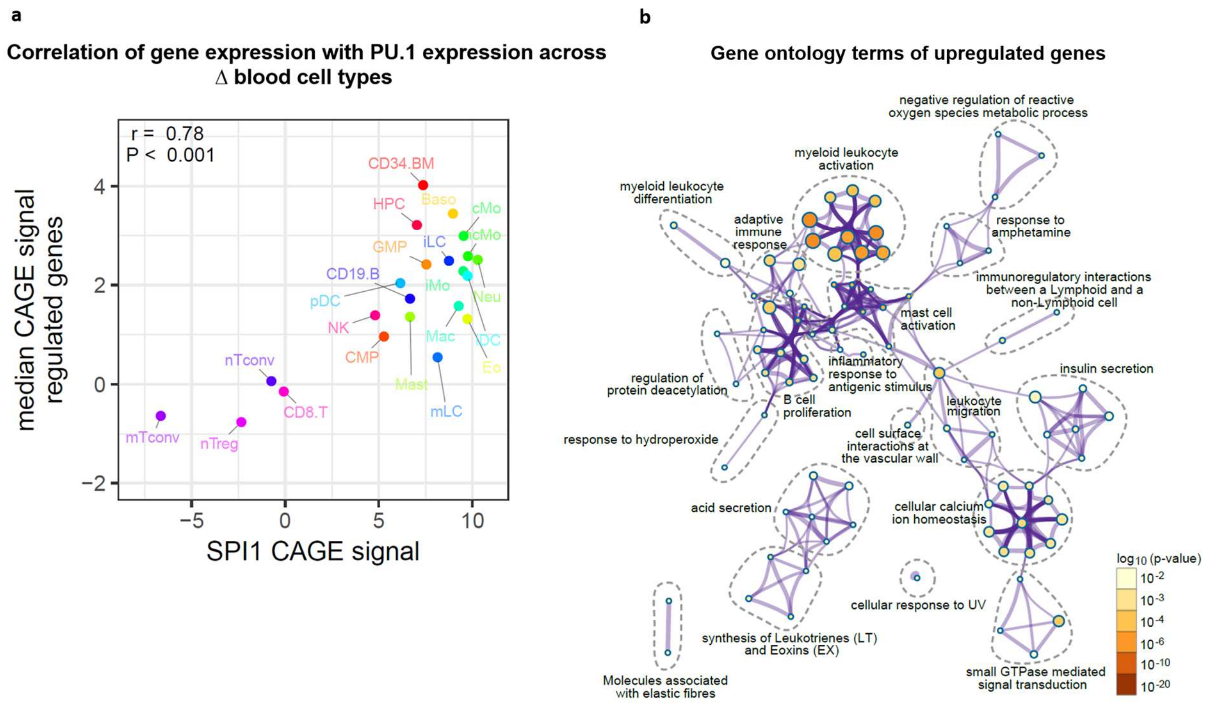


**Figure 5-18 - Correlation of induced gene expression with PU.1 expression in human blood cells & gene ontology analysis**

**a |** Comparison of PU.1-induced upregulated genes (3-fold) with the SPI1 CAGE (Cap Analysis of Gene Expression) expression data from the FANTOM Consortium across various human blood cell types. The median CAGE signal of the PU.1-regulated genes is plotted on the y-axis, whereas the SPI1 CAGE signal of different blood cells is plotted on the x-axis. Correlated cell types share similar colors. Cell types with high PU.1 expression levels are located in the upper right-hand side of the plot. The Pearson correlation as well as the p-value are given. **b |**Gene ontology (GO) analysis of PU.1-induced upregulated genes in CTV-1 cells ran with Metascape. The significance of the enrichment of a particular term is depicted with the log10 of the p-value and indicated by the node coloring.

As exemplified in the scatter plot, genes induced at least 3-fold by PU.1 expression in CTV-1 cells correlate with the PU.1 expression in diverse blood cell types. Expression data was obtained from CAGE data of the FANTOM Consortium. CAGE is used to comprehensively map the vast majority of human TSSs and their promotors, thus correlating with active gene expression. On the y-axis, the median CAGE signal of the upregulated genes is plotted across distinct human blood cells, whereas the SPI1 (PU.1) CAGE signal is plotted on the x-axis. Myeloid cells like MO or neutrophils (green), as well as MAC and DC (turquoise) show the highest PU.1 expression, which is significantly correlated with the expression of the genes upregulated by PU.1 in lymphoid CTV-1 cells (r=0.78). This indicates that induced genes show a positive correlation with the endogenous PU.1 expression in the myeloid lineage. A gene ontology (GO) analysis of these upregulated genes ran with Metascape strengthens this result (see **Figure 5-18b**), since PU.1-induced genes are mostly enriched for terms corresponding to the myeloid lineage, like myeloid leukocyte activation or differentiation respectively. Taken together, PU.1 expression in the lymphoid cells leads to the upregulation of genes belonging to the myeloid lineage, thus changing the endogenous gene expression profile of the non-PU.1-expressing cells.

## 5.3.2 PU.1-induced Chromatin Accessibility in Lymphoid CTV-1 Cells

Besides gene expression changes, we also aimed to analyze the influence of PU.1 expression on the existing chromatin landscape. Hence, we mapped PU.1 binding sites as well as the accessibility of the chromatin in CTV-1 cells using ChIP and ATACseq respectively (see sections 4.2.6-4.2.8). Detailed computational analysis is listed in section 10.1.5 ('Basic analysis of ATAC & ChIPseq data'). To analyze the ATACseq signals, ChIPseq obtained binding sites as well as the deposition of the active histone mark H3K27ac, generated raw read data was mapped to the human reference genome (hg19; see sections 10.1.1.1-10.1.1.2), CNV-normalized and data of corresponding replicates was merged (three replicates each; two for ETS-1 and FLI-1). For visualization, bigWigs were generated and the signals were averaged across replicates. The peak finding was essentially done as already described before for the ATACseq and ChIPseq data sets resulting in roughly 45.000 PU.1 peaks in the lymphoid cell line (see section 5.2.2). The localization of the open chromatin regions (ATACseq signal) and the H3K27ac and PU.1 peaks is already depicted above **(Figure 5-15)**. To study the binding preferences of the TF PU.1 with regard to the chromatin landscape of lymphatic CTV-1 cells, generated ATACseq signals of PU.1- and PU.1mut-transfected cells were annotated, merged and used for an explorative K-means clustering approach. This unsupervised machine learning approach is based on a cluster algorithm, which splits the data into a set of clusters based on the distances between each data point and the center location of each cluster. The goal of this algorithm is to find groups in the data, with their number represented by the variable K. The algorithm works iteratively to assign each data point to one of the K-groups based on the features that are provided so that the data points are clustered based on feature similarity. For our approach, we set the variable K from 9 to 16 and annotated all merged ATACseq and ChIPseq (PU.1 and H3K27ac) reads of PU.1- and PU.1mut-transfected cells with the generated cluster data to draw histogram plots for visualization of the different cluster sizes and the corresponding grouping of our data set. Based on this analysis, the K-means cluster solution splitting the data into 14 distinct groups was used for all further computational analysis. Consequently, all relevant data sets were subjected to the 14 K-means cluster solution and the corresponding ChIPseq and ATACseq peaks were annotated generating merged peak files. After the size of each cluster was determined and the 14 generated clusters were ordered in terms of chromatin accessibility based on their individual ATACseq signals. Furthermore, the average remodeling index of each cluster was calculated as the ratio of the ATACseq signal of PU.1-expressing vs. non-expressing CTV-1 cells. This index was used as a criterion for PU.1-induced changes in the chromatin landscape of the lymphoid cells, since it correlates with the accessibility of the chromatin upon PU.1 expression when compared to conventional PU.1mut-transfected CTV-1 cells. To visualize the correlation of the obtained ATAC and ChIPseq data of PU.1-transfected vs. PU.1mut-transfected cells, histogram plots across the 14 ATACseq K-means cluster were compiled across determined PU.1 binding sites **(Figure 5-19)**.

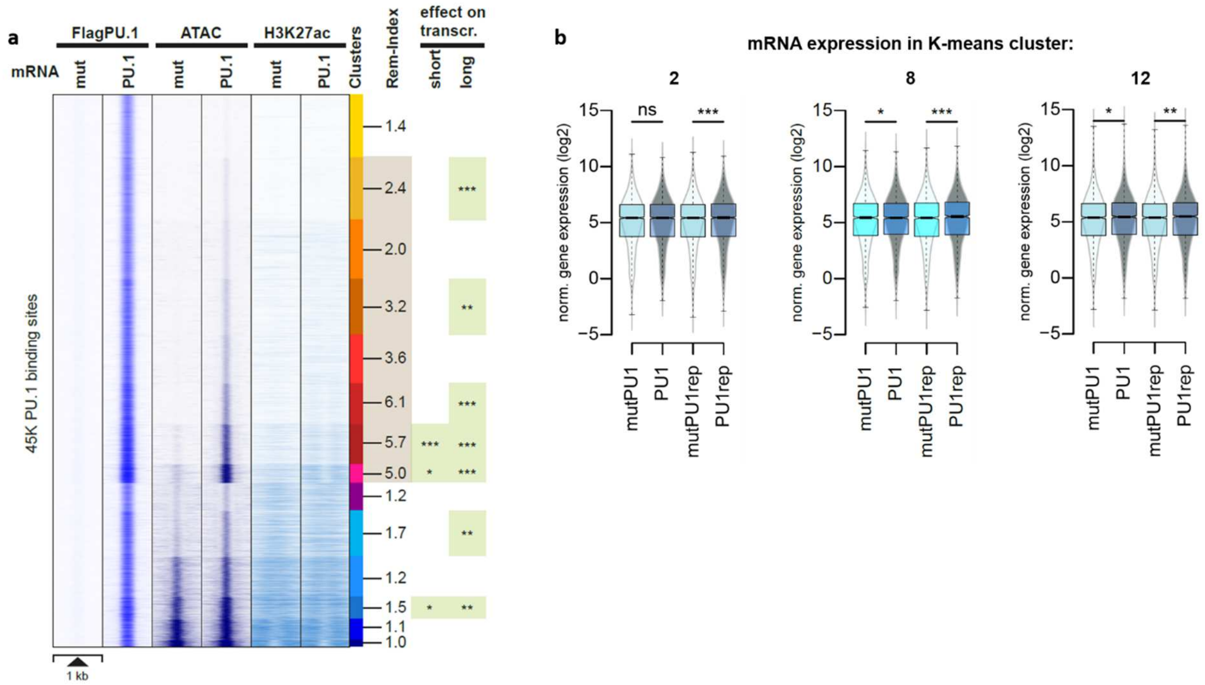


**Figure 5-19 - K-means clustering of ATACseq, ChIPseq & RNAseq data of PU.1-transfected CTV-1 cells**

**a |** The distribution of the PU.1 ChIPseq (mut & PU.1), ATACseq (mut & PU.1) and H3K27ac ChIPseq (mut & PU.1) signals of PU.1- vs. PU.1mut transfected cells are plotted across the 45.000 PU.1 binding sites obtained in CTV-1 cells. The 14 K-means cluster generated out of the ATACseq signal are shown with the color bar on the right, next to the corresponding remodeling index (Rem-Index) of each cluster. The PU.1-induced effect in terms of transcription is also depicted for each cluster. Statistical significance was calculated with a paired Wilcox test ('*' p-value < 0.05, '**' p-value < 0.01, '***' p-value < 0.001). **b |** The mRNA expression of CTV-1 cells transfected with PU.1 or PU.1mut mRNA is shown in association to the generated K-means cluster of the ATACseq signals of transfected cells for three representative clusters. Short-term (mutPU1, PU1) and repetitive long-term transfections (mutPU1rep, PU1rep) used for RNAseq are depicted.

As seen in Figure **5-19** the explorative generated clusters split up according to pre-existing accessible vs. non-accessible chromatin states as depicted by the ATACseq signal of PU.1mut-transfected cells. In addition, PU.1-transfected cells gain an ATACseq signal in certain regions, when compared to control transfected cells (clusters 2-8, dark-yellow until pink).  This effect is also visible with the above mentioned remodeling index, which is depicted on the right and which is mainly increased in clusters 6-8 (red until pink). Moreover, the deposition of the active histone mark H3K27ac is also strongly associated with open chromatin regions in PU.1- and PU.1mut-transfected cells, respectively. Furthermore, we screened for expressed genes associated with our 14 K-means cluster to obtain the corresponding gene expression data for each cluster. For this approach, the normalized and corrected read count table containing the logCPM values of PU.1-transfected CTV-1 cells was used (see section 5.3.1). This table was joined with the list of expressed genes obtained out of the ATACseq data, which was generated via overlapping the data with all annotated human enhancer regions.  After the mRNA expression was plotted across the 14 clusters and histogram plots were drawn as seen in **Figure 5-19b** for 3 selected clusters (remaining plots are depicted in **supplementary Figure 10-4**). As seen here, cluster 2 for example represents a cluster, where the PU.1 expression in the lymphoid cells induces the opening of the chromatin (Rem-Index=2.4), which is further accompanied with a significant effect on the corresponding transcriptome of the cells, when treated in a long-term approach.

In cluster 8 on the other hand, also a cluster where PU.1 induces severe changes in the local chromatin architecture, the induced transcriptome is already affected by short-term PU.1 expression. In contrast, cluster 12 represents a cluster, which is already accessible in non-PU.1-expressing CTV-1 cells (Rem-Index=1.5), however this cluster is also associated with significant gene expression changes upon PU.1 expression in short- and long-term studies. Collectively the data suggest, that PU.1 expression seems to be involved in chromatin remodeling, which is further accompanied with specific gene expression changes in the lymphoid CTV-1 cell line as ATAC-sequencing revealed extensive remodeling of chromatin upon PU.1 expression, which is partially associated with the deposition of histone modification H3K27ac and the enhanced expression of neighboring genes.

Next, the impact of motif cooperativeness across the 14 ATAC-based K-means clusters was analyzed. For this purpose, a motif scan for all 14 clusters using the HOMER suite (Heinz et al. 2010) was conducted. Known motifs as well as *de novo* motifs were searched and the obtained data of all clusters was merged, creating two files containing all known or *de novo* motifs found in PU.1 peaks of the specific ATACseq K-means cluster. Those files were further reduced to the top motif hits (p-value of at least 1e$^{-12}$) including only relevant motifs. To visualize the distribution of those motifs a balloon plot was generated and the MAC-derived PU.1 motif score distribution of each cluster was analyzed in addition **(Figure 5-20)**.

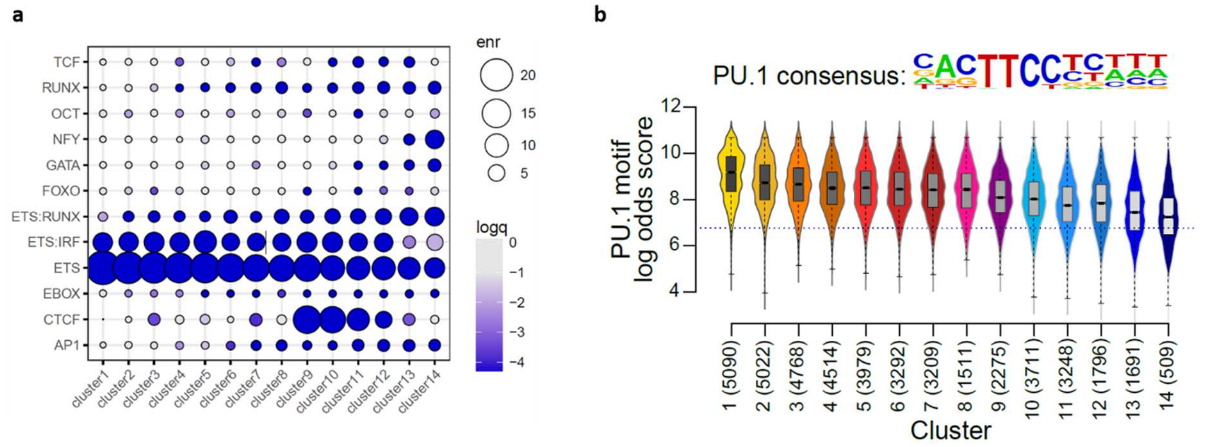


**Figure 5-20 - Motif cooperativeness & PU.1 motif log odds score distribution across the 14 K-means cluster**

**a |** The balloon plot depicts the motif enrichment of the selected motifs (y-axis) in each cluster (x-axis). The node thickness represents the enrichment and the coloring correlates with the p-value of the motif co-occurrence in PU.1 peaks of each K-means cluster. **b |** The PU.1 motif log odds score distribution of the consensus MAC PU.1 motif (depicted above the plot), generated earlier, is shown for each cluster in a combined bean- and boxplot (the color code for each cluster is depicted in the bean plot). The quantity of PU.1-specific ATACseq sites is illustrated. The median of the specific distribution across all clusters is depicted inside the bean with a conventional boxplot.

As seen in the balloon plot **(Figure 5-20a)** only a few cooperative motifs are significantly co-associated with PU.1 binding motifs across the K-means cluster. Among them mainly ETS and composite ETS:IRF sites are enriched in almost all clusters. However, in cluster 14 for example the promotor-specific TF NFY is co-enriched as well, whereas in clusters 9-11 CTCF motifs are found in association to PU.1 motifs. Moreover, the MAC-derived PU.1 motif log odds score distribution, mentioned earlier, across the clusters is illustrated in **Figure 5-20b** and the combined bean- and boxplot shows that pre-existing accessible sites (9-14) are associated with lower motif log odds scores in comparison to non-accessible sites (1) and sites with a high remodeling index (2-8). Besides, networks of motif co-associations were examined to investigate the fraction of PU.1-specific peaks overlapping with cooperative TF motifs. For this analysis, the motif search was repeated with the PU.1 motif masked to include PU.1-PU.1 co-associations (clustered PU.1 binding sites; see also section 10.1.6 ‘Analysis of Homotypic PU.1 Clusters’) in ATACseq-annotated PU.1-bound regions, before the networks were generated and plotted. The motif co-association for each cluster is represented in the following figure **(Figure 5-21a)**.

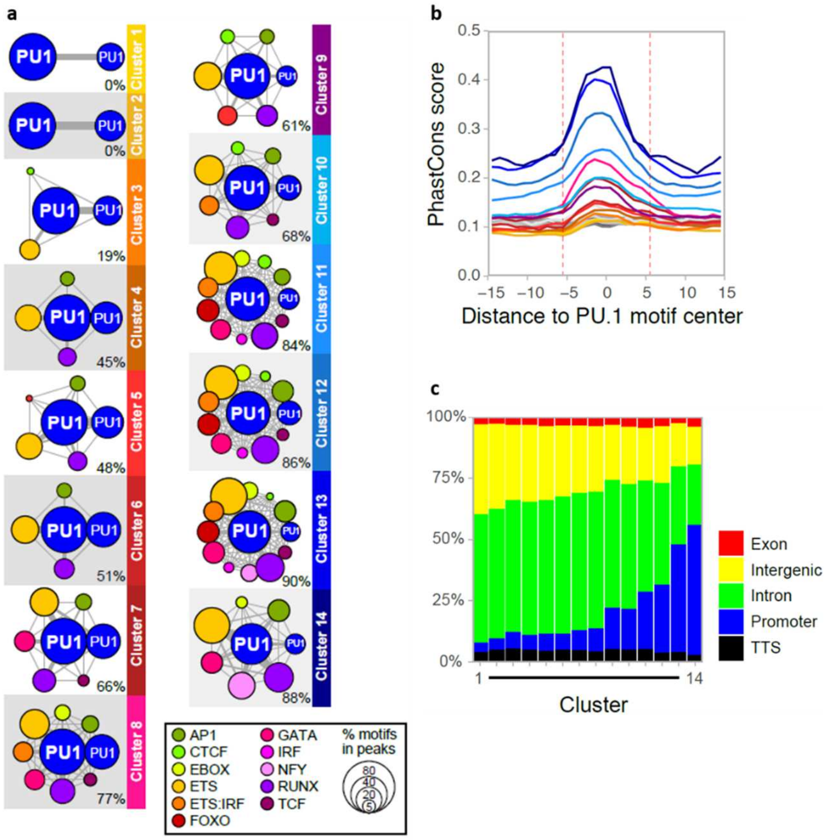


**Figure 5-21 - Motif co-enrichment, evolutional conservation & gene ontology across K-means cluster**

**a |** Motif co-enrichment networks for all ATACseq-annotated PU.1 binding clusters are displayed. The fraction of PU.1 peaks overlapping with co-associated TF motifs other than PU.1 is shown below each network. The size of each node represents the according motif enrichment in percent and the coloring indicates the co-associated TF motif. PU.1-PU.1-associations are drawn in blue. **b |** Evolutionary conservation of the PU.1 motif across the 14 ATACseq K-means cluster is illustrated in the histogram with the PhastCons score. Clusters are colored according to their individual color code (see **a**). Non-bound sites are shown in grey. **c |** Gene ontology analysis across the 14 K-means cluster is depicted in the stacked bar chart. The association with the individual regions is given in percent.

Pre-existing accessible regions show a high degree of motif co-enrichment among their PU.1 binding sites (cluster 13 for example). Among them mainly ETS and RUNX TF motifs are significantly enriched, however, promotor-specific TF (e.g. NFY) are also co-associated with PU.1-motifs of those regions. Moreover, those regions show the lowest enrichment of additional PU.1 binding sites compared to all other clusters. Non-accessible regions on the other hand (see cluster 1), are not associated with any other TF motifs than PU.1 motifs itself, which is correlated with higher motif log odds scores of the PU.1 motif as depicted in **Figure 5-20b**.  Sites with a high remodeling index in contrast (see cluster 6-8), are mainly co-enriched for additional PU.1 binding sites as well as for additional ETS family member TF motifs similar to pre-existing accessible ATACseq sites. Moreover, the evolutionary conservation of the consensus PU.1 motif was analyzed across all clusters. Therefore, the peaks of each cluster were centered on the PU.1 motif, before the ATACseq cut sites were annotated. Those peaks were then further used to map the fraction of PU.1-bound vs. non-bound regions. For this approach, a random control set of unbound regions was generated, before the ATACseq signals before and after PU.1-transfection were annotated and re-centered on the PU.1 motif. PU.1 motif-centered peaks as well as PU.1-unbound random control sets across all clusters were than used to annotate the PhastCons score distribution. The PhastCons program can be used for the identification of evolutionarily conserved elements in a multiple alignment, where conserved elements are identified based on a two-state phylogenetic hidden Markov model. The corresponding conservation of the PU.1 motif across the PU.1-bound and non-bound ATACseq clusters is illustrated in **Figure 5-21b** and the histogram plot shows that pre-existing accessible sites are associated with a higher degree of conservation of the PU.1 motif, whereas non-accessible as well as sites with a high remodeling index are associated with a lower degree of conservation. Likewise, we also studied the gene ontology of our 14 K-means cluster by annotating the PU.1-bound ATACseq peaks with the GTF (gene transfer format) file of the human reference genome (hg19), which lists information about gene structure. The resulting bar chart emphasizes that PU.1 bound ATACseq regions becoming accessible upon PU.1-transfection are mainly enriched for intronic and intergenic regions, while binding sites in pre-existing accessible regions are mainly enriched for promotor regions (see **Figure 5-21c**).

To investigate the relationship between homotypic PU.1 clusters and PU.1-induced *de novo* binding, homotypic indices across the PU.1-bound ATACseq K-means cluster were generated and associated with the aforementioned remodeling index to investigate the percentage of co-associated PU.1 binding for each ATACseq cluster. This analysis revealed that there is a significant correlation (r=0.98) between homotypic PU.1 binding sites and *de novo* PU.1-induced chromatin remodeling, since K-means cluster with a high remodeling index (red, 6; dark-red, 7; pink, 8;) correlated the best with a high percentage of homotypic binding sites (see **Figure 5-22a**).

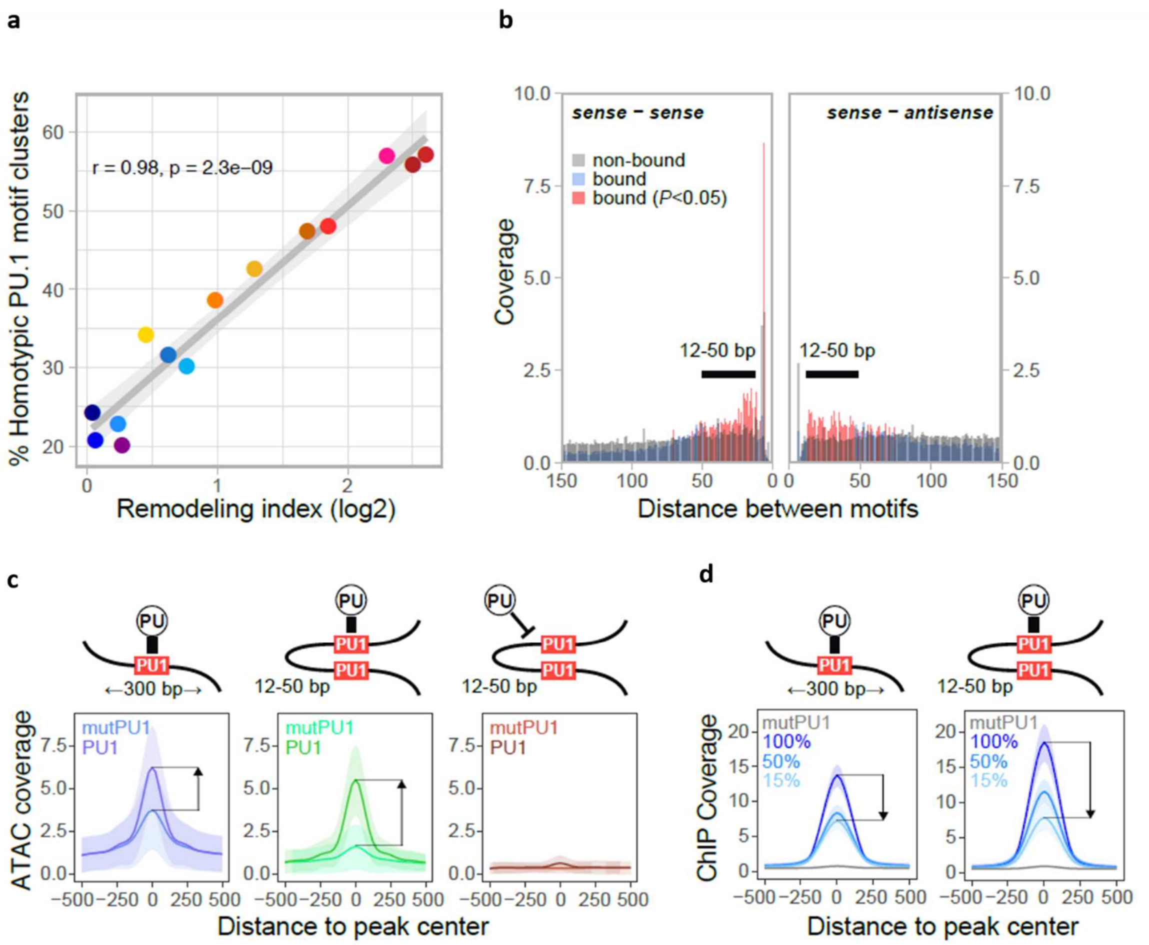


**Figure 5-22 - Relationship of homotypic clusters & PU.1 binding site selection**

**a |** Scatter plot depicting the correlation between homotypic PU.1 motif clusters (y-axis; in %) and the PU.1-induced remodeling index (x-axis) across the 14 ATACseq K-means cluster. The Pearson correlation as well as the p-value are illustrated. **b |** The PU.1 coverage is plotted against the distance between PU.1 motifs across PU.1-bound vs. non-bound ChIPseq binding sites. Significantly enriched PU.1-bound regions occurring between PU.1 motifs are depicted in orange (p-value < 0.05), whereas the overall coverage of PU.1-bound regions is depicted in blue vs. non-bound in grey. **c |** Histogram plots illustrating the ATACseq coverage of PU.1 vs. PU.1mut in bound and unbound regions containing single and clustered PU.1 motifs (from left to right: PU.1 (purple) vs. PU.1mut (blue) coverage across single binding sites; PU.1 (green) vs. PU.1mut (turquoise) coverage across clustered binding sites; PU.1 (brown) vs. PU.1mut (light-brown) coverage across unbound clustered binding sites). The 95% confidence interval is depicted. **d |** Histogram plots illustrating the PU.1 ChIPseq coverage across single (left) vs. clustered (right) binding sites for different PU.1 titration levels (percentage of transfected PU.1 (100%, 50%, 15%) is indicated with different shades of blue; PU.1mut, grey). The 95% confidence interval is depicted.

Moreover, we analyzed if there is a preferred distance between PU.1 motifs in pairs allowing for efficient binding of PU.1 (see section 10.1.5 ‘preferred distance between PU.1 motifs in pairs’). For this purpose, homotypic motifs were overlapped with the PU.1 peaks obtained in the lymphatic cell line, split into bound and unbound sites and the frequency of enriched motifs across all PU.1 peaks was calculated in sense as well as antisense orientation. As depicted in **Figure 5-22b** the PU.1 coverage was significantly enriched in regions, in which the distance between two PU.1 motifs ranged from 12-50 bp across a 300 bp window (sense-antisense). In all other regions, bound vs. non-bound sites showed a similar coverage with no significant differences detectable between those peak sets. In addition, the ATACseq coverage of CTV-1 cells transfected with PU.1 or PU.1mut mRNA, respectively, was annotated across all PU.1 binding sites with either single or clustered PU.1 motifs, bound or unbound by the TF.

This analysis demonstrated that the differential ATACseq coverage between the WT and the mutant protein is highest in regions harboring clustered PU.1 binding motifs within a 12-50 bp distance, whereas this difference is lost in unbound regions. To study, if this binding site selection is dependent on synergistic or additive binding events, the ChIPseq coverage of different PU.1 titration levels, which will be further elucidated in the next chapter, was also associated with the presence of clustered PU.1 binding sites. As seen in **Figure 5-22d,** the ChIPseq coverage obtained for lower PU.1 levels (15%) was decreased the most in regions comprised of clustered binding sites in comparison to regions with a single PU.1 motif. Therefore, this analysis favors a synergistic rather than an additive binding site selection of the hematopoietic TF in regions comprised of clustered PU.1 motifs.

Taken together, we found that *de novo* remodeled PU.1 binding sites are characterized by high PU.1 motif log odds scores and with the enrichment of co-associated clustered PU.1 motifs. Moreover, the consensus PU.1 motif of those sites is evolutionary less conserved and the binding sites are mainly enriched for intergenic regions. Pre-existing accessible sites on the other hand, comprise lower PU.1 motif log odds scores, the highest degree of motif co-enrichment, are mainly associated with promotor-regions and their PU.1 recognition sequence is more conserved.

### 5.3.3 Cooperative Binding Events of ETS-Family Transcription Factors

In addition, we also studied the binding site selection of two additional ETS-family members called ETS-1 and FLI-1 (Friend leukemia integration 1 transcription factor), respectively (detailed computational analysis is listed in section 10.1.5, 'comparison of ETS1, FLI1 and PU1 binding sites' and in section 10.1.7 'Analysis of Heterotypic PU.1 clusters'). These cell type-specific TFs belong to the first class of the ETS-family and are endogenously expressed in the lymphoid cell line as seen in **Figure 5-14** and we asked, if they occupy the same binding sites as PU.1 upon its overexpression in the conventional cell line. To analyze their different binding sites, ChIPseq raw read data was mapped to the human reference genome (hg19; see sections 10.1.1.1-10.1.1.2), CNV-normalized and data of corresponding replicates was merged (two replicates each). For visualization, bigWigs were generated and the signals were averaged across replicates. The peak finding was essentially done as already described before and besides stringent peak finding, a more lenient approach was also included to compare the overlap of ETS binding sites. For this approach, the standard FDR cut-off was set and the reads per peak were not limited, so that the peak calling was essential carried out with the standard settings of the HOMER suite (Heinz et al. 2010). An example for the distribution of the binding sites for each TF is depicted in the following IGV genome browser track **(Figure 5-23a)** and it illustrates that there exist common sites between all TFs, which are bound by ETS-1 and FLI-1 independently of the PU.1 expression level in the cell line. However, there are also sites only bound by all TFs after PU.1-induction as well as PU.1-specific binding sites, which are not bound by ETS-1 or FLI-1, respectively. To investigate the overlap of all binding sites on a global scale, the enrichment of the ETS TF binding sites across the PU.1 ATACseq K-means cluster was studied next. Therefore, histogram plots for all TF binding sites before and after PU.1 transfection, as well as the ATACseq signal of PU.1mut-transfected across all PU.1 binding sites obtained in the cell line were compiled (see **Figure 5-23b**) and the overlap of ETS-1 and FLI-1 binding sites compared to PU.1 was calculated across all clusters (see **Figure 5-23c**).

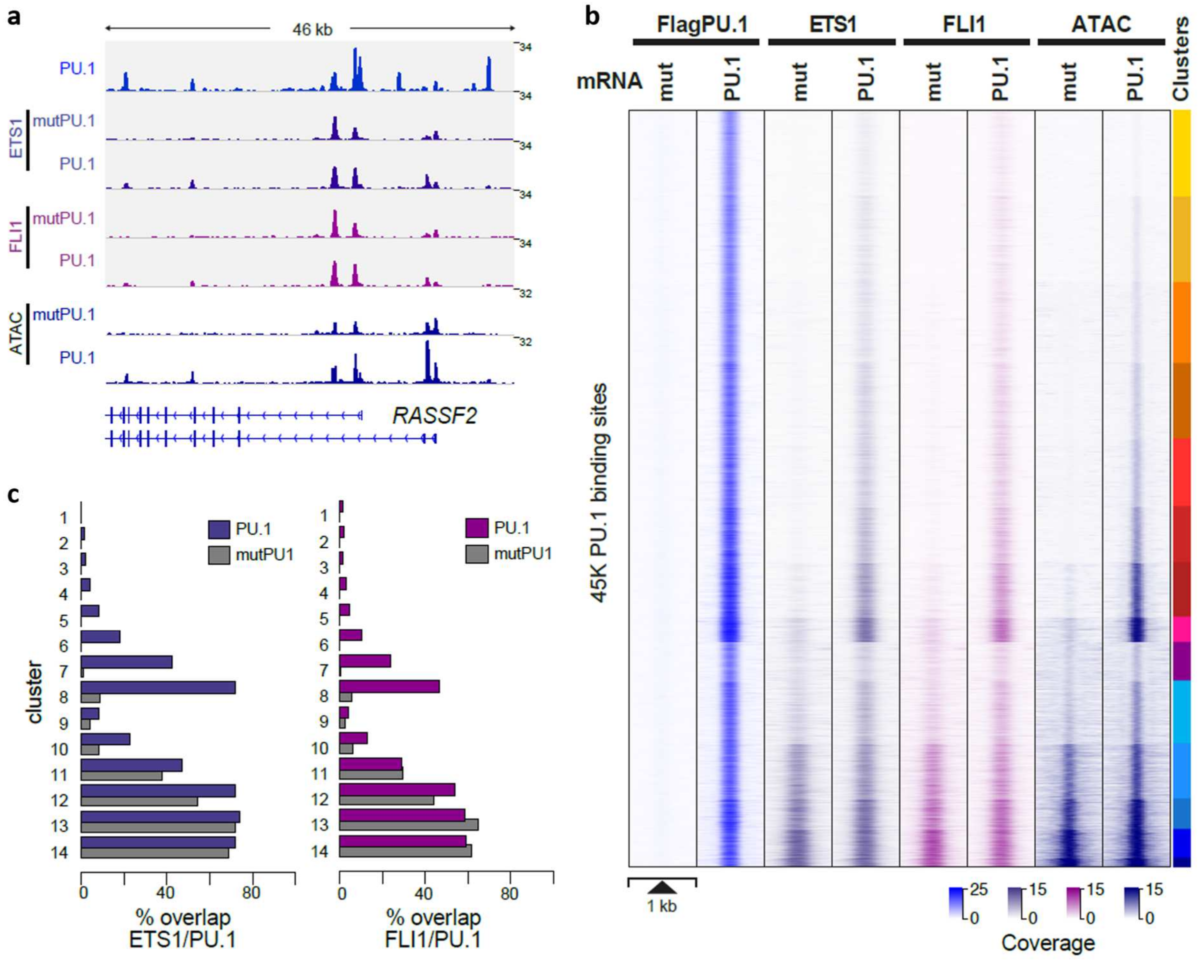


**Figure 5-23 - Comparison of PU.1-, ETS-1- and FLI-1-specific binding sites in PU.1-expressing and conventional CTV-1 cells**

**a |** IGV genome browser track of the PU.1 (blue), ETS-1 (purple) and FLI-1 (pink) ChIPseq reads of PU.1-transfected (upper track) and conventional (lower track) CTV-1 cells across the RASSF2 locus. ATACseq reads of PU.1-transfected and conventional cells are depicted in blue below the ChIPseq tracks. All read counts are CNV-normalized and averaged across replicates. The exact location of the gene is depicted (chr20:4,766,326-4,813,717). **b |** The distribution of the PU.1 ChIPseq signal (PU.1-transfected & and conventional CTV-1 cells), the ETS-1 ChIPseq signal (PU.1-transfected & and conventional CTV-1 cells), the FLI-1 ChIPseq signal and the ATACseq signal (PU.1-transfected & and conventional CTV-1 cells) plotted across the 14 K-means ATACseq cluster is illustrated. The specific clusters are shown with the color bar on the right. **c |** Bar plots displaying the overlap of ETS-1 (left; blue) and FLI-1 (right; purple) binding sites with PU.1 binding sites of PU.1-transfected and conventional CTV-1 cells (grey bars) across the 14 K-means ATACseq clusters. The particular overlap is given in percent.

As the plots demonstrate, the strongest overlap of the ETS-family TFs compared to PU.1 is seen in regions, which are already accessible in the lymphoid cell line independently of the PU.1 expression status in the cells (cluster 10-14, light blue until dark blue). Non-accessible regions (cluster 1, yellow) as well as almost all regions *de novo* remodeled by PU.1 (cluster 2-8, dark-yellow until pink) on the other hand, show only an overlap of all three factors upon PU.1-induction in the cell line, implying that those regions cannot be accessed by ETS-1 and FLI-1, when PU.1 is not present in the lymphoid cells. This is further supported in **Figure 5-23c**, which illustrates that the overlap of the ETS binding sites with PU.1 binding sites in clusters 11-14 is at least 40% no matter if PU.1 is expressed or not. Moreover, the major overlap of both ETS class 1 TFs is found in cluster 8 (pink) upon PU.1 expression. This cluster has a high remodeling index (see **Figure 5-19**) and ETS TF binding to *de novo* remodeled regions seems mostly to be established by PU.1 expression in the CTV-1 cell line. To further investigate if PU.1-bound and unbound regions show an exchange with the ETS class 1 TFs or rather a competitive binding pattern, the HOMER (Heinz et al. 2010) *de novo* derived ETS specific consensus sequence was analyzed first and overlapped with the PU.1 consensus motif to define heterotypic PU.1/ETS clusters. The preferred distance between ETS and PU.1 motifs across a 300 bp window (sense-antisense) was calculated and the highest enrichment was gained for a 12-50 bp distance between PU.1 and ETS class 1 consensus motifs. Single binding sites bound by PU.1 only or PU.1 and ETS TFs, as well as paired sites bound by PU.1 only or PU.1 and ETS TFs were obtained and annotated across the 14 ATACseq K-means cluster. The corresponding histogram is shown in **Figure 5-24a** which illustrates that single, as well as paired binding sites, bound by ETS-1 or FLI-1 together with PU.1 are mainly enriched in *de novo* remodeled (brown and pink) and accessible (blue to dark-blue) regions, whereas all other regions are mainly mutually exclusively enriched for single or homotypic PU.1 binding sites. The relationship between those differential binding sites is also depicted in the histogram plots in **Figure 5-24b**, which illustrate that the remodeling capacity of PU.1 is further enhanced by the synergistic binding of ETS-family TFs, leading to an even higher ATACseq coverage upon PU.1 expression at single and paired binding sites. Moreover, this correlates with the corresponding PU.1 ChIPseq coverage across those regions, which is also enhanced in single regions bound by ETS-1 and at heterotypic PU.1 binding sites. In addition, the PU.1 and ETS class 1 motif score distribution correlates well with the particular binding pattern of PU.1 and ETS-1/FLI-1, respectively, with lower PU.1 motif scores primarily present at shared single and paired binding sites as well as at ETS-specific binding sites, which on the other hand show higher ETS class 1 motif scores (see **Figure 5-24c**).

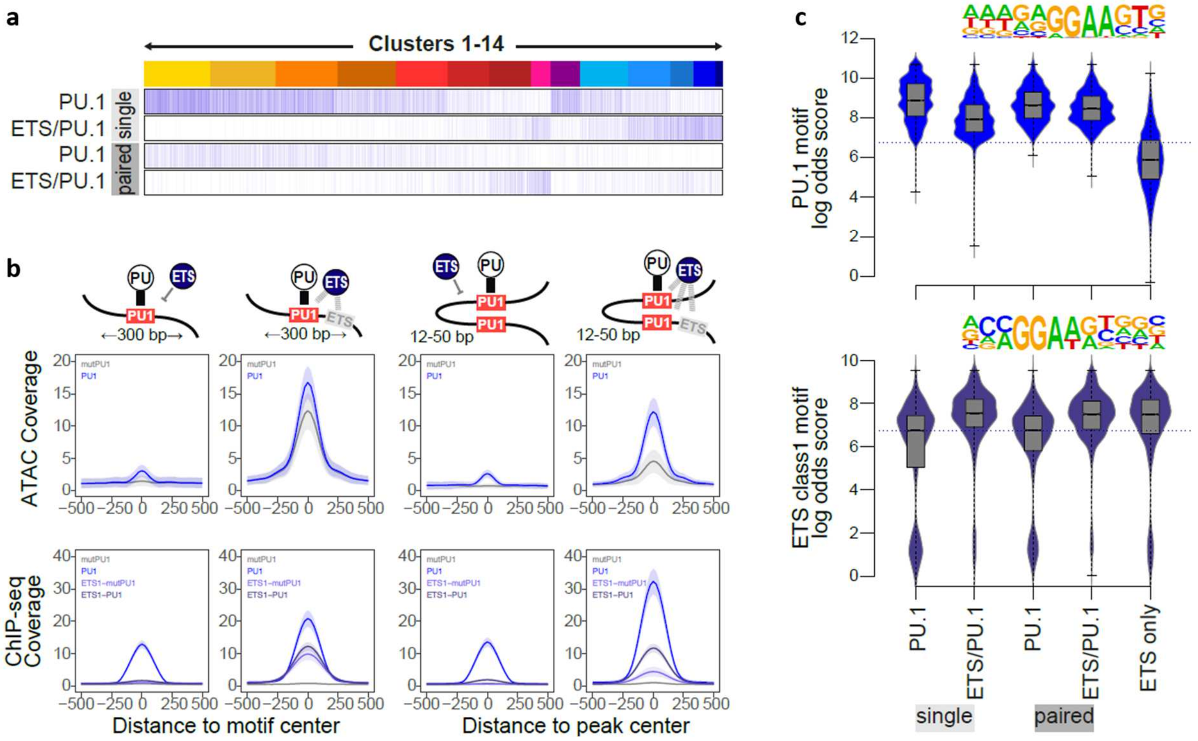


**Figure 5-24 - Relationship of ETS class 1 and PU.1 binding sites at heterotypic PU.1 clusters**

**a |** The distribution of the ChIPseq signal across the 45.000 PU.1 binding sites obtained in CTV-1 cells in correlation with the ATACseq K-means clusters is plotted. PU.1 binding site selection in regions with single or paired heterotypic clusters is depicted below the 14 K-means cluster. **b |** Histogram plots illustrating the ATACseq coverage (PU.1, blue; PU.1mut, grey) as well as the ChIPseq coverage of PU.1 (blue), mutant PU.1 (grey), ETS-1 (dark-purple) and ETS-1 in PU.1mut CTV-1 cells (light-purple) across 300 bp regions (left) and 12-50 bp regions (right) with single or paired heterotypic clusters. The 95% confidence interval is indicated for each peak set. **c |** The motif score distribution of the consensus MAC PU.1 motif (upper plot), generated earlier, and the defined ETS class 1 specific motif (lower plot) is shown for single and paired heterotypic PU.1 binding sites in a combined bean- and boxplot. The median of the specific distribution across all peak sets is depicted inside the bean with a conventional boxplot.

In summary, these data suggests that the PU.1-induced chromatin remodeling in the lymphoid cell line can be further enhanced by the synergistic binding of its ETS-family members ETS-1 and FLI-1, probably resulting in a more stable opening of the chromatin. However, motif analyses (data not shown) revealed that ETS TF-bound sites are also enriched for additional TF consensus sequences of the RUNX or GATA TF family, respectively, so that besides a synergistic binding model, cooperativeness between different TF families is likely involved in this process as well. Nevertheless, the ETS class 1 TFs are not able to bind to *de novo* remodeled regions without PU.1 being expressed in those cells, suggesting that PU.1 is the major driver in opening those sites.

### 5.3.4 Binding Site Selection of PU.1-deletion Constructs in Lymphoid CTV-1 Cells

Since chromatin remodeling seems to be one of the major components controlling the access to novel binding sites in the lymphoid cell line upon PU.1 expression, we further analyzed which N-terminal protein domains of the cell type-specific TF could be involved in opening up its binding sites. For this purpose, we deleted PU.1's known protein interaction domains and used these deletion constructs for subsequent ChIPseq analyses. The design of the deletion mutants is shown in **Figure 5-25a** above their particular expression pattern in CTV-1 cells.

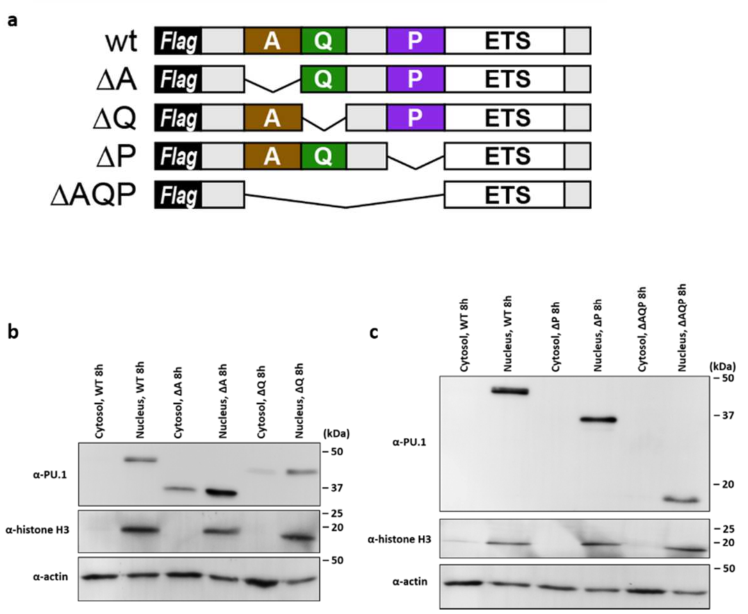


**Figure 5-25 - Design and expression pattern of PU.1-deletion mutants**

**a |** Design of PU.1-deletion mutants. The acidic (delA; aa 10-73), glutamine-rich (delQ; aa 74-100), the PEST domain (delP; aa 118-160) or all three protein-interaction domains (delAQP; aa 10-160) were deleted. The DNA-binding domain (ETS domain) was kept in all constructs and a 3x-FLAG-tag was added C-terminal for detection. **b |** CTV-1 cells were transfected with WT PU.1 mRNA (approx. 45 kDa) and corresponding amounts of PU.1-delA (ΔA; approx. 26 kDa) or PU.1-delQ (ΔQ; approx. 31 kDa) mRNA. Nuclear extracts were prepared 8 h after transfection and analyzed using WB. Proteins were detected with an anti-PU.1, anti-histone H3 and an anti-actin antibody. 15 µg of the corresponding cytosolic and nuclear fractions were loaded per lane. **c |** CTV-1 cells were transfected with WT PU.1 mRNA and corresponding amounts of PU.1-delP (ΔP; approx. 30 kDa) or PU.1-delAQP (ΔAQP; approx. 17 kDa) mRNA. Nuclear extracts were prepared 8 h after transfection and analyzed using WB. Proteins were detected with an anti-PU.1, anti-histone H3 and an anti-actin antibody. 15 µg of the corresponding cytosolic and nuclear fractions were loaded per lane.

PU.1's interaction domains were deleted in a stepwise manner as illustrated in the above figure. Four distinct deletion constructs were designed (PU.1-delA, without acidic transactivation domain; PU.1-delQ, without glutamine-rich domain; PU.1-delP, without PEST-domain; PU.1-delAQP, without all three domains) and ordered as gBlock® gene fragments (IDT, Coralville, USA; see section 3.8) for subsequent Gibson assembly (see section 4.2.2.4) and *in vitro* transcription (see section 4.2.5.2) as already described in section 5.2.2. Successful protein expression as well as nuclear localization of the mutant proteins was assessed first using WB analysis (see 4.3.5). For this purpose, CTV-1 cells were transfected with either 5’-capped, poly-adenylated PU.1 mRNA (1 µg/1x$10^6$ cells), or the different PU.1-deletion constructs (amount corresponding to their size in relation to the WT PU.1 protein). Nuclear extracts, as well as cytosolic fractions were harvested 8 h after transfection (see 4.3.2) and further analyzed using WB analysis. For protein expression analysis an anti-PU.1 antibody was used (epitope within the ETS-domain), whereas the nuclear localization of the proteins was verified using an anti-histone H3 antibody. The anti-actin antibody used in addition served as a general loading control (for a list of used antibodies see section 3.5). As seen in the WB images in **Figure 5-25b**, all PU.1-deletion mutants are almost equally expressed in the lymphoid cell line and successfully located to the nucleus, since their expression correlates with the nuclear expression of histone H3 (approx. 18 kDa; middle blot each). To analyze the particular impact of each PU.1 protein domain on its binding site preferences, ChIPseq (see section 4.2.7-4.2.8) was performed eight hours after transient mRNA transfection (see 4.1.3.1). Detailed computational analysis is listed in section 10.1.8. Obtained ChIPseq raw read data was mapped to the human reference genome (hg19; see section 10.1.1), CNV-normalized and the data sets of corresponding replicates were merged (two replicates each). For visualization, bigWigs were generated and the signals were averaged across replicates. Peak finding was essentially done as already described before, to filter for peaks with at least 15 reads with a FDR below 0.05 (see section 5.2.2). An example for the distribution of the binding sites for each deletion mutant is depicted in the following IGV genome browser track **(Figure 5-26a)**.

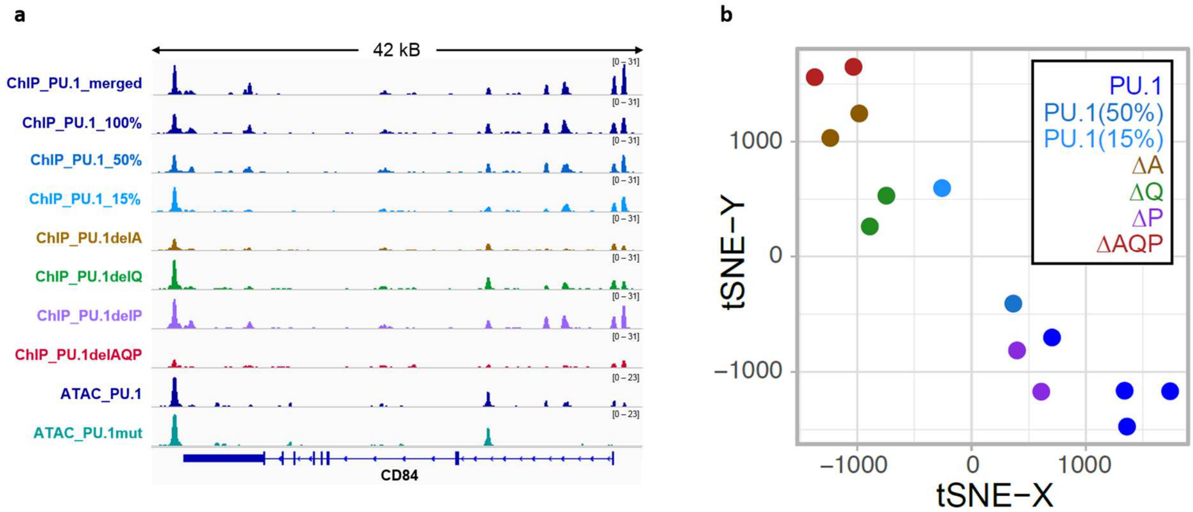


**Figure 5-26 - Distribution of PU.1 peaks from deletion constructs**

**a |** ChIPseq PU.1 reads of CTV-1 cells transfected with different amounts of PU.1 mRNA (PU.1, blue; PU.1_50%, medium blue; PU.1_15%, light blue) as well as PU.1 reads of deletion mutants (delA, brown; delQ, green; delP, purple; delAQP, red) are shown across the CD84 locus. ATACseq signal of CTV-1 cells transfected with PU.1 (blue) and PU.1mut (turquoise) mRNA is depicted in the last two rows. All ATACseq and ChIPseq signals are CNV-normalized and averaged across replicates (chr1:160,508,522-160,551,288). **b |** 2-dimensional visualization of the distribution of annotated PU.1 peaks across analyzed samples using t-SNE embedding. Correlated samples are illustrated with identical colors (PU.1, blue; PU.1_50%, medium blue; PU.1_15%, light blue; delA, brown; delQ, green; delP, purple; delAQP, red).

As already seen in the IGV genome browser track, the individual deletion mutants show a diverse binding pattern to PU.1-specific binding sites. The mutant lacking only the PEST-domain (delP; purple) comprises a similar binding pattern as the WT PU.1 protein, whereas the binding is reduced in the absence of the glutamine-rich domain (delQ; green) and even more diminished in the absence of the acidic domain (delA; brown). Moreover, the mutant lacking all three N-terminal PU.1 domains (delAQP; red) almost completely loses its ability to bind to the nuclear DNA. In addition, a titration of PU.1 levels (PU.1 100%, blue; PU.1 50%, medium blue; PU.1 15%, light blue; one replicate each) is also shown in the IGV track in **Figure 5-26a** below the merged PU.1 peak set (three replicates; see also section 5.3.2). This set was used to compare the impact of the deletion mutants with respect to lower PU.1 expression in the cell line (see also **Figure 5-5**) and it implicates, that reduced PU.1 levels are comparable to the binding capacity of the delQ- and delA-mutant protein, respectively. To further analyze the binding preferences of the mutant PU.1 proteins a merged peak set containing all PU.1 binding sites was generated and normalized using the r-log transformation method already described in section 5.1.1. The distribution of the annotated peaks for each sample is shown in the t-SNE plot of **Figure 5-26b** and it further confirms the diverse binding site distribution already seen in the genome browser track. In detail, PU.1 replicates (PU1.1, PU1.2, PU.1.3 & PU.1.4; blue) cluster together in close proximity to each other and correlate with binding sites of less expressed PU.1 (PU.1 50%, medium blue; PU.1 15%, light blue). Moreover, the binding pattern of the delP-mutant (purple) seems to be highly correlated with the WT protein, since it is also located in close proximity to WT PU.1. Binding sites of the remaining mutant proteins however, cluster together further away from WT PU.1, with the delAQP-protein (red) being the one with the highest distance. Besides, differential peak sets between the different samples were generated to further study the impact of each particular domain. Histogram plots of the PU.1 ChIPseq coverage in selected differential regions were compiled and are illustrated in **Figure 5-27** (for remaining histograms see **supplementary Figure 10-5**). Interestingly, we did not find any significant differences between the WT PU.1 and the delP-mutant protein, but 52%, 72% and 77% of peaks were upregulated in the WT PU.1 set in comparison to the delQ-, delA- and delAQP-construct, respectively (see 10.1.8, 'detection of differential peaks' and **Figure 5-27**). Moreover, there was no difference between 100% and 50% PU.1, but 29% of PU.1 peaks were upregulated in the WT compared to even less PU.1 (15%). However, these data needs to be handled with care, since only one replicate for each titration level was generated so far. In addition, we found a small fraction of peaks (about 3% each) enriched in the delQ- and delA-variant when compared to the WT, but the major fraction of differential PU.1 binding sites in both sets was downregulated in comparison to the WT PU.1 protein. The delAQP-construct on the other hand didn't comprise any detectable enriched peaks when compared to peaks of the full-length PU.1 protein. Strikingly, no differential peaks were detectable between the PU.1-

delQ-construct and the peak set of less PU.1 (15%), however 21.5% of peaks were significantly enriched in the glutamine-rich domain lacking protein (delQ) in comparison to the delA-construct.

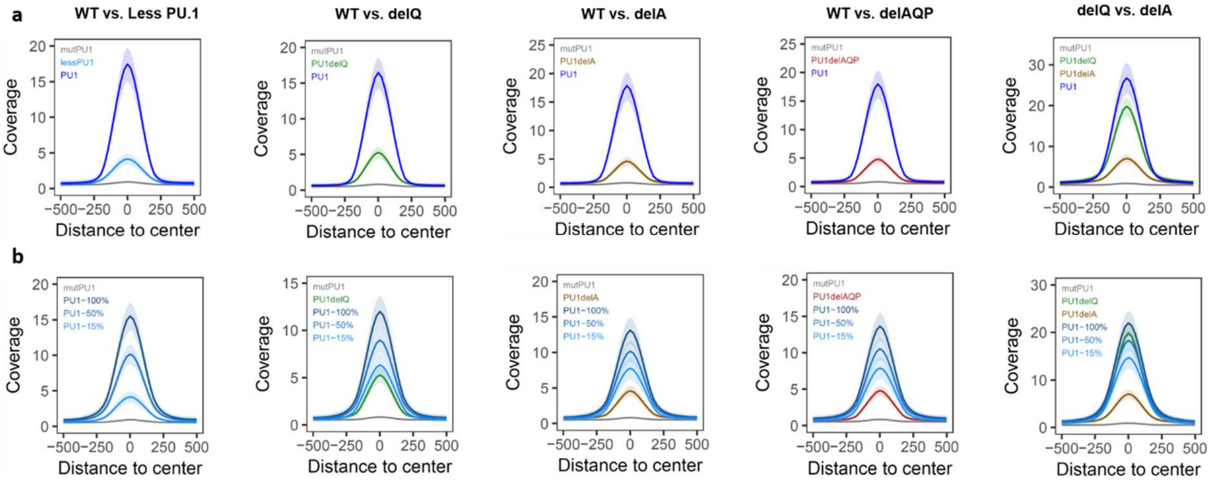


**Figure 5-27 - ChIPseq coverage across differential peaks of PU.1-mutants in CTV-1 cells**

**a |** Histogram plots showing the ChIPseq coverage (y-axis) of PU.1 reads of PU.1 (blue)- and PU.1mut (grey)-transfected CTV-1 cells, as well as of cells transfected with less PU.1 (15%, light blue), PU.1-delQ (green), PU.1-delA (brown) and PU.1-delAQP (red) across differential peak sets (WT vs. less PU.1; WT vs. delQ; WT vs. delA; WT vs. delAQP; delQ vs. delA). The distance to the PU.1 peak center is indicated on the x-axis and the 95% confidence interval is shown. **b |** Same as in **a |** just that PU.1 read counts of cells transfected with reduced PU.1 levels are depicted in comparison (100%, blue; 50%, medium blue; 15%, light blue).

Furthermore, differential PU.1 peaks of preferred sets (WT vs. less PU.1; WT vs. delQ; WT vs. delA; WT vs. delAQP; delQ vs. delA) were joined with the obtained ATACseq K-means cluster of PU.1- and PU.1mut-transfected CTV-1 cells as well as with the corresponding H3K27ac data to analyze the position of these binding sites across the different chromatin states defined earlier (see section 5.3.2). **Figure 5-28a** illustrates that efficient binding of PU.1 to *de novo* remodeled sites seems to be concentration dependent, since titration of PU.1 (less, 15%) reduced its binding ability to clusters with a high remodeling index (see cluster 2-8, dark-yellow until pink) in comparison to the WT PU.1 protein (100%). Differential peaks between the WT and the delAQP- as well as the delQ-mutant were distributed in a very similar fashion with *de novo* remodeled binding sites being the ones significantly enriched in WT PU.1 peaks compared to the mutant proteins. However, the PU.1 mutant lacking all N-terminal domains showed a reduced coverage across all obtained WT binding sites. In the absence of the acidic transactivation domain (delA) binding to *de novo* remodeled sites was even more diminished and differential PU.1 peaks between the delQ- and delA-mutant were also mainly found in regions with high remodeling indices, suggesting that the PU.1-delA-mutant is even more dependent on accessible chromatin states in comparison to the delQ-variant. Moreover, single PU.1 motifs were mainly located in pre-existing open chromatin regions (cluster 9-14; purple until blue) whereas clustered motifs were primarily enriched in *de novo* bound regions (cluster 5-8; brown until pink).

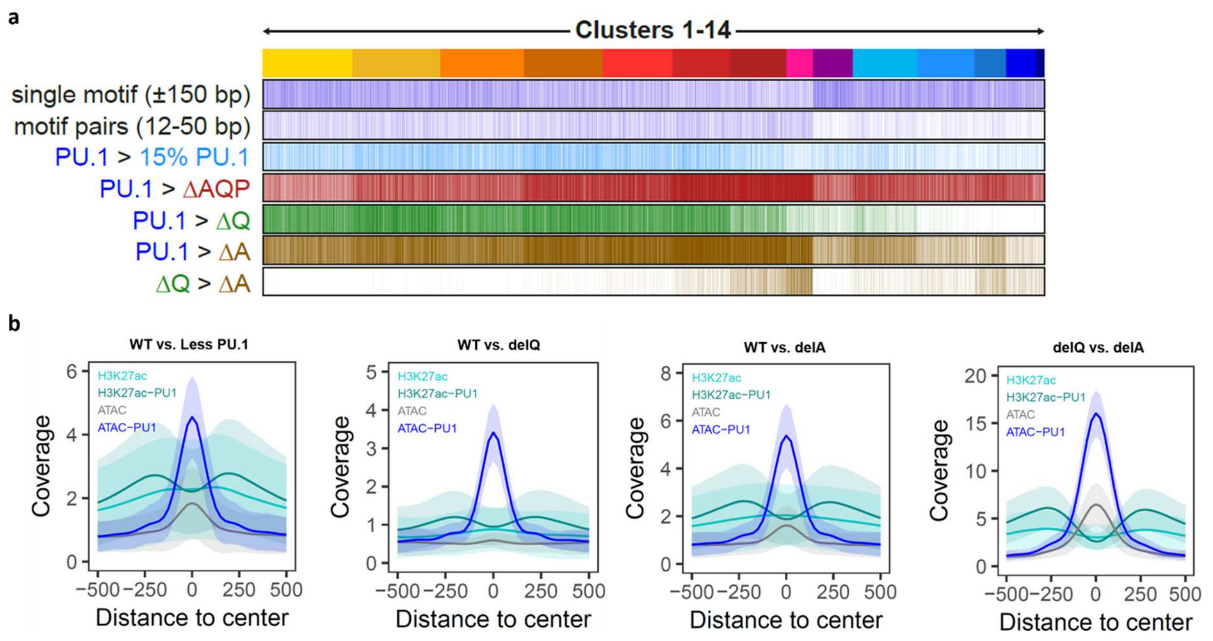


**Figure 5-28 - K-means clustering of ATAC & ChIPseq data of differential peak sets from PU.1-deletion mutants**

**a |** The distribution of the ChIPseq signal across the 45.000 PU.1 binding sites obtained in CTV-1 cells in correlation with the ATACseq K-means clusters is plotted. PU.1 binding site selection in regions with single vs. clustered motifs is depicted below the K-means cluster. Moreover, differential ChIPseq peaks of WT vs. less PU.1 (15%, light blue), WT vs. delAQP (red), WT vs. delQ (green) and WT vs. delA (brown) as well as differential peaks between delQ vs. delA (brown) are shown across the 14 K-means cluster generated out of the ATACseq signal of transfected CTV-1 cells. The corresponding color bar is located on top. **b |** Histogram plots showing the coverage of annotated H3K27ac and ATACseq data of PU.1 (ATACseq, blue; H3K27ac, green)- vs. PU.1mut (ATACseq, grey; H3K27ac, turquoise)-transfected CTV-1 cells across depicted differential bound regions. The distance to the PU.1 peak center is indicated on the x-axis and the 95% confidence interval is illustrated.

In addition, histogram plots supporting above findings are also depicted in **Figure 5-28b**. Differential peaks enriched in CTV-1 cells transfected with the WT PU.1 protein in comparison to less PU.1, or the deletion constructs (delQ, delA) show a similar ATACseq coverage, with PU.1-induced open chromatin (as measured by ATACseq-PU.1, blue) mainly overlapping with these differentially bound regions. This trend is also supported by the enrichment of the active histone mark H3K27ac of PU.1-transfected cells, however the coverage is relatively low across all analyzed peak sets. Likewise, PU.1 peaks differentially bound between the glutamine-rich (delQ) and acidic transactivation (delA) domain lacking protein mainly reside in regions overlapping with a high ATACseq-PU.1 coverage (blue). Nonetheless some differential peaks also correlate with the ATACseq signal of PU.1mut-transfected conventional CTV-1 cells, thus with pre-existing open chromatin states. Moreover, we analyzed the occurrence of clustered PU.1 binding sites in comparison to binding sites with a single PU.1 motif, respectively, across the differential bound regions in more detail. Therefore, WT PU.1 bound peaks, which overlapped with or didn't overlap with clustered PU.1 motifs, were annotated (see section 10.1.6 'Analysis of Homotypic PU.1 clusters') before histogram plots across differential bound regions were generated **(Figure 5-29a)**.

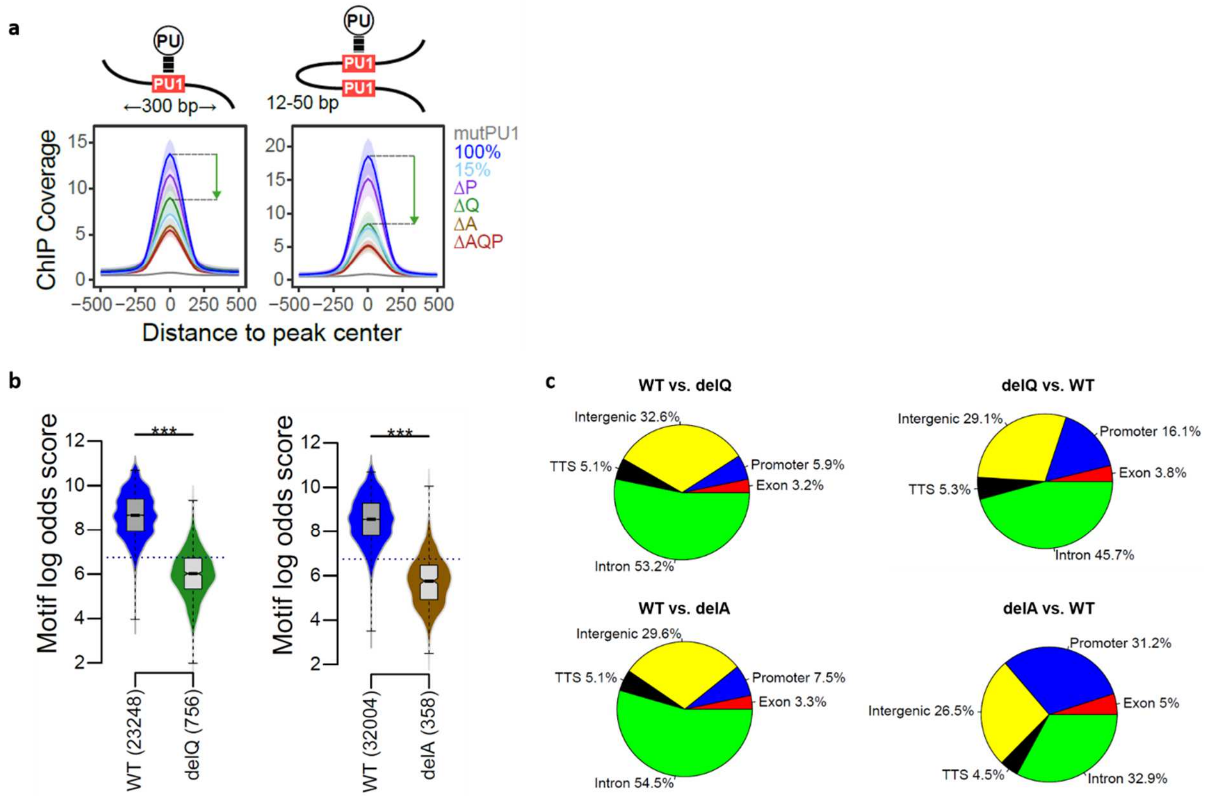


**Figure 5-29 - Motif score distribution of enriched peaks and gene ontology terms of selected PU.1-deletion mutants**

**a |** Histogram plots illustrating the ChIPseq coverage of reads for PU.1 (100%, blue; 15%, light-blue), PU.1mut (grey), PU.1-delP (purple), PU.1-delQ (green), PU.1-delA (brown) and PU.1-delAQP (red) in bound regions with single or clustered PU.1 motifs. The distance to the PU.1 peak center is indicated on the x-axis and the 95% confidence interval is depicted. **b |** MAC-derived PU.1 motif log odds score distribution of the delQ(green)- and the delA(brown)-construct in comparison to WT PU.1. The fraction of enriched peaks for each differential peak set is shown. Statistical significance was calculated with a paired Wilcox test ('*' p-value < 0.05, '**' p-value < 0.01, '***' p-value < 0.001).
**c |** Gene ontology analysis across differential bound regions as indicated.

As seen in **Figure 5-29a** the ChIPseq coverage of the mutant proteins to peaks harboring a single PU.1 binding site is decreased for all deletion constructs compared to WT PU.1 with the delP-variant being enriched the most and the delAQP-variant the less. The same trend is seen for clustered PU.1 binding sites, however the ChIPseq coverage of the PEST-domain lacking protein almost reaches the WT protein level here and both are even more enriched in homotypic regions compared to regions harboring single motifs. The remaining deletion mutants on the other hand, showed a decreased coverage across clustered binding sites when compared to single binding sites. Notably, the difference between the ChIPseq coverage of the WT PU.1 protein and the glutamine-rich domain-lacking mutant PU.1 was predominantly decreased in regions containing clustered PU.1 binding motifs as indicated by the green arrows. In addition, we studied the PU.1 motif score distribution as well as the gene ontology of corresponding binding sites for the glutamine-rich and acidic domain lacking PU.1 protein, since they showed the strongest influence on PU.1 binding site selection. As nicely illustrated in the combined bean- and boxplots in **Figure 5-29b** both mutants show a significant decrease of the PU.1 motif score in their particular enriched binding sites in comparison to WT specific regions. In addition, their enriched binding sites (delQ vs. WT; delA vs. WT) are mainly located in promoter regions with the correlated loss in intergenic and intronic regions respectively. Together these data show that the binding of PU.1 to *de novo* remodeled sites is impaired by reduced PU.1 levels, as well as by the absence of the glutamine-rich and the acidic transactivation domain of PU.1 respectively.

## 5.4 Protein-Protein Interaction Analysis

Since we found severe changes in the local organization of the chromatin of lymphatic CTV-1 cells upon PU.1-induction, we further asked which features could be involved in PU.1's binding site selection beyond protein-DNA interactions. Therefore, we aimed to map protein-protein interactions of the nuclear protein and selected deletion mutants using the so-called proximity-dependent biotinylation approach (BioID) in diverse PU.1-expressing and non-expressing cell lines. This method is based on the fusion of the protein of interest to a promiscuous biotin ligase (BirA) of *E.coli*, which - after introduction into the cells and the subsequent addition of biotin - leads to the biotinylation of proximal target proteins (Roux et al. 2013) that can be analyzed by mass spectrometry analyses. For this purpose, gBlock® gene fragments for WT PU.1 fused over a flexible linker to the *E.coli* BirA were designed. The PU.1-delQ- and delA-mutant were also fused to the ligase and a construct containing only the ligase fused to a nuclear localization sequence was designed as well and served as negative control for all experiments. As the PU.1-delP-mutant was not enriched for differential binding sites in comparison to the WT protein (see **chapter 5.3.4**) and the del-AQP-mutant lacks all N-terminal protein domains possibly involved in establishing PU.1 protein-protein interactions, these mutants were not used for subsequent proteomic analyses. All gBlock® gene fragments are listed in section 3.8 and were used for Gibson assembly (see 4.2.2.4) and consecutive *in vitro* transcription (see 4.2.5.2) as already described before (see section 5.2.2). Selected 5'-capped, poly-adenylated mRNAs were transiently introduced into the myeloid leukemia cell line THP-1 using electroporation (see 4.1.3.1) first. To check the expression levels of the fusion proteins whole cell lysates were harvested (see 4.3.1) after indicated time-points and analyzed using WB analysis (see section 4.3.5 and **Figure 5-30**).

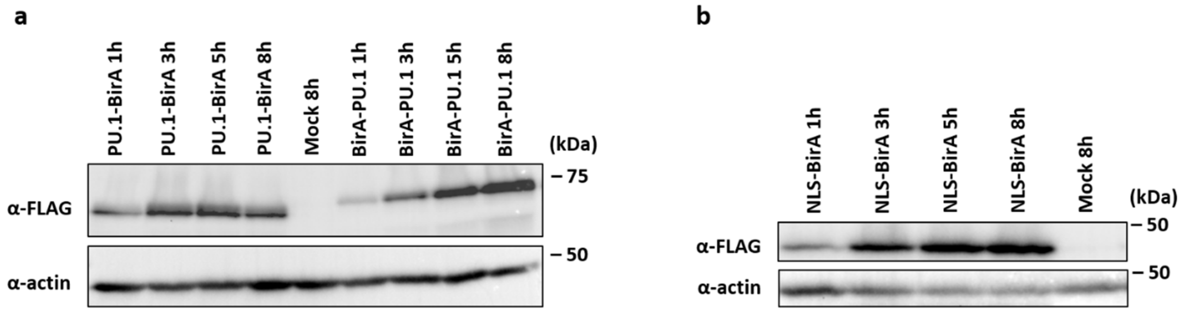


**Figure 5-30 - Expression of PU.1-fusion proteins and controls in THP-1 cells**

**a |** Time course (1h, 3h, 5h, 8h) of PU.1-BirA (left) and BirA-PU.1 (right) expression pattern in myeloid THP-1 cells. A mock control is also depicted after 8 h. Blots were stained with an anti-FLAG and anti-actin antibody respectively. **b |** Time course (1h, 3h, 5h, 8h) of NLS-BirA expression pattern in myeloid THP-1 cells. A mock control is also depicted after 8 h. Blots were stained with an anti-FLAG and anti-actin antibody respectively.

The WB images depict, that all selected fusion proteins are expressed in the myeloid leukemia cell line. Moreover, the protein expression of all constructs was most abundant between three and five hours after transfection. According to this, we decided to stimulate the biotinylation of vicinal proteins three hours after transfection and to harvest the corresponding lysates eight hours after transfection (see section 4.3.7 for detailed methodology). Since we aimed to analyze cell type-specific differences in the interactome of PU.1, erythroid K-562 cells as well as the non-expressing, already extensively discussed cell line CTV-1 were used in addition to myeloid THP-1 cells. ChIPseq analysis was also conducted 5 hours after transfection (see section 4.2.6-4.2.7) to verify that the fusion proteins did not interfere with regular PU.1 binding patterns. Computational analysis for depicted samples is illustrated in section 10.1.9 ('Analysis of the binding properties of PU.1-fusion Proteins') and an IGV genome browser section of transfected CTV-1 cells is shown below.

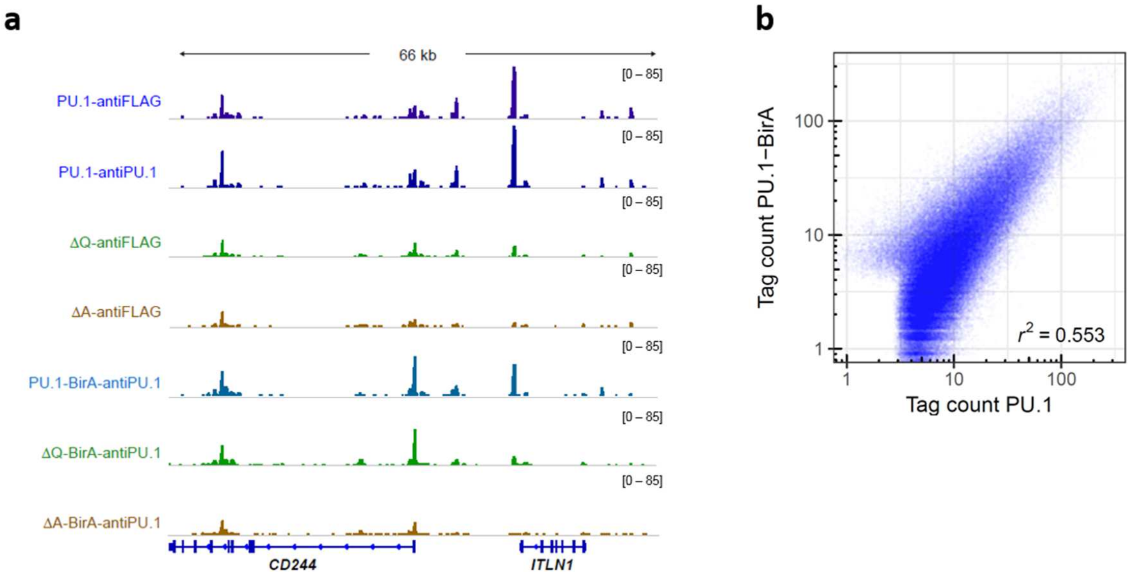


**Figure 5-31 - IGV genome browser track and comparison of ChIPseq reads of PU.1-fusion proteins in CTV-1 cells**

**a |** ChIPseq signals of PU.1 (obtained with an anti-FLAG or anti-PU.1 antibody respectively; blue), PU.1-BirA (light blue), PU.1-delQ/PU.1-delQ-BirA (green) and PU.1-delA/PU.1-delA-BirA (brown) across depicted genomic loci are shown. All reads are CNV-normalized and averaged across replicates. The exact localization of the PU.1 binding sites is depicted for each locus (chr1:160,800,594-160,867,530).
**b |** Read counts obtained in PU.1-BirA-transfected CTV-1 cells (y-axis) vs. cells transfected with WT PU.1 (x-axis) obtained with an anti-PU.1 antibody are depicted in the scatter plot. The Pearson correlation (coefficient of determination) is depicted ($r^2$=0.553).

The C-terminal tagged PU.1-fusion protein showed an equal binding pattern compared to the WT PU.1 protein, however its enrichment seemed to be slightly reduced as also illustrated in the scatter plot ($r^2$=0.553). The same was seen for the mutant fusion proteins, though their ChIPseq enrichment was even lower than the one of the WT-fusion protein **(Figure 5-31)**. As all proteins were expressed in the hematopoietic THP-1 cell line and their fusion to the biotin ligase didn't impair their major function, which is binding to nuclear DNA, we went on with the BioID approach. Therefore, THP-1 cells were used first and transfected with PU.1-BirA, as well as with the NLS-BirA control. Biotinylated proteins were bound on Streptavidin-coated magnetic beads and subsequent trypsin digestion and mass spectrometry analysis was carried out in cooperation with Prof. Dr. Axel Imhof and Dr. Andreas Schmidt at the Zentrallabor für Proteinanalytik of the University of Munich (ZfP, LMU). All experiments were performed in triplicates allowing for robust Student's t-test statistics. In order to interpret mass spectrometry raw data, data was mapped against a combined forward/reverse human protein database (Uniprot, vs. Feb. 2015) employing the Andromeda algorithm within the MaxQuant software suite (vs. 1.6.0.1) at the ZfP. Proteins were quantified, if at least two razor or unique peptides were identified and the corresponding iBAQ (intensity-based absolute quantiation) values were reported. In order to facilitate sample comparison proteins with less than two valid values per sample in the protein hits category of the reversed database were removed. For sample comparison, iBAQ values were log2-transformed and subsequently median normalized and only proteins with a FDR cut-off of 0.05 were used for subsequent analysis.

Obtained results for the myeloid THP-1 cell line are summarized in the following figure. In total, we found 147 proteins, which were significantly enriched in the PU.1-IP compared to the negative control (NLS-BirA; **Figure 5-32**).

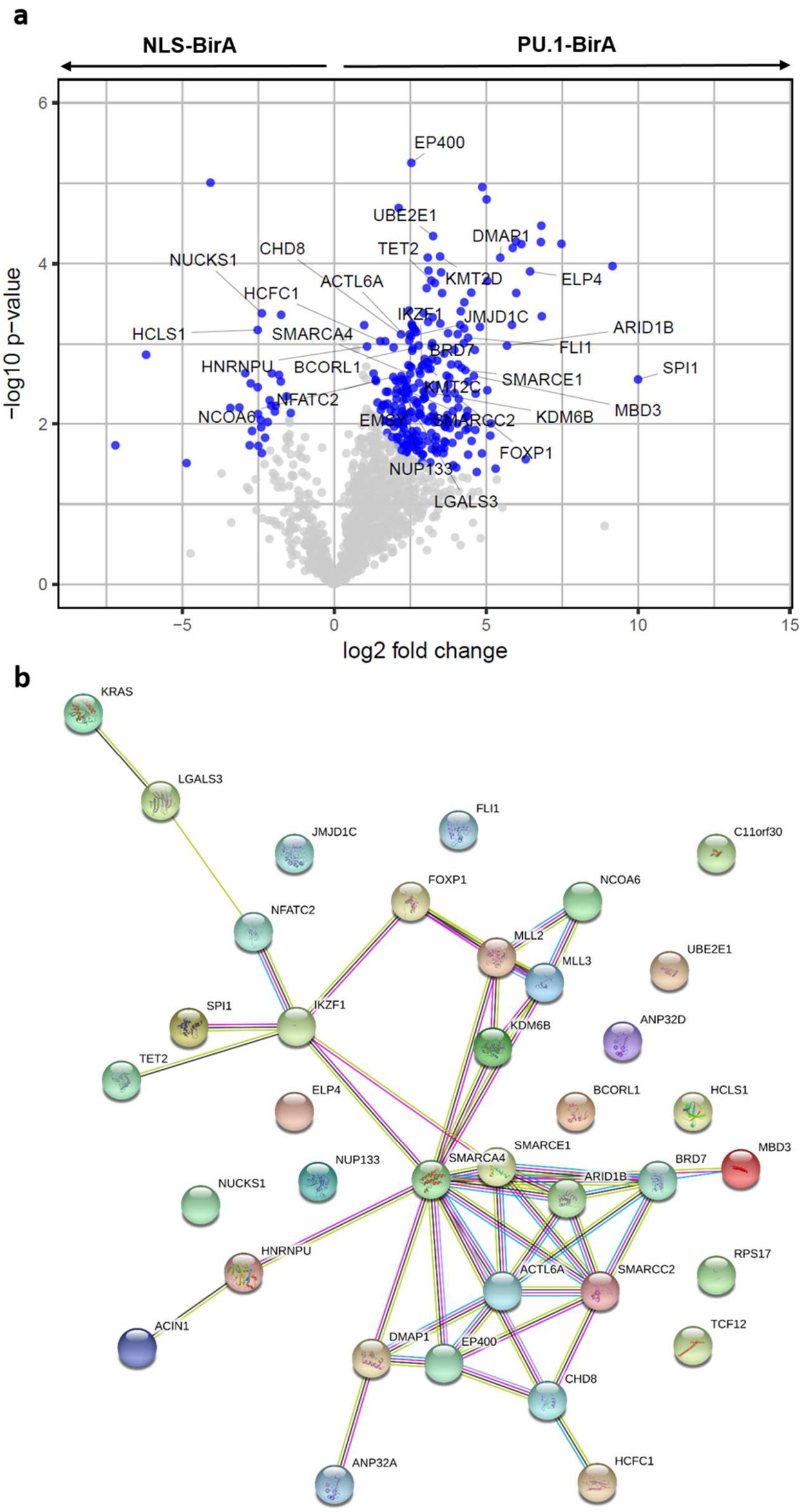


**Figure 5-32 - PU.1 protein-protein interactions in myeloid THP-1 cells**

**a |** Volcano plot illustrating proteins significantly enriched in PU.1-BirA-transfected THP-1 cells, when compared to NLS-BirA-transfected control cells. The significance (-log10 p-value) is plotted against the difference (log2 fold-change) of the normalized intensities. The more significant the difference, the smaller the p-value and thus the higher the -log10 of the p-value. Therefore, points for features with highly significant differences lie high in the plot. Features of interest are those, which change significantly and by a certain magnitude. Blue dots represent proteins with a FDR < 0.05 and a logFC > 2. Only PU.1-specific proteins of GO-terms for chromatin organization and myeloid cell differentiation are highlighted. **b |** STRING analysis illustrating the functional protein association network. The network view summarizes predicted associations for selected proteins significantly enriched in the PU.1-IP. The network nodes are the proteins, the edges represent the predicted functional associations. Seven types of evidence are used in predicting these associations (i.e.co-occurrence, experimental & database evidence).

Out of all significantly enriched PU.1-specific proteins, we further filtered the obtained gene list using Metascape for better visualization of interesting terms. Only genes associated with chromatin remodeling and myeloid differentiation are highlighted in **Figure 5-32** (detailed computational analysis is listed in section 10.1.9 'Analysis of obtained mass spectrometry data'). As seen in the volcano plot among these proteins mainly components of the SWI/SNF remodeler complex like ARID1B, SMARCA4 (BRG1) and SMARCE1 among others, are associated with PU.1 in the myeloid cell line. Moreover, cell type-specific TFs like FLI-1 and FOXP1 (Forkhead box protein P1) are enriched in the PU.1-IP, as well as its known interactor TET2. In addition, enriched GO-terms strengthen these results as primarily chromatin remodeling, as well as regulation of transcription and myeloid cell differentiation terms are found for the PU.1-specific set (see **supplementary Figure 10-6**). A functional protein association network summarizes these findings (see **Figure 5-32b**) and since we found that PU.1 is associated with the SWI/SNF complex in the myeloid cell line, we wondered if this would also hold true in other hematopoietic lineages. Therefore, we performed the same experiments in the erythroid K-562 cell line. However, the efficiency of the IP in these cells was much lower when compared to THP-1 cells. Nevertheless, a significant association of PU.1 with proteins of the SWI/SNF remodeler complex was also seen in those cells (see **Figure 5-33**). Moreover, the cell type-specific TF TAL-1 (T-cell acute lymphocytic leukemia protein 1) was significantly enriched. This factor also plays an important role in hematopoietic differentiation and serves as positive regulator of erythroid differentiation, thus obtained results might be reliable even with the small amount of enriched genes.

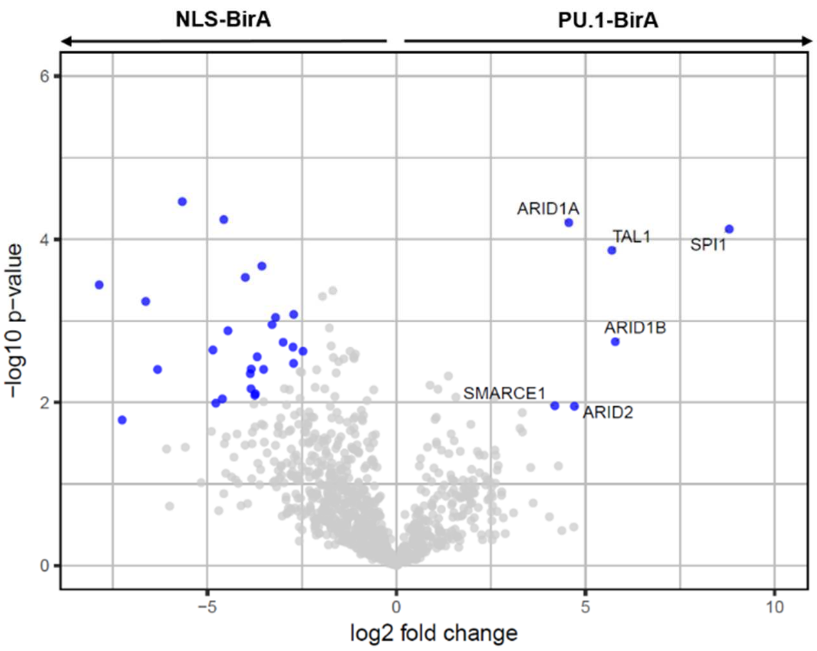


**Figure 5-33 - PU.1 protein-protein interactions in erythroid K-562 cells**

Volcano plot illustrating proteins significantly enriched in PU.1-BirA-transfected K-562 cells, when compared to NLS-BirA-transfected control cells. The significance (-log10 p-value) is plotted against the difference (log2 fold-change) of the normalized intensities. The more significant the difference, the smaller the p-value and thus the higher the -log10 of the p-value. Therefore, points for features with highly significant differences lie high in the plot. Features of interest are those, which change significantly and by a certain magnitude. Blue dots represent proteins with a FDR < 0.05 and a logFC > 2. Only proteins enriched in the PU.1-IP are highlighted.

To further verify these data, the lymphoid CTV-1 cell line already analyzed in detail above (see section 5.3) was used in addition. Since this cell line does not endogenously express PU.1, we also introduced the two mutant PU.1-proteins, which showed the strongest effect on binding site selection (see section 5.3.4) as BirA-fusion proteins into those cells (PU.1-delA-BirA, PU.1-delQ-BirA). This allowed us to study, if PU.1 might physically interact with components of the remodeler complex as well as with other proteins in terms of its specific protein-interacting domains. For this approach the BioID protocol was slightly modified. In brief, the amount of transfected cells and corresponding mRNA was doubled and a dialysis step to remove residual, unbound biotin was included. This was thought to be helpful, because of the lacking endogenous PU.1 protein expression in the lymphatic CTV-1 cells.

The obtained results for the WT PU.1-fusion protein compared to the negative control are summarized in **Figure 5-34**.

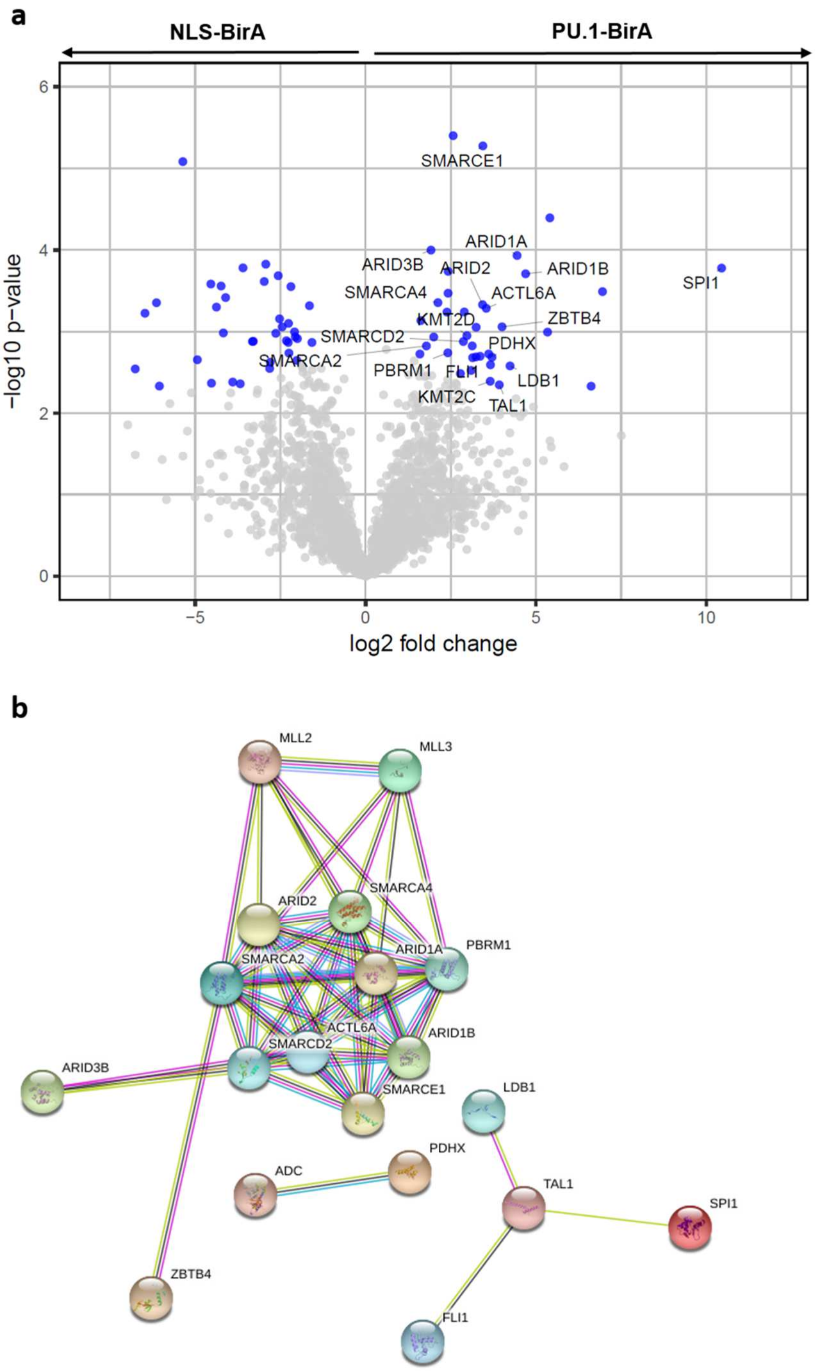


**Figure 5-34 - PU.1 protein-protein interactions in lymphoid CTV-1 cells**

**a |** Volcano plot illustrating proteins significantly enriched in PU.1-BirA-transfected CTV-1 cells, when compared to NLS-BirA-transfected control cells. The significance (-log10 p-value) is plotted against the difference (log2 fold-change) of the normalized intensities. The more significant the difference, the smaller the p-value and thus the higher the -log10 of the p-value. Therefore, points for features with highly significant differences lie high in the plot. Features of interest are those, which change significantly and by a certain magnitude. Blue dots represent proteins with a FDR < 0.05 and a logFC > 2. Only PU.1-specific proteins of GO-terms for chromatin organization and interesting transcriptional regulators are highlighted. **b |** STRING analysis illustrating the functional protein association network. The network view summarizes predicted associations for selected proteins significantly enriched in the PU.1-IP. The network nodes are the proteins, the edges represent the predicted functional associations. Seven types of evidence are used in predicting these associations (i.e.co-occurrence, experimental & database evidence) and only connected nodes are depicted.

First, we filtered the PU.1-specific obtained gene list using Metascape for better visualization of interesting terms. Therefore, only genes associated with chromatin remodeling and transcription are highlighted in **Figure 5-34**. The generated volcano plot illustrates that again mainly components of the SWI/SNF remodeler complex like ARID1B, SMARCA4 (BRG1) and SMARCD2 among others, are associated with PU.1 in the lymphoid cell line. Moreover, cell type-specific TFs like FLI-1 and TAL-1 already found to be enriched in THP-1 and K-562 cells, respectively, are associated with PU.1. In addition, enriched GO-terms strengthen these results as primarily chromatin and nucleosome disassembly, as well as regulation of transcription terms are enriched in the PU.1-specific set (see **supplementary Figure 10-7**). Furthermore, the functional protein association network properly illustrates the enriched SWI/SNF complex and PU.1's (SPI1) association with the ETS-family member FLI-1 as well as with the cell type-specific TF TAL-1 (see **Figure 5-34b**). Since the IP with the PU.1-delQ-BirA fusion protein didn't interfere with the interactome compared to WT PU.1 (solely eleven differential proteins involved in translation were detected), only the results showing the comparison of PU.1 with its acidic transactivation domain lacking mutant are depicted in **Figure 5-35**. FDR-corrected mass spectrometry data were filtered first by comparison to the corresponding negative control (NLS-BirA), before differential enriched proteins between both sets were analyzed. Remarkably, the resulting data showed that proteins, which were significantly enriched in the WT PU.1-IP in contrast to the mutant PU.1, were exactly the components of the chromatin remodeler complex found in association with PU.1 in all analyzed cell types (e.g. ARID1A, ARID1B and SMARCA4/BRG1). This implies that especially the interactions with the SWI/SNF complex are lost upon deletion of PU.1's acidic transactivation domain (see **Figure 5-35**), which is also illustrated in the functional STRING network for PU.1-specific associations in relation to the interactions of the mutant protein. Enriched gene ontology terms are depicted in **supplementary Figure 10-8**.

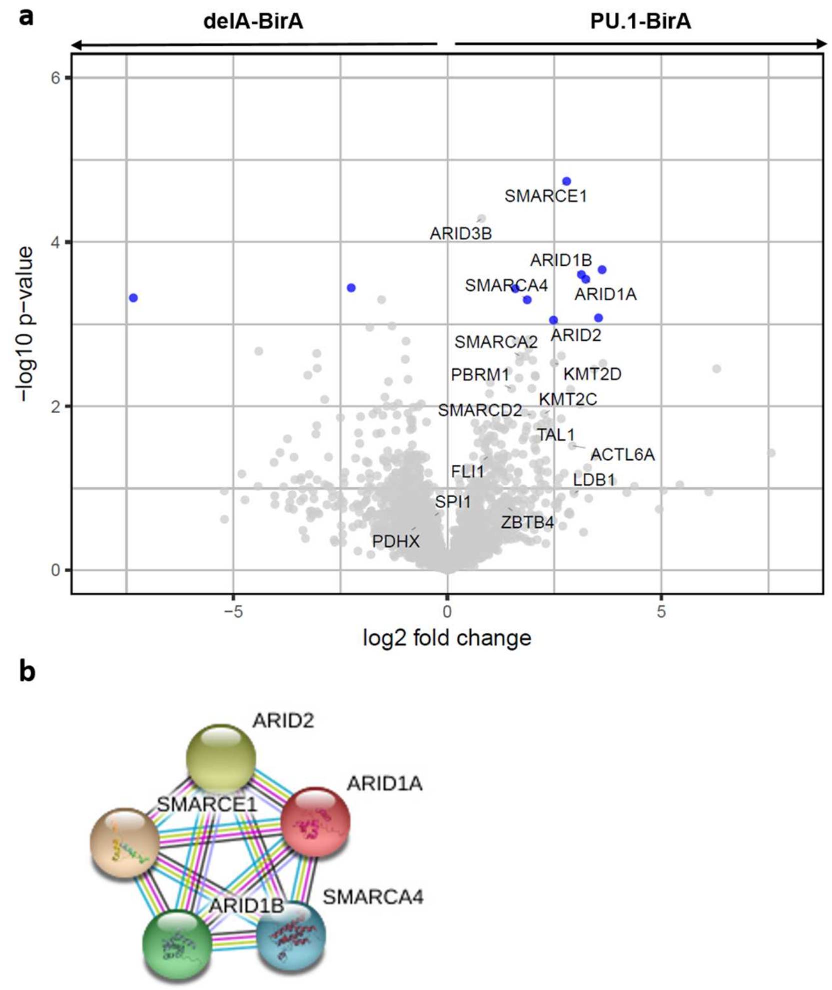


**Figure 5-35 - PU.1 vs. PU.1-delA protein-protein interactions in lymphoid CTV-1 cells**

**a |** Volcano plot illustrating proteins significantly enriched in PU.1-BirA-transfected CTV-1 cells, when compared to PU.1-delA-BirA-transfected cells. The significance (-log10 p-value) is plotted against the difference (log2 fold-change) of the normalized intensities. The more significant the difference, the smaller the p-value and thus the higher the -log10 of the p-value. Therefore, points for features with highly significant differences lie high in the plot. Features of interest are those, which change significantly and by a certain magnitude. Blue dots represent proteins with a FDR < 0.05 and a logFC > 2. Only PU.1-specific (vs. NLS-BirA) proteins of GO-terms for chromatin organization and interesting transcriptional regulators are highlighted. **b |** STRING analysis illustrating the functional protein association network. The network view summarizes predicted associations for selected proteins significantly enriched in the PU.1-IP compared to the PU.1-delA-IP. The network nodes are the proteins, the edges represent the predicted functional associations. Seven types of evidence are used in predicting these associations (i.e.co-occurrence, experimental & database evidence) and only connected nodes are depicted.

Taken together the protein-protein interaction analysis could show that PU.1 is mainly associated with components of the SWI/SNF remodeler complex in all examined lineages, including myeloid, erythroid and lymphoid cell lines. Moreover, cell type-specific already described known and functional interactors were found in close proximity to PU.1 in the tested cell types. Finally yet importantly, the physical interaction with the chromatin remodeler complex could be proven indirectly, since this association was significantly lost in the PU.1-fusion protein lacking the acidic transactivation domain. To analyze these results in relation to PU.1's physiological function, which is binding to nuclear DNA and consequent transcriptional regulation, we aimed to study its binding site selection in the lymphoid cell line upon inhibition of one of the major components of the remodeler complex, namely SMARCA4 or BRG1, found in all cell lines. Therefore, we made use of the selective SMARCA2/4 and polybromo-1 inhibitor PFI-3 (Sigma-Aldrich, Taufkirchen, Germany). CTV-1 cells transiently transfected with WT PU.1 were cultured in PFI-3-containing medium with varying concentrations directly after electroporation and cell lysates for subsequent ChIPseq analysis were harvested after 8 h (only one replicate each). The detailed computational analysis for normalization and peak finding is already listed in section 10.1.8 ('Analysis of NGS data of PU.1-deletion constructs'). The binding preferences of PU.1 upon BRG1-inhibition are shown in the IGV genome browser track below (**Figure 5-36**).

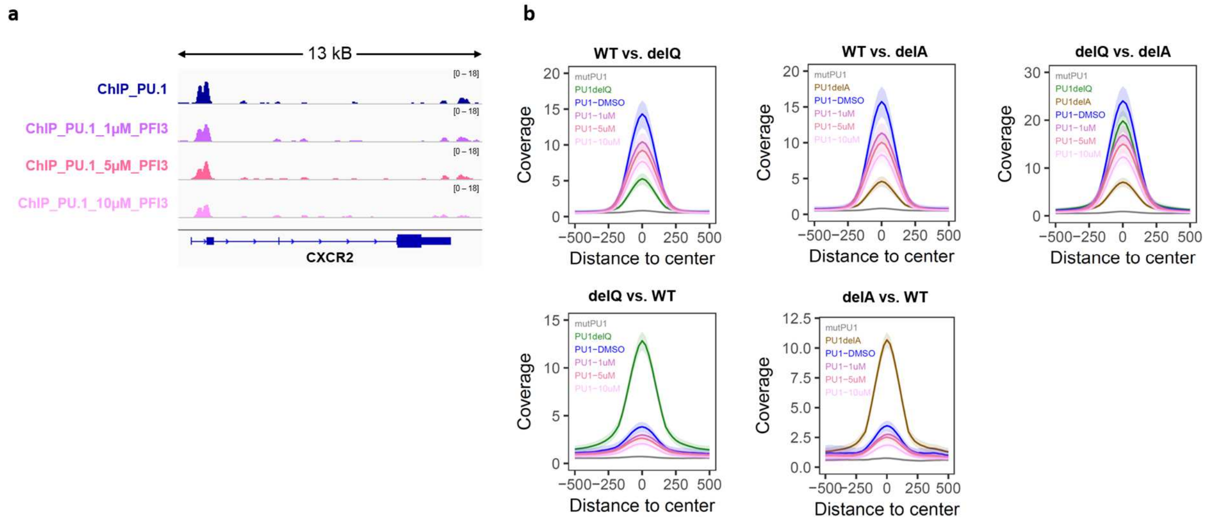


**Figure 5-36 - IGV genome browser track and comparison of ChIPseq reads of PU.1-fusion proteins in CTV-1 cells**

**a |** ChIPseq reads of PU.1 in CTV-1 cells treated with different doses of BRG1-inhibitor (DMSO ctrl., blue; 1 µM PFI-3, orchid; 5 µM PFI-3, violet-red; 10 µM PFI-3, plum) across the CXCR2 locus are shown. All reads are CNV-normalized. The exact localization of the PU.1 binding sites is depicted (chr2:218,989,810-219,003,592). **b |** Histogram plots showing the ChIPseq coverage (y-axis) of PU.1 (DMSO, blue)- and PU.1mut (grey)-transfected cells, as well as of transfected cells treated with different doses PFI-3 (1 µM PFI-3, orchid; 5 µM PFI-3, violet-red; 10 µM PFI-3, plum) across differential peak sets (WT vs. delQ; WT vs. delA; delQ vs. delA; delQ vs. WT; delA vs. WT). The ChIPseq coverage for cells transfected with PU.1-delQ (green) and PU.1-delA (brown) is depicted in addition. The distance to the PU.1 peak center is indicated on the x-axis and the 95% confidence interval is illustrated.

As illustrated in the genome browser track the binding of PU.1 to its specific binding sites in the lymphoid cell line is diminished upon BRG1-inhibition in a dose-dependent manner. Moreover, histogram plots illustrating the PU.1 ChIPseq coverage of the WT PU.1 protein as well as of the delQ- and the delA-mutant proteins across their particular differential bound regions confirm that the coverage of PU.1 is reduced upon PFI-3 treatment in WT-specific regions (upper left histograms). In addition, treatment with 10 µM PFI-3 implies that the inhibiton of BRG1 reduces PU.1's binding capacity to a level comparable to the enrichment of the mutant proteins in WT-enriched regions. In contrast, the binding of PU.1 in delQ- or delA-specific regions is even more reduced upon BRG1-inhibition in a concentration-dependent turn (lower histograms). Strikingly, in regions differential bound by PU.1-delQ in comparison to PU.1-delA the corresponding ChIPseq coverage of PU.1 of treated cells is still higher than the one of PU.1 lacking the acidic domain (upper right histogram). In summary, this analyses suggests that the inhibition of one of the components of the SWI/SNF remodeling complex partly resembles the binding preferences of the mutant proteins. However, since these experiments were only performed once, the data is not statistical significant yet. Furthermore, one should keep in mind that PFI-3-treatment does not completely inhibit BRG1-activity (Fedorov et al. 2015; Gerstenberger et al. 2016) and that its inhibition is not PU.1-specific, but rather occurs on a genome-wide scale, thus interfering with several cellular and transcriptional processes.

# 6 DISCUSSION AND PERSPECTIVES

Regulation of transcription in eukaryotic cells, which determines cell type-specific gene expression programs, involves the dynamic interplay between the chromatin and basal, as well as cell type-specific TFs, which bind to enhancers, upstream activator elements and proximal promotor sequences. Molecular mechanisms that mediate these cell type-specific transcriptional programs, however, remain poorly understood (Gottesfeld and Carey 2018; Heinz et al 2015). So-called pioneer TFs are thought to have the unique and indispensable role of unmasking 'closed' heterochromatin domains to allow for the implementation of new cellular transcriptional programs. This activity implies that they must be able to trigger the remodeling of the local chromatin structure, thereby providing accessibility for additional lineage-determining TFs. During this process, several subsequent steps have been uncovered so far, like the rapid but weak initial binding to closed chromatin, the stabilization of TF-binding followed by chromatin opening and the subsequent loss of CpG methylation, among others (Mayran and Drouin 2018). The exact mechanisms, which allow such master regulators to reorder nucleosomes, and hence initiate the remodeling of the chromatin allowing for efficient binding are only partially understood yet.

A well-studied example of a master regulator is the pioneer TF PU.1. This ETS-family TF is restricted to the hematopoietic system and its binding patterns significantly change during differentiation processes. Prerequisites regulating its binding preferences include its affinity to motif sequences, the cooperativeness between adjacent TF binding sites, as well as its nuclear protein concentration (Pham et al. 2013). The capability to engage chromatin remodeling complexes or epigenetic modifiers, thus, could be shown for the first time in the presented work. However, the contribution of each component and the detailed mechanisms controlling and stabilizing PU.1's access and binding to nuclear DNA are not well understood.  This is why this thesis focused on mechanisms of PU.1 binding site selection with regard to the cooperativeness with additional lineage-determining TFs, the epigenetic regulation of its binding site selection, as well as on the impact of the local chromatin structure, and on its cell type-specific interactome. Diverse high throughput techniques were used to analyze its binding sites in different PU.1-expressing and non-expressing cell types (ChIPseq), its influence on the local chromatin structure (ATACseq), the resulting gene expression profile (RNAseq), as well as its interaction landscape with nuclear factors (BioID). All of the mentioned analyses lead to a more detailed understanding of how this important master regulator gains access to its binding sites on nuclear DNA and how bound sites can be distinguished from non-bound sites in complex mammalian systems.

## 6.1 Cell Type-Specific PU.1 Binding Site Selection

In the eukaryotic system, a single TF is hardly sufficient to control the expression of a target gene alone. To achieve precise gene regulation, cell type-specific and general TFs act together in large cooperative networks. Thus, ChIPseq analyses on TF binding preferences were already extensively studied. However, genome-wide binding profiles for any given factor are still only available for a limited set of human cell types and only few systematic studies exist, which mapped binding sites of multiple TFs across a diverse set of cell types, or of a given TF across several cell types, respectively (Lee et al. 2012). The human hematopoietic system analyzed in this thesis, comprises an ideal model for the analysis of cooperative TF-networks that coordinate the development of distinct blood cell types out of HSCs (Heinz et al. 2010). Here, the lineage-determining TF PU.1 plays a major role in the generation of common myeloid progenitor cells as well as of later stages of MO/MAC and B cells. Previous studies analyzing PU.1 binding site selection, however, were only focusing on few PU.1-expressing cell types (Ghisletti et al. 2010; Heinz et al. 2010; Mullen et al. 2011; Pham et al. 2012/2013; Zarnegar et al. 2010), which is insufficient to study cell specificity on a global scale and thus, to predict cell type-specific TF binding. To bypass this, the genome-wide localization of PU.1 binding sites in diverse blood cell types was analyzed in a comprehensive manner in this thesis with regard to its motif cooperativity with additional cell type-specific TFs. Using this approach, we were able to determine the entire set of *in vivo* occupied sites in 23 different blood cell types. In addition, the influence of motif cooperativeness and motif affinity was analyzed and these analyses unraveled additional mechanisms driving cell type-specific PU.1 binding. Besides, cell type-specific PU.1 binding sites, a huge fraction of common binding sites across correlated cell types, like MO and MAC as well as related myeloid cell lines, could be detected (see **Figure 5-1 and 5-2**). Moreover, a large fraction of cell type-specific PU.1 sites was found to be enriched for additional TF motifs. Cell type-specific PU.1 binding sites of primary B cells for instance, are mainly characterized by cooperative binding motifs for EBOX, ETS and IRF TFs, which is in line with previous studies from Heinz et al. (2010). Here, E2A and CEBP TF motifs were shown to co-occur with PU.1 motifs in murine B cells. However, we could also identify certain cells types, which were less dependent on motif cooperativeness like CD15-positive neutrophils for example. Cell type-specific PU.1 binding sites of those cells were only enriched for additional TF motifs in 40% of their PU.1 peaks (see **Figure 5-3**). This suggests that motif cooperativeness is only able to explain a small fraction of PU.1 binding events and that it is just one factor increasing the likelihood of PU.1 binding. Indeed, the affinity of PU.1 to its recognition sequence is also an important prerequisite determining the PU.1 binding site selection (Pham et al. 2013), which could also be shown in our systematic approach as the distribution of PU.1's motif affinities in cell type-specific peaks was highly cell type-dependent. High motif scores were present in almost all analyzed myeloid cell lines, whereas primary CD19-positive B cells, mainly characterized by cooperative binding sites, showed a lower PU.1 motif score enrichment (see **Figure 5-2**). Indeed, cell type-specific cooperativeness and motif affinity can only explain some binding events, as only a small fraction of all putative binding sites throughout the genome is actually occupied by PU.1 in all analyzed cell types. Therefore, PU.1 binding site selection has to be dependent on additional features like the restriction through epigenetic mechanisms for instance.

## 6.2 Restriction of PU.1 Binding Site Selection

To study the impact of epigenetic regulation, *in vitro* and *in vivo* experiments were conducted to unveil if DNA methylation is able to inhibit PU.1 binding or rather, if PU.1 binding leads to the subsequent demethylation of its binding sites. Preliminary work already showed that there is little overlap between PU.1 binding and DNA methylation, at least in human MO, as the cell type-specific recruitment of PU.1 is associated with the active demethylation of its binding sites during MO to MAC differentiation (Pham et al. 2013). However, gene bodies of highly expressed genes are normally heavily methylated (Yin et al. 2017), which implies that DNA methylation is able to restrict TF binding (Domcke et al. 2015; Moore et al. 2013). In line with this, *in vitro* MST-measurements with hemi-methylated PU.1 consensus sites uncovered that the binding ability of PU.1 is impaired in an asymmetric manner, when DNA methylation occurs next to its core recognition sequence (see **Table 5-2** & **Table 5-3**). This was also shown previously by the group of Gregory Poon using fluorescently labeled DNA to study interrogated site-specific binding of PU.1 (Stephens et al. 2016). Since asymmetric CpG-demethylation normally only occurs during replication, this implies that PU.1-induced cell fate decisions could directly be made during mitosis, the only time-point when hemi-methylated DNA is present in somatic cells. How transcriptional gene regulation at different cell-cycle stages occurs is still largely unknown. Although a fraction of TFs including the mixed-lineage leukemia (MLL), bromo-domain containing four (BRD4), Forkhead box A1 (FOXA1) and the GATA-binding protein 1 (GATA1) are known to be retained on mitotic chromosomes (Liu et al. 2017), studies investigating the retention of the pioneer TF PU.1 haven't been made so far. Taken together, own as well as published *in vitro* analysis of PU.1 binding to methylated DNA suggests, that PU.1 is at least in part able to bypass epigenetic restriction, however it is still not clear how PU.1 is able to bypass or compete with epigenetic restriction *in vivo*. To address this question, we also studied its binding preferences on methylated vs. unmethylated DNA *in vivo*. We made use of the lymphoid CTV-1 cell line, which does not express PU.1 and likely contains a high degree of methylated PU.1 binding sites in comparison to human primary MO as analyzed previously by MCIp-sequencing (depicted in **Figure 5-6**). Global demethylation was achieved by DAC-treatment and methylated vs. unmethylated cells were compared regarding their particular gene expression profile and their corresponding chromatin accessibility. This analysis revealed that a large fraction of *de novo* binding sites was mainly induced by the overexpression of PU.1, rather than through the global demethylation of the genomic DNA upon DAC-treatment (25406 vs. 6572 binding sites respectively as seen in the Venn diagram in **Figure 5-11**). *De novo* PU.1-bound regions coming along with its expression were less correlated with CpG-methylation and showed a lower overall GC-enrichment compared to bound regions in which the access was gained by PU.1 in combination with the DAC-treatment (**Figure 5-12**). This correlates with own published work in PU.1-expressing cells, were PU.1 binding sites were shown to be rarely associated with nearby CpG-methylation (Pham et al. 2013). Similar observations could recently be made regarding the hematopoietic master TF RUNX1, which was shown to contribute to DNA demethylation in a site-directed manner in human hematopoietic cells. Overexpression of this pioneer TF led to direct DNA demethylation and co-immunoprecipitation assays uncovered physical interactions between RUNX1 and DNA demethylation machinery enzymes like TET2 and TET3 for instance (Suzuki et al. 2017). This was also shown for PU.1 before, as the recruitment of important epigenetic modifiers like DNMT3B, as well as the recruitment of TET2 to genes that become demethylated during MO to OC differentiation, was shown to be dependent on PU.1 (de la Rica et al. 2013). Although, RUNX1 overexpression was carried out in human embryonic kidney 293T cells, rather than in hematopoietic cell lines, which likely differ regarding their particular methylation and gene expression profile, the enrichment of its consensus motif was mainly found in regions, which get demethylated during early hematopoietic differentiation and strongly correlated with its endogenous expression level in the analyzed cell types (Suzuki et al. 2017).  The PU.1-induced transcriptome of DAC-treated CTV-1 cells, however, only showed a small overlap between the expression of myeloid genes and endogenous PU.1 expression as seen in the scatter plot of **Figure 5-9**. PU.1-induced genes of DAC-treated CTV-1 cells were mainly involved in immune response and interferon-gamma (INF-γ) mediated signaling pathways. This of course could be due to the induced stress upon DAC-treatment and the transfection procedure itself, but is also in line with published data stating that PU.1 is able to regulate innate immune responses under certain circumstances. In this context, PU.1 was shown to mediate GM-CSF-dependent effects on terminal differentiation of alveolar MAC (AMs; Shibata et al. 2001) and was found to be involved in IL-4 mediated alternative alveolar MAC (AAM) polarization and eosinophilic asthmatic inflammation (Qian et al. 2015), which is coupled to the upregulation of individual immune signaling genes. However, cause and consequence of DNA demethylation and PU.1 binding still cannot be distinguished with our model system, but are likely initiated by the concerted activity of sequence-specific TFs. In line with this, we found that the global demethylation of nuclear CTV-1 DNA leads to the establishment of novel co-associations, as GATA TF motifs were mainly found to co-occur with PU.1 binding sites in DAC-specific regions (see **Figure 5-11**).

The interaction of GATA-1 and PU.1 was already analyzed in 2000 by Nerlov et al., who proposed that GATA-1 represses myeloid gene expression when expressed in myeloid cells via the direct interference with PU.1's function in an antagonistic manner. Nonetheless, not all *de novo* bound regions showed this co-association, suggesting that additional mechanisms might exist, which distinguish bound vs. non-bound sites. To bypass the limitations of our model system, additional approaches like whole-genome bisulfite sequencing of conventional and PU.1-expressing CTV-1 cells treated with DAC or left untreated may unravel further mechanisms involving PU.1 binding and the impact of epigenetic restriction.

## 6.3 *De novo* Binding Site Selection & Chromatin Remodeling

PU.1 binding analysis in PU.1-expressing cell types only provides a static view of TF binding in the context of the chromatin structure, since the pre-existing chromatin landscape established before the appearance of PU.1 is not taken into account. Therefore, one is not able to distinguish if the analyzed TF induces local changes in the chromatin structure to gain access to its binding sites, or if pre-existing accessible chromatin domains are bound instead.  To study TF occupancy in an unbiased manner, we made use of the lymphoid CTV-1 cell line, which does neither express PU.1, nor its ETS-family members SPIB and SPIC as analyzed by RNA-sequencing. Moreover, its known hetero-dimerization partners IRF4 and IRF8 were not found to be expressed in those cells (see **Figure 5-14**). CTV-1 cells thus, provide the ideal model to study PU.1 binding dynamics without the interference of its conventional binding prerequisites and with regard to the particular chromatin landscape of these cells. For overexpression proposes, a transient mRNA-transfection model was used. As already described 1990, most current gene transfer methods only function satisfactorily in specialized systems primarily involving adherent cell lines, however non-adherent primary hematopoietic cells and hematopoietic cell lines are still notoriously difficult to transfect, because of their rapid immune response to extrinsic signals mediated by several antigens expressed on these cells (Zenke et al.). Therefore, it is almost impossible to generate stable, inducible cell lines without the need of lentiviral transduction procedures for the hematopoietic system. To avoid this, we made use of the direct electroporation of PU.1-mRNA, which did not evoke an immune response as measured by RT-qPCR for the MX1 target gene (data not shown). MX1 is a dynamin like GTPase that is part of the antiviral response induced by type I and II interferons, inhibits virus infection by blocking viral transcription and replication (Verhelst et al. 2012), and is highly expressed in activated immune cells upon recognition of foreign nucleic acids. Therefore, it serves as the ideal marker to investigate the degree of activation of hematopoietic cells and cell lines upon mRNA transfection procedures. Since the ideal method should also have a high transfection efficiency and cell viability, as well as low cell toxicity, transfection efficiency and cell viability were also analyzed and shown to be extremely high with our methodology in all tested cells. This fits well with published data on mRNA-transfected DCs (Van Tendeloo et al. 2001; see **Figure 5-5**). As mRNA transfection efficiency is cell cycle-independent, there is no need of generating inducible vectors in addition. Moreover, it leads to an adjustable and rapid protein expression (Kim and Eberwine 2010). Nevertheless, induced gene expression changes are only of a short duration and longitudinal overexpression effects cannot be studied using mRNA transfections. Therefore, additional transfection procedures have been developed including the inducible PUER expression system. PU.1 overexpression in these cells derived from the fetal liver of PU.1$^{-/-}$ mice, was achieved by the fusion of PU.1 to the ligand-binding domain of the estrogen receptor, which is preferentially regulated by tamoxifen. These progenitor cells were shown to have the capacity to differentiate into macrophages upon tamoxifen treatment as long-term differentiation studies can be easily carried out using inducible and stable cell lines. This system thus, is also limited as the basal expression of the PUER protein was shown to alter the developmental capacity of the cells, likely because the PUER expression is considerably higher than PU.1 expression after retroviral transduction for example (Walsh et al. 2002). Heinz et al. also showed the effect of basal PUER expression in 2010, where low levels of PU.1 activity and DNA binding were observed by ChIPseq even in the absence of tamoxifen. Although, PU.1 binding patterns and its impact on the chromatin can be studied in a long-term approach with this system, it is still limited to murine cells and currently not available for the human system. With our approach, we were able to expand our studies to human hematopoietic cells and cell lines, therefore bypassing the time-consuming and complex strategy to generate stable and inducible PU.1-expressing cells through lentiviral transduction.

### 6.3.1 *De novo* Binding Site Selection in Lymphoid CTV-1 Cells

Differential gene expression analysis in CTV-1 cells transfected with PU.1 and PU.1mut mRNA unveiled, that the PU.1-induced transcriptome of these cells was highly associated with gene expression patterns observed in myeloid cells **(Figure 5-17)**. Moreover, the induced gene expression profile highly correlated with the PU.1 expression in different hematopoietic cell types as the Pearson correlation between PU.1-induced genes in CTV-1 cells and its expression in distinct blood cells was 0.78 with myeloid cells being the ones correlating the best (see **Figure 5-18**). Nishiyama et al. already made similar observations in 2004 showing that overproduction of PU.1 in mouse bone marrow-derived mast cell progenitors induces the expression of MO-specific genes. Besides rapid and severe gene expression changes from a lymphoid to a myeloid phenotype, we also studied PU.1's *de novo* binding site selection. PU.1 bound roughly 45.000 sites in the lymphoid cell line, which were either correlated with closed chromatin domains, pre-existing accessible ones or *de novo* opened sites as revealed by ATAC-sequencing. The extensive, PU.1-induced remodeling of the chromatin was partially associated with the deposition of the active histone modification H3K27ac and the enhanced expression of neighboring genes **(Figure 5-19)**. Furthermore, *de novo* remodeled sites were significantly associated with higher PU.1 motif scores (see **Figure 5-20**) and with the co-occurrence of clustered PU.1 binding sites (see **Figure 5-22**), suggesting that homotypic consensus sites and high affinity motifs are responsible for a large fraction of *de novo* remodeled PU.1 binding sites in this cell line, which is in line with previous studies (Pham et al. 2013). Additional reports even showed that homotypic motif clusters are often found at genes regulated by environmental TFs, which operate independently and regulate Pol II recruitment additively (Giorgetti et al. 2010). Furthermore, such homotypic TF binding sites were found to be able to buffer the effect of genetic variation among various generations of lymphatic cell lines (Kilpinen et al. 2013). PU.1 binding in pre-existing open chromatin, however, was predominantly found at single PU.1 binding sites with lower motif scores and many surrounding consensus motifs for other transcription factor families, including GATA, ETS, RUNX and AP-1 **(Figure 5-21)**. This cooperativeness likely enhances or even enables PU.1 binding at low affinity sites as already shown by several groups (Eeckhoute et al. 2009; Heinz et al. 2010; Pham et al. 2012/2013) for murine and human MO, MAC and B cells. In addition, PU.1-bound motifs of pre-existing accessible sites showed a higher degree of conservation and were primarily located in promotor regions when compared to *de novo* bound sites, which were less conserved and primarily located in intergenic regions (see **Figure 5-21**).

Interestingly, we found evidence that the ETS-related family members ETS-1 and FLI-1 endogenously expressed in CTV-1 cells (see **Figure 5-14**) are not able to induce the same extensive reorganization of the chromatin at *de novo* remodeled PU.1 binding sites. The overlap of binding sites of the ETS TFs was mainly occurring in already accessible regions, but not in regions, in which *de novo* access was gained upon PU.1 expression (see **Figure 5-23**). In terms of *de novo* remodeled chromatin, this is in contradiction to previous studies, reporting that in several cases (e.g. for IRF and AP-1 TFs) a TF from one homotypic family can be joined or replaced by another factor from the same family, thereby maintaining a binding site and serving as a docking point for additional related factors (Garber et al. 2012). However, this holds true at accessible sites, since we observed an exchange or rather an overlap of ETS-1, FLI-1 and PU.1 at these binding sites, which likely stabilizes TF binding and subsequent chromatin opening. Combinatorial transcriptional control in blood progenitor cells for ten key TFs including PU.1 and FLI-1 was already analyzed 2010 on a genome-wide scale. These analyses demonstrated that the median distance between heterotypic, cooperative binding sites ranges from 22 bp to 85 bp (Wilson et al. 2010). This is in line with our data, as the optimal distance between PU.1 and ETS class 1 motifs ranged from 12 bp to 50 bp. Interestingly, shared sites were mainly located in open chromatin, with lower PU.1 motif scores and higher ETS motif scores co-associated with additional cell type-specific TF motifs like RUNX1 sites (see **Figure 5-24**). In summary, this suggests that the ETS-family members enhance efficient chromatin remodeling upon PU.1 expression, and that PU.1 seems to be the major driver regarding *de novo* chromatin remodeling.

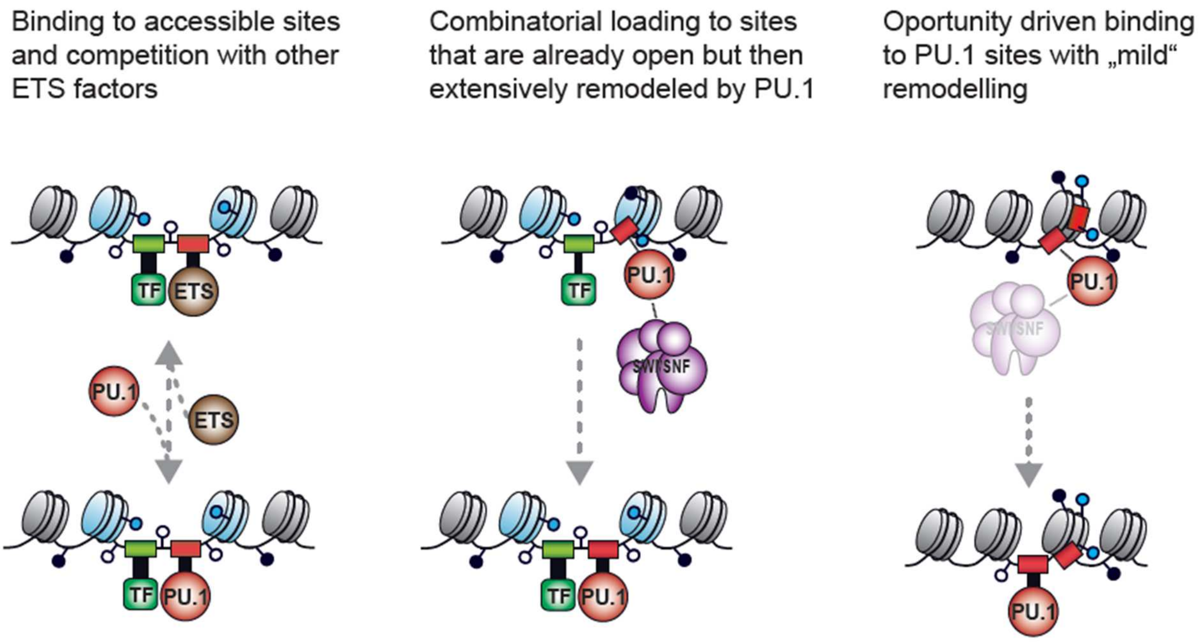
To analyze the impact of the nuclear concentration of PU.1 and its potential interaction with other nuclear factors, helping to establish stable binding on *de novo* remodeled sites, the titration of PU.1 levels and the analysis of several deletion mutants was studied as well. This analysis showed that the efficient binding of PU.1 to *de novo* remodeled sites is PU.1 concentration dependent, since lower levels of PU.1 reduce its binding, with the most drastic effect observed in regions of newly induced open chromatin. Moreover, lower PU.1 concentration mostly led to the loss of PU.1 binding in regions comprised of clustered binding sites, which are primarily located in *de novo* remodeled regions (see **Figure 5-28**). This observation further suggests that homotypic binding sites are crucial for PU.1's binding site selection in chromatin-restricted regions. In addition, we analyzed its binding using particular PU.1-deletion mutants lacking either parts of the N-terminal TAD or of the PEST domain. The nuclear localization and DNA binding activity of several PU.1-deletion constructs has already been extensively studied before using *in vitro* Gel shift assays or immunohistochemistry (Delva et al. 2004; Kwok, 2007; Yee et al. 1998) and could be further elucidated during this study with the current advantage of high-throughput sequencing, being much more sensitive and therefore informative than classical biochemical approaches. Hence, PU.1's binding *in vivo* was significantly reduced in the absence of the glutamine-rich domain and even more diminished in the absence of the acidic domain both together forming its so-called TAD, as illustrated in **Figure 5-27**. The loss of the PEST-domain on the other hand, did not impair PU.1's binding site selection at all, with no differential peaks detected between the WT and this deletion mutant. This observations can be confirmed by experiments of Bruce Torbett`s group from 2006, demonstrating that an immature myeloid cell line derived from *Sfpi1*-null mice lacking the acidic or the glutamine-rich PU.1 domain showed severe impacts on the differentiation of these cells (Foos et al. 2006). Another study showed that the acidic region is required for the expression of MO-specific genes in transfected mast cells and for an enhanced IL-6 production in response to LPS, whereas the glutamine-rich region is involved in the expression of MHC class II genes and the PEST-domain even had a negative role in MO-specific gene expression (Nishiyama et al. 2004). Our data supplements several studies, which analyzed the involvement of PU.1's protein domains in different protein-protein interactions and showed that PU.1 is physically interacting with general TFs like TFIID and TBP, with early hematopoietic TFs like GATA2 and RUNX1, and with other cell type-specific TFs as IRF4/8 or c-Jun (Burda et al. 2010; Tenen 2003) for instance. The effect of each particular domain in the context of chromatin accessibility though has not been analyzed in detail yet, so that our analyses contribute to a better understanding of the involvement of the non-DNA binding PU.1 domains in the context of its pioneering activity. For this purpose, the impact of the different deletion mutants in terms of *de novo* remodeling induced by the expression of the WT protein was studied with particular interest. Interestingly, we found the strongest effect induced by the loss of the delQ- or delA-domain in regions harboring a high remodeling index, thus in regions were WT PU.1 expression led to the opening of the chromatin in CTV-1 cells (see **Figure 5-28**). Strikingly, differential PU.1 peaks between the delQ- compared to delA-construct were also primarily associated with *de novo* remodeled sites, suggesting that PU.1 is even more dependent on its acidic transactivation domain in terms of chromatin remodeling. Moreover, the glutamine-rich domain lacking protein did not differ much from the WT protein titrated to 15% in terms of PU.1 binding sites, indicating that this domain might rather mimic concentration-dependent binding events, which seem mainly important at homotypic binding sites. In line with this, the PU.1 ChIPseq coverage of the delQ-mutant was decreased the most in regions with clustered, homotypic PU.1 binding sites whereas the coverage of the delA-mutant was lost in almost all *de novo* remodeled sites independently of the motif composition (see **Figure 5-28** and **Figure 5-29**). Together, these data shows that the efficient binding of PU.1 to *de novo* remodeled sites is concentration-dependent, significantly reduced in the absence of the glutamine-rich transactivation domain and even more diminished in the absence of the its acidic domain. This suggests that the latter one could be required for the access to binding sites located in closed chromatin.

A study by Ellen Rothenberg's group recently published, described similar observations on pro-T cells. Here they showed that PU.1 is be able to bind closed as well as open chromatin in different developmental T cell stages in concordance with its own site-specificity, its motif affinity and cooperativity, and its local concentration in these cells. In addition, its non-DNA binding domains were shown to be important for efficient and stable chromatin remodeling (Ungerbäck et al. 2018). However, these analyses were mainly performed in PU.1-expressing cells, thereby likely interfering with its own endogenous protein expression. Furthermore, the deletion constructs used did not distinguish between its described protein-interactions domains, but were just separated into the N-terminal protein domain and the C-terminal DNA-binding domain, which is not sufficient to unveil the particular impact of each protein-interaction domain. In addition, the effect of this loss of function constructs was only studied regarding PU.1's genome occupancy and its pioneering activity, but functional implications on the required nuclear interactome were not followed up. Therefore, the experiments performed in this thesis have several advantages as the pioneering activity of PU.1 was studied in cells, which do not express PU.1 endogenously. Moreover, nuclear interactions of WT PU.1 and its acidic protein domain were studied for the first time using a quantitative mass spectrometry approach, which will be discussed in the following chapter.

### 6.3.2 Interaction of PU.1 and Chromatin Modifying Enzymes

Since we uncovered severe changes in the local organization of the chromatin induced by PU.1 overexpression in lymphatic CTV-1 cells, we further studied, which features besides protein-DNA interactions might be involved in PU.1's binding site selection. Therefore, protein-protein interactions of the nuclear protein were mapped using the proximity-dependent BioID-approach. Since the loss of the acidic transactivation domain led to a decreased binding affinity to *de novo* remodeled sites, we also anaylzed the interactome of this particular PU.1 mutant protein. These analyses revealed cell type-specific interactions with other lineage-determining TFs and regulatory proteins in all tested hematopoietic cell lines covering the myeloid, erythroid and lymphoid lineage. In myeloid THP-1 cells for instance, PU.1 was strongly associated with the active DNA-demethylation enzyme TET2 as already described before in human MO (de la Rica et al. 2013), whereas in lymphoid CTV-1 and erythroid K-562 cells PU.1 was significantly associated with the cell type-specific TF TAL-1. FLI-1 was also found to be enriched in the PU.1-specific interactome of THP-1 and CTV-1 cells (see **Figures 5-32 - 5-34**), suggesting that this ETS-family TF might interact with PU.1 or even adopt to its function in cell types lacking endogenous PU.1 expression like CTV-1 cells. Furthermore, PU.1 was associated with several components of the SWI/SNF family of chromatin remodeling complexes, including ARID1A, SMARCD2 and SMARCA4 (BRG1) among others in all analyzed cell lines **(Figures 5-32 - 5-34)**. Since the SWI/SNF complex itself is known to have only limited ability to bind to sequence-specific elements, its recruitment to target loci is believed to require its interaction with sequence-specific TFs *in vivo*. These kind of interactions could already be shown using quantitative affinity purification coupled to mass spectroscopy for the homeodomain TF CDX2 as an example, which was associated with multiple members of the BRG1-containing SWI/SNF complex (Nguyen et al. 2017). Furthermore, BRG1 itself has been implicated in the activation and repression of gene expression in various tissues via its association with several TFs and histone-modifying enzymes (Trotter and Archer 2008). Likewise, the human SWI/SNF complex was found to participate in nuclear receptor-mediated transcriptional regulation. In addition, the glucocorticoid receptor (GR) and the estrogen receptor (ER) have been shown to recruit SWI/SNF to responsive promoters. Several diverse transcriptional activators like C/EBPβ, c-Myc and the heat shock factor 1 (HSF1) for instance have been found to recruit SWI/SNF to specific promoters, which was at least in some cases coupled to the activation of gene expression. Although the molecular details of the association between SWI/SNF and transcription activators remain largely unknown, there is evidence that points to the importance of specific features of transcriptional activation domains (Narlikar et al. 2002). Interestingly an *in vitro* transcription study performed in 1999, utilizing the yeast SWI/SNF remodeler complex as well as yeast-specific master TFs, already indicated that this remodeler complex is dependent on the direct interaction with an acidic activation domain to facilitate efficient transcription from nucleosome templates (Neely et al.). This view is supported by additional reports that suggest that SWI/SNF-targeting to nuclear DNA is reduced or even eliminated by mutations disrupting the acidic residues of transcriptional activators (Hassan et al. 2001; Peterson and Workman 2000). Remarkably, our observed PU.1-interactions with components of the SWI/SNF remodeler complex were specifically lost in CTV-1 cells transfected with the PU.1 mutant lacking its acidic transactivation domain **(Figure 5-35)**. This strongly suggests that the direct interaction of PU.1's acidic transactivation domain with certain components of the chromatin-remodeling complex is responsible for gaining access to its binding sites located in closed chromatin domains as observed in the analyzed CTV-1 model system. Along with this, several studies recently published in Nature Genetics could demonstrate similar examinations involving the SMARCD2 subunit of the SWI/SNF complex, which was able to mediate granulopoiesis through a C/EBPε-dependent mechanism (Michael and Kadoch 2017; Priam et al. 2017; Witzel et al. 2017).

In addition, BRG1-inhibiton experiments performed in this study further strengthen the idea of the direct interaction of PU.1 and a subunit of the SWI/SNF complex as its binding ability is strongly reduced in cells treated with the small molecule inhibitor PFI-3 (see **Figure 5-36**).



Binding to accessible sites and competition with other ETS factors

Combinatorial loading to sites that are already open but then extensively remodeled by PU.1

Oportunity driven binding to PU.1 sites with „mild“ remodelling

**Figure 6-1 - Schematic view of PU.1 binding site selection in terms of the chromatin structure**

Conclusively, the results obtained in this thesis expand our current understanding of PU.1’s binding site selection by supporting a mechanism that, besides motif cooperativeness, epigenetic regulation and the nuclear concentration of PU.1, involves the interaction with the SWI/SNF complex to overcome chromatin restriction and thus allows for stable *de novo* binding. Our current hypothesis on the binding site selection of this important master regulator suggests that three different classes of PU.1 consensus recognition sites exist depending on the higher-order chromatin structure of nuclear DNA. Pre-existing accessible sites are defined by the exchange or rather, competition with other ETS-family related TFs as illustrated for ETS-1 and FLI-1 in this thesis. Less wide open chromatin regions on the other hand, are bound by PU.1 in combination with additional cell type-specific TFs including several TF groups like the RUNX, GATA or ETS family for example. Combinatorial binding likely stabilizes these sites, which are then extensively remodeled by PU.1 upon its interaction with the SWI/SNF remodeler complex. Closed chromatin domains lastly, represent an opportunity for PU.1 binding likely driven by its association with the SWI/SNF remodeler complex as summarized in **Figure 6-1**.

## 6.4 Perspectives

The presented analyses elucidated a couple of interesting and novel aspects contributing to our current understanding of how a pioneering TF is able to interfere with the higher-order chromatin structure, thus allowing for efficient and stable binding to nuclear DNA. Different aspects involved in the binding site selection of the master regulator of the hematopoietic system PU.1 have been analyzed using sequencing-based methods as well as mass spectrometry all leading to a more detailed characterization of PU.1's binding preferences. Motif cooperativeness was found to be important at low affinity binding sites, where additional TF binding enhances PU.1's binding affinity in an additive manner. DNA methylation was shown to impair efficient TF binding at least to some extent. However, demethylation analyses should be repeated using whole-genome bisulfite sequencing to further study cause and consequence of TF binding and DNA demethylation. Moreover, the TF's dependency on homotypic clusters within *de novo* remodeled binding sites comprised of high-affinity PU.1 motifs was analyzed and will be followed up with special interest to unravel if PU.1 binding follows an additive, or rather a synergistic pattern. As the optimal distance between two PU.1 motifs could be cut down to 12-50 bp **(Figure 5-22)**, its binding affinity to these elements will be analyzed *in vitro* using conventional Gel-shift experiments or MST-affinity measurements, respectively, and CTV-1 nuclear extracts of the WT protein and its two transactivation domain mutant proteins. This will be of special interest regarding the glutamine-rich mutant protein, since the loss of this domain seems to impair PU.1's binding affinity in a concentration-dependent manner, rather than through the interaction with additional nuclear factors. The acidic-domain lacking protein will be further examined regarding its association with the BRG1-containing SWI/SNF remodeler complex. In this context, the inhibition of BRG1 might lead to the loss of PU.1 especially in *de novo* remodeled regions as already suggested by the presented data, although replicates of these experiments have to be performed to allow for statistical significance. Nonetheless, one should keep in mind that the inhibition of this important chromatin-modifying enzyme, is not only influencing the binding of a single TF like PU.1, but also mediates several steps involved in normal gene expression, so that this approach will be biased by interfering with a lot of nuclear transcriptional pathways. To bypass this limitation or rather in addition, we also aim to analyze BRG1's binding site selection *in vivo* in PU.1-transfected vs. conventional CTV-1 cells using ChIP- and ATAC-sequencing. This will allow for the comparison of PU.1- and BRG1-bound sites, and possibly further strengthen our findings on the *de novo* binding site selection of the hematopoietic TF. As already discussed, ETS-1 and FLI-1 were not able to gain access to *de novo* remodeled sites, without PU.1 being expressed in those cells. However, the overlap of binding sites of these factors with PU.1 binding sites might stabilize *de novo* induced chromatin remodeling, why the occupancy of all factors on heterotypic binding sites will be analyzed in more detail. In addition, we will aim to overexpress SPIB in this cell line and to compare its binding patterns as well as its interactome with the one obtained for PU.1 expression. SPIB also belongs to class 3 of ETS proteins and binds the same consensus sequence as PU.1. This TF is not expressed in the lymphoid cell line and its domain structure is very similar to the one of PU.1, also harboring a N-terminal acidic transactivation domain, which will be deleted as well (Carlsson et al. 2003). The analyses of the overexpression of SPIB and its protein-domain mutants in comparison to the presented data obtained for PU.1, will thus be used to further characterize PU.1-induced *de novo* chromatin remodeling in the lymphatic cell line.

# 7 ZUSAMMENFASSUNG

Transkriptionsfaktoren definieren sich durch ihre Fähigkeit spezifische Basenabfolgen in genomischer DNA zu erkennen und zu binden. Diese Sequenzmotive sind im Vergleich zur Länge des humanen Genoms erstaunlich klein, weshalb es für viele Transkriptionsfaktoren einen großen Anteil an nicht-funktionalen Motiven gibt, an denen ihre Bindung verhindert werden soll. Der myeloische und B-zellspezifische Transkriptionsfaktor PU.1 stellt ein hervorragendes Modell zur Untersuchung globaler, dynamischer Bindungsprozesse dar. Dieser Transkriptionsfaktor ist ein zentraler Regulator während der hämatopoetischen Zelldifferenzierung und spielt durch seine Regulation von zelltypspezifischen Genen eine entscheidende Rolle in verschiedenen hämatopoetischen Zelllinien. Bisher ist nur wenig verstanden, wie dieser Master-Transkriptionsfaktor Zugang zu seinen Bindestellen im Kontext der Chromatinstruktur erlangt. Neben der Analyse seiner Motivkooperativität und der epigenetischen Regulierung, haben wir ein transientes mRNA-Transfektionsmodel verwendet, um seine *de novo* Bindung in der lymphatischen Leukämie-Zelllinie CTV-1 zu untersuchen. Diese Zelllinie exprimiert weder PU.1, noch seine Verwandten SPIB und SPIC der ETS-Familie. Die PU.1 Expression initiierte hier rasch ein spezifisches Genexpressionsprogramm, welches vor allem von myeloiden Genen dominiert wurde, die mit der PU.1-Expression in verschiedenen hämatopoetischen Linien korrelierten. ATACseq-Analysen enthüllten eine umfangreiche Umstrukturierung des Chromatins in Folge der PU.1-Expression, welche teils mit dem Vorhandensein der aktiven Histonmodifizierung H3K27ac, sowie mit der gesteigerten Expression benachbarter Gene assoziiert war. Umstrukturierte Regionen waren signifikant mit homotypischen PU.1 Bindestellen und/oder höheren Motiv-*Scores* assoziiert, was vermuten lässt, dass homotypische Bindestellen und hochaffine Konsensusmotive für eine große Anzahl der Bindeereignissen in diesen Regionen verantwortlich sind. Außerdem scheinen heterotypische Bindestellen zwischen PU.1 und seinen Verwandten der ETS-Familie ETS-1 und FLI-1, die Fähigkeit von PU.1, Chromatinstrukturen zu öffnen, zu verstärken, was möglicherweise durch neue ETS-abhängige Co-assoziationen in weniger offenem Chromatin vermittelt wird. Die Bindung an bereits offenes Chromatin jedoch, wurde hauptsächlich in PU.1-Regionen gefunden, welche mit niedrigeren Motiv-*Scores* und vielen benachbarten Konsensus-Motiven anderer Transkriptionsfaktorfamilien, wie GATA, RUNX und AP-1, assoziiert waren. Diese ermöglichen wahrscheinlich die Bindung von PU.1 an niedrigaffine Bindestellen. PU.1-Titrations-Experimente und die Analyse mehrerer Deletions-Mutanten haben gezeigt, dass die erfolgreiche Bindung an neue Bindestellen konzentrationsabhängig ist und dass PU.1-Mutanten mit fehlender Glutamin-reicher Domäne und noch stärker, mit fehlender saurer Domäne, den Zugang zu Bindestellen in geschlossenem Chromatin verlieren. Die *in vivo* Biotinylierung benachbarter Proteine (BioID) konnte zeigen, dass PU.1 mit verschiedenen

Komponenten des SWI/SNF-Komplexes, wie zum Beispiel ARID1A, SMARCD2 und SMARCA4 (BRG1), assoziiert ist. Diese Interaktionen waren speziell in der PU.1-Mutante ohne die saure Domäne nicht zu detektieren. Zusammenfassend konnte also gezeigt werden, dass die *de novo* Bindung von PU.1 an seine Bindestellen in nukleärer DNA rasche und weitreichende Änderungen in der Chromatinstruktur von lymphatischen CTV-1 Zellen verursacht, welche durch die Interaktion seiner sauren Domäne mit dem SWI/SNF-Komplex vermittelt werden.

# 8 REFERENCES

**Adcock IM, Caranori G. 2009.** Transcription factors. *Asthma and COPD.* 2009, Vol. 2.

**Akbari OS, Bae E, Johnsen H, Villaluz A, Wong D, Drewell RA. 2008.** A novel promoter-tethering element regulates enhancer-driven gene expression at the bithorax complex in the Drosophila embryo. *Development.* 2008, Vol. 135, 1, pp. 123-131.

**Andreesen R, Picht J, Löhr GW. 1983.** Primary cultures of human blood-born macrophages grown on hydrophobic teflon membranes. *Journal of immunological methods.* 1983, Vol. 56, 3, pp. 295-304.

**Arinobu Y, Mizuno S, Chong Y, Shigematsu H, Iino T, Iwasaki H, Graf T, Mayfield R, Chan S, Kastner P Akashi K. 2007.** Reciprocal activation of GATA-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. *Cell Stem Cell.* 2007, Vol. 1, 4, pp. 416-427.

**Bannister AJ, Kouzarides T. 2011.** Regulation of chromatin by histone modifications. *Cell Research.* 2011, Vol. 21, 3, pp. 381-395.

**Bell O, Tiwari VK, Thomä NH, Schübeler D. 2011.** Determinants and dynamics of genome accessibility. *Nature Reviews Genetics.* 2011, Vol. 12, 8, pp. 554-564.

**Buecker C, Wysocka J. 2012.** Enhancers as information integration hubs in development: lessons from genomics. *Trends in Genetics.* 2012, Vol. 28, 6, pp. 276-284.

**Bulger M, Groudine M. 2011.** Functional and mechanistic diversity of distal transcription enhancers. *Cell.* 2011, Vol. 144, 3, pp. 327-339.

**Burda P, Laslo P, Stopka T. 2010.** The role of PU.1 and GATA-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia.* 2010, Vol. 24, 7, pp. 1249-1257.

**Carlsson R, Persson C, Leanderson T. 2003.** SPI-C, a PU-box binding ETS protein expressed temporarily during B-cell development and in macrophages, contains an acidic transactivation domain located to the N-terminus. *Molecular immunology.* 2003, Vol. 39, 16, pp. 1035-1043.

**Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute J, Shao W, Hestermann EV, Geistlinger TR, Fox EA, Silver PA, Brown M. 2005.** Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell.* 2005, Vol. 122, 1, pp. 33-43.

**Choukrallah MA, Matthias P. 2014.** The Interplay between Chromatin and Transcription Factor Networks during B Cell Development: Who Pulls the Trigger First? *Frontiers in Immunology.* 2014, Vol. 5, 156.

**Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B, Kathiria A, Cho SW, Mumbach MR, Carter AC, Kasowski M, Orloff LA,

**Risca V, Kundaje A, Khavari PA, Montine TJ, Greenleaf WJ, Chang HY. 2017.** An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods.* 2017, Vol. 14, 10, pp. 959-962.

**de la Rica L, Rodríguez-Ubreva J, García M, Islam AB, Urquiza JM, Hernando H, Christensen J, Helin K, Gómez-Vaquero C, Ballestar E. 2013.** PU.1 target genes undergo Tet2-coupled demethylation and DNMT3b-mediated methylation in monocyte-to-osteoclast differentiation. *Genome Biology.* 2013, Vol. 14, 9.

**Deaton AM, Bird A. 2011.** CpG islands and the regulation of transcription. *Genes & Development.* 2011, Vol. 25, 10, pp. 1010-1022.

**Delva L, Gallais I, Guillouf C, Denis N, Orvain C, Moreau-Gachelin F. 2004.** Multiple functional domains of the oncoproteins Spi-1/PU.1 and TLS are involved in their opposite splicing effects in erythroleukemic cells. *Oncogene.* 2004, Vol. 23, 25, pp. 4389-4399.

**Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schübeler D. 2015.** Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature.* 2015, Vol. 528, 7583, pp. 575-590.

**Donaldson LW, Petersen JM, Graves BJ, McIntosh LP. 1996.** Solution structure of the ETS domain from murine Ets-1: a winged helix-turn-helix DNA binding motif. *The EMBO Journal.* 1996, Vol. 15, 1, pp. 125-134.

**Dulac C. 2010.** Brain function and chromatin plasticity. *Nature.* 2010, Vol. 465, 7229, pp. 728-735.

**Eeckhoute J, Métivier R, Salbert G. 2009.** Defining specificity of transcription factor regulatory activities. *Journal of Cell Science.* 2009, Vol. 122, 22, pp. 4027-4034.

**Erdel F, Krug J, Längst G, Rippe K. 2011.** Targeting chromatin remodelers: signals and search mechanisms. *Biochimica et biophysica acta.* 2011, Vol. 1809, 9, pp. 497-508.

**FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, et al. and Hayashizaki Y. 2014.** A promoter-level mammalian expression atlas. *Nature.* 2014, Vol. 507, 7493, pp. 462-470.

**FANTOM Consortium, Suzuki H, Forrest ARR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, Lassmann T, Ravasi T, et al. and Hayashizaki Y. 2009.** The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genetics.* 2009, Vol. 41, 5, pp. 553-562.

**Fedorov O, Castex J, Tallant C, Owen DR, Martin S, Aldeghi M, Monteiro O, Filippakopoulos P, Picaud S, Trzupek JD, Gerstenberger BS, Bountra C, Willmann D, Wells C1, Philpott M, Rogers C, Biggin PC, et al. and Müller S. 2015.** Selective targeting of the BRG/PB1 bromodomains impairs embryonic and trophoblast stem cell maintenance. *Science Advances.* 2015, Vol. 10, 1.

**Feng R, Desbordes SC, Xie H, Tillo ES, Pixley F, Stanley ER, Graf, T. 2008.** PU.1 and C/EBPalpha/beta convert fibroblasts into macrophage-like cells. *Proceedings of the National Academy of Sciences of the United States of America.* 2008, Vol. 105, 16, pp. 6057-6062 .

**Foos GE, Fischer KM, Savage J, Reddy V and Torbett BE. 2006.** Identification of PU.1 Target Genes That Are Dependent on Specific Functional Domains of the Transcription Factor PU.1. *Blood.* 2006, Vol. 108, p. 1174.

**Friedman AD. 2007.** Transcriptional control of granulocyte and monocyte development. *Oncogene.* 2007, Vol. 26, 47, pp. 6816–6828.

**Garber M, Yosef N, Goren A, Raychowdhury R, Thielke A, Guttman M, Robinson J, Minie B, Chevrier N, Itzhaki Z, Blecher-Gonen R, Bornstein C, Amann-Zalcenstein D, Weiner A, Friedrich D, Meldrim J, Ram O, Cheng C, Gnirke A, Fisher S, et al. and Amit I. 2012.** A High-Throughput Chromatin Immunoprecipitation Approach Reveals Principles of Dynamic Gene Regulation in Mammals. *Molecular Cell.* 2012, Vol. 47, 5, pp. 810-822.

**Gerstenberger BS, Trzupek JD, Tallant C, Fedorov O, Filippakopoulos P, Brennan PE, Fedele V, Martin S, Picaud S, Rogers C, Parikh M, Taylor A, Samas B, O'Mahony A, Berg E, Pallares G, Torrey AD, Treiber DK, et al. and Owen DR. 2016.** Identification of a Chemical Probe for Family VIII Bromodomains through Optimization of a Fragment Hit. *Journal of medicinal chemistry.* 2016, Vol. 59, 10, pp. 4800-4811.

**Gertz J, Savic D, Varley KE, Partridge EC, Safi A, Jain P, Cooper GM, Reddy TE, Crawford GE, Myers RM. 2013.** Distinct properties of cell-type-specific and shared transcription factor binding sites. *Molecular Cell.* 2013, Vol. 52, 1, pp. 25-36.

**Ghisletti S, Barozzi I, Mietton F, Polletti S, De Santa F, Venturini E, Gregory L, Lonie L, Chew A, Wei CL, Ragoussis J, Natoli G. 2010.** Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages program in macropha. *Immunity.* 2010, Vol. 32, 3, pp. 317-328.

**Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA 3rd, Smith HO. 2009.** Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods.* 2009, Vol. 6, 5, pp. 343-345.

**Giorgetti L, Siggers T, Tiana G, Caprara G, Notarbartolo S, Corona T, Pasparakis M, Milani P, Bulyk ML, Natoli G. 2010.** Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs. *Molecular Cell.* 2010, Vol. 37, 3, pp. 418-428.

**Gottesfeld JM, Carey MF. 2018.** Introduction to Thematic Minireview Series: Chromatin and Transcription. *The Journal of Biological Chemistry.* 10.1074, 2018.

**Graw RG Jr, Herzig GP, Eisel RJ, Perry S. 1971.** Leukocyte and platelet collection from normal donors with the continuous flow blood cell separator. *Transfusion.* 1971, Vol. 11, 2, pp. 94-101.

**Hardison RC, Taylor J. 2012.** Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics.* 2012, Vol. 13, 7, pp. 469-483.

**Hassan AH, Neely KE, Vignali M, Reese JC, Workman JL. 2001.** Promoter targeting of chromatin-modifying complexes. *Frontiers in Bioscience.* 2001, Vol. 6, pp. 1054-1064.

**Hassan AH, Prochasson P, Neely KE, Galasinski SC, Chandy M, Carrozza MJ, Workman JL. 2002.** Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. *Cell.* 2002, Vol. 111, 3, pp. 369-379.

**He Y, Fang J, Taatjes DJ, Nogales E. 2013.** Structural visualization of key steps in human transcription initiation. *Nature.* 2013, Vol. 495, 7442, pp. 481-486.

**Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010.** Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell.* 2010, Vol. 38, 4, pp. 576-589.

**Heinz S, Romanoski CE, Benner C, Glass CK. 2015.** The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology.* 2015, Vol. 16, 3, pp. 144-154.

**Hergeth SP, Schneider R. 2015.** The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle. *EMBO reports.* 2015, Vol. 16, pp. 1439-1453.

**Huang J, Liu X, Li D, Shao Z, Cao H, Zhang Y, Trompouki E, Bowman TV, Zon LI, Yuan GC, Orkin SH, Xu J. 2016.** Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. *Developmental Cell.* 2016, Vol. 36, 1, pp. 9-23.

**Iwasaki H, Somoza C, Shigematsu H, Duprez EA, Iwasaki-Arai J, Mizuno S, Arinobu Y, Geary K, Zhang P, Dayaram T, Fenyus ML, Elf S, Chan S, Kastner P, Huettner CS, Murray R, Tenen DG, Akashi K. 2005.** Distinctive and indispensable roles of PU.1 in maintenance of hematopoietic stem cells and their differentiation. *Blood.* 2005, Vol. 106, 5, pp. 1590-1600.

**John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA. 2011.** Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics.* 2011, Vol. 43, 3, pp. 264-268.

**Johnson WD Jr, Mei B, Cohn ZA. 1977.** The separation, long-term cultivation, and maturation of the human monocyte. *Journal of experimental medicine.* 1977, Vol. 146, 6, pp. 1613-1626.

**Kadoch C, Crabtree GR. 2015.** Mammalian SWI/SNF chromatin remodeling complexes and cancer: Mechanistic insights gained from human genomics. *Science Advances.* 2015, Vol. 1, 5.

**Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI, Yurovsky A, Lappalainen T, Romano-Palumbo L, Planchon A, Bielser D, Bryois J, Padioleau I, et al. and Dermitzakis ET. 2013.** Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science.* 2013, Vol. 342, 6159, pp. 744-747.

**Kim TK, Eberwine JH. 2010.** Mammalian cell transfection: the present and the future. *Analytical and bioanalytical chemistry.* 2010, Vol. 397, 8, pp. 3173-3178.

**Kim TK, Shiekhattar R. 2015.** Architectural and Functional Commonalities between Enhancers and Promoters. *Cell.* 2015, Vol. 162, 5, pp. 948-959.

**Koschmieder S, Rosenbauer F, Steidl U, Owens BM, Tenen DG. 2005.** Role of transcription factors C/EBPalpha and PU.1 in normal hematopoiesis and leukemia. *International Journal of Hematology.* 2005, Vol. 81, 5, pp. 368-377.

**Kwok JC, Perdomo J, Chong BH. 2007.** Identification of a monopartite sequence in PU.1 essential for nuclear import, DNA-binding and transcription of myeloid-specific genes. *Journal of cellular biochemistry.* 2007, Vol. 101, 6, pp. 1456-1474.

**Laiosa CV, Stadtfeld M, Xie H, de Andres-Aguayo L, Graf T. 2006.** Reprogramming of committed T cell progenitors to macrophages and dendritic cells by C/EBP alpha and PU.1 transcription factors. *Immunity.* 2006, Vol. 25, 5, pp. 731-744.

**Laird PW. 2005.** Cancer Epigenetics. *Human Molecular Genetics.* 2005, Vol. 14, 1, pp. 65-76.

**Lawrence M, Daujat S, Schneider R. 2016.** Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics.* 2016, Vol. 32, 1, pp. 42-56.

**Lee BK, Bhinge AA, Battenhouse A, McDaniell RM, Liu Z, Song L, Ni Y, Birney E, Lieb JD, Furey TS, Crawford GE, Iyer VR. 2012.** Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome research.* 2012, Vol. 22, 1, pp. 9-24.

**Lefterova MI, Steger DJ, Zhuo D, Qatanani M, Mullican SE, Tuteja G, Manduch E, Grant GR, Lazar MA. 2010.** Cell-Specific Determinants of Peroxisome Proliferator-Activated Receptor γ Function in Adipocytes and Macrophages. *Molecular and Cellular Biology.* 2010, Vol. 30, 9, pp. 2078-2089.

**Li E, Beard C, Jaenisch R. 1993.** ROLE FOR DNA METHYLATION IN GENOMIC IMPRINTING. *Nature.* 1993, Vol. 366, 6453, pp. 362-365.

**Liu Y, Chen S, Wang S, Soares F, Fischer M, Meng F, Du Z, Lin C, Meyer C, DeCaprio JA, Brown M, Liu XS, He HH. 2017.** Transcriptional landscape of the human cell cycle. *Proceedings of the National Academy of Scienes of the USA.* 2017, Vol. 114, 13, pp. 3473-2478.

**Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M. 2008.** FoxA1 translates epigenetic signatures into enhancer driven lineage-specific transcription. *Cell.* 2008, Vol. 132, 6, pp. 958-970.

**MacQuarrie KL, Fong AP, Morse RH, Tapscott SJ. 2011.** Genome-wide transcription factor binding: beyond direct target regulation. *Trends in Genetics.* 2011, Vol. 27, 4, pp. 141–148.

**Mayran A, Drouin J. 2018.** Pioneer transcription factors shape the epigenetic landscape. *The Journal of Biological Chemistry.* 10.1074, 2018.

**McKercher SR, Torbett BE, Anderson KL, Henkel GW, Vestal DJ, Baribault H, Klemsz M, Feeney AJ, Wu GE, Paige CJ, Maki RA. 1996.** Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *EMBO Journal.* 1996, Vol. 15, 20, pp. 5647-5658.

**Meierhoff G1, Krause SW, Andreesen R. 1998.** Comparative analysis of dendritic cells derived from blood monocytes or CD34+ hematopoietic progenitor cells. *Immunobiology.* 1998, Vol. 198, 5, pp. 501-513.

**Michel BC, Kadoch C. 2017.** A SMARCD2-containing mSWI/SNF complex is required for granulopoiesis. *Nature Genetics.* 2017, Vol. 49, 5, pp. 655-657.

**Moore SPG, Toomire KJ, Strauss PR. 2013.** DNA modifications repaired by base excision repair are epigenetic. *DNA Repair.* 2013, Vol. 12, 12, pp. 1152-1158.

**Mullen AC, Orlando DA, Newman JJ, Lovén J, Kumar RM, Bilodeau S, Reddy J, Guenther MG, DeKoter R, Young RA. 2011.** Master Transcription Factors Determine Cell-Type-Specific Responses to TGF-β Signaling. *Cell.* 2011, Vol. 147, 3, pp. 565-576.

**Mullen AC, Orlando DA, Newman JJ, Lovén J, Kumar RM, Bilodeau S, Reddy J, Guenther MG, DeKoter RP, Young RA. 2011.** Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell.* 2011, Vol. 147, 3, pp. 565-576.

**Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. 1986.** Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology.* 1986, Vol. 51, pp. 263-273.

**Mund C, Hackanson B, Stresemann C, Lübbert M, Lyko F. 2005.** Characterization of DNA demethylation effects induced by 5-Aza-2'-deoxycytidine in patients with myelodysplastic syndrome. *Cancer Research.* 2005, Vol. 65, 16, pp. 70856-7090.

**Narlikar GJ, Fan HY, Kingston RE. 2002.** Cooperation between complexes that regulate chromatin structure and transcription. *Cell.* 2002, Vol. 108, 4, pp. 475-487.

**Natoli G. 2010.** Maintaining cell identity through global control of genomic organization. *Immunity.* 2010, Vol. 33, 1, pp. 12-24.

**Neely KE, Hassan AH, Wallberg AE, Steger DJ, Cairns BR, Wright AP, Workman JL. 1999.** Activation domain-mediated targeting of the SWI/SNF complex to promoters stimulates transcription from nucleosome arrays. *Molecular Cell.* 1999, Vol. 4, 4, pp. 649-655.

**Nerlov C, Querfurth E, Kulessa H, Graf T. 2000.** GATA-1 interacts with the myeloid PU.1 transcription factor and represses PU.1-dependent transcription. *Blood.* 2000, Vol. 95, 8, pp. 2543-2551.

**Nguyen TT, Savory JG, Brooke-Bisschop T, Ringuette R, Foley T, Hess BL, Mulatz KJ, Trinkle-Mulcahy L, Lohnes D. 2017.** Cdx2 Regulates Gene Expression through Recruitment of Brg1-associated Switch-Sucrose Non-fermentable (SWI-SNF) Chromatin Remodeling Activity. *The Journal of biological chemistry.* 2017, Vol. 292, 8, pp. 3389-3399.

**Nishiyama C, Nishiyama M, Ito T, Masaki S, Masuoka N, Yamane H, Kitamura T, Ogawa H, Okumura K. 2004.** Functional analysis of PU.1 domains in monocyte-specific gene regulation. *FEBS letters.* 2004, Vol. 561, 1-3, pp. 63-68.

**Oikawa T, Yamada T. 2003.** Molecular biology of the Ets family of transcription factors. *Gene.* 2003, Vol. 303, pp. 11-34.

**Ong CT, Corces VG. 2011.** Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics.* 2011, Vol. 12, 4, pp. 283-293.

**Palii CG, Perez-Iratxeta C, Yao Z, Cao Z, Dai F, Davison J, Atkins H, Allan D, Dilworth FJ, Gentleman R, Tapscott SJ, Brand M. 2011.** Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. *The EMBO Journal.* 2011, Vol. 30, 3, pp. 494-509.

**Peterson CL, Workman JL. 2000.** Promoter targeting and chromatin remodeling by the SWI/SNF complex. *Current opinion in genetics & development.* 2000, Vol. 10, 2, pp. 187-192.

**Pham TH, Benner C, Lichtinger M, Schwarzfischer L, Hu Y, Andreesen R, Chen W, Rehli M. 2012.** Dynamic epigenetic enhancer signatures reveal key transcription factors. *Blood.* 2012, Vol. 119, 24, pp. 161-171.

**Pham TH, Minderjahn J, Schmidl C, Hoffmeister H, Schmidhofer S, Chen W, Längst G, Benner C, Rehli M. 2013.** Mechanisms of in vivo binding site selection of the hematopoietic master transcription factor PU.1. *Nucleic Acids Research.* 2013, Vol. 41, 13, pp. 6391-6402.

**Pinello L, Canver MC, Hoban MD, Orkin SH, Kohn D, Bauer DE, Yuan GC. 2016.** Analyzing CRISPR genome-editing experiments with CRISPResso. *Nature Biotechnology.* 2016, Vol. 34, 7, pp. 695-697.

**Poon GMK, Kim HM. 2017.** Signatures of DNA target selectivity by ETS transcription factors. *Transcription.* 2017, Vol. 8, 3, pp. 193-203.

**Priam P, Krasteva V, Rousseau P, D'Angelo G, Gaboury L, Sauvageau G, Lessard JA. 2017.** SMARCD2 subunit of SWI/SNF chromatin-remodeling complexes mediates granulopoiesis through a CEBPε dependent mechanism. *Nature Genetics.* 2017, Vol. 49, 5, pp. 753-764.

**Qian F, Deng J, Lee YG, Zhu J, Karpurapu M, Chung S, Zheng JN, Xiao L, Park GY, Christman JW. 2015.** The transcription factor PU.1 promotes alternative macrophage polarization and asthmatic airway inflammation. *Journal of molecular cell biology.* 2015, Vol. 7, 6, pp. 557-567.

**Rao SSP, Huang SC, Glenn St Hilaire B, Engreitz JM, Perez EM, Kieffer-Kwon KR, Sanborn AL, Johnstone SE, Bascom GD, Bochkov ID, Huang X, Shamim MS, Shin J, Turner D, Ye Z, Omer AD, Robinson JT, Schlick T, Bernstein BE, Casellas R, Lander ES, Aiden EL. 2017.** Cohesin Loss Eliminates All Loop Domains. *Cell.* 2017, Vol. 171, 2, pp. 305-320.

**Rishi V, Bhattacharya P, Chatterjee R, Rozenberg J, Zhao J, Glass K, Fitzgerald P, Vinson C. 2010.** CpG methylation of half-CRE sequences creates C/EBPalpha binding sites that activate some tissue-specific genes. *Proceedings of the Natonial Academy of Sciences of the United States of America.* 2010, Vol. 107, 47, pp. 20311-20316.

**Robinson MD, McCarthy DJ, Smyth GK. 2010.** edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010, Vol. 26, 1, pp. 139-140.

**Ross IL, Yue X, Ostrowski MC, Hume DA. 1998.** Interaction between PU.1 and another Ets family transcription factor promotes macrophage-specific Basal transcription initiation. *Journal of Biological Chemistry.* 1998, Vol. 273, 12, pp. 6662-6669.

**Roux KJ, Kim DI, Burke B. 2013.** BioID: a screen for protein-protein interactions. *Current protocols in protein science.* 2013, Vol. 74.

**Sanders JS, Mason CE. 2016.** The Newly Emerging View of the Genome. *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry.* 2016.

**Sanderson RJ, Shepperdson RT, Vatter AE, Talmage DW. 1977.** Isolation and enumeration of peripheral blood monocytes. *Journal of immunology.* 1977, Vol. 118, 4, pp. 1409-1414.

**Scott EW, Simon MC, Anastasi J, Singh H. 1994.** Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science.* 1994, Vol. 265, 5178, pp. 1573-1577.

**Seki M, Kimura S, Isobe T, Yoshida K, Ueno H, Nakajima-Takagi Y, Wang C, Lin L, Kon A, Suzuki H, Shiozawa Y, Kataoka K, Fujii Y, Shiraishi Y, Chiba K, Tanaka H, Shimamura T, Masuda K, Kawamoto H, Ohki K, Kato M, Arakawa Y, Koh K, et al. and Takita J. 2017.** Recurrent SPI1 (PU.1) fusions in high-risk pediatric T cell acute lymphoblastic leukemia. *Nature Genetics.* 2017, Vol. 49, 8, pp. 1274-1281.

**Shibata Y, Berclaz PY, Chroneos ZC, Yoshida M, Whitsett JA, Trapnell BC. 2001.** GM-CSF Regulates Alveolar Macrophage Differentiation and Innate Immunity in the Lung through PU.1. *Immunity.* 2001, Vol. 15, 4, pp. 557-567.

**Shlyueva D, Stampfel G, Stark A. 2014.** Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics.* 2014, Vol. 15, 4, pp. 272-286.

**Siersbæk R, Nielsen R, John S, Sung M, Baek S, Loft A, Hager GL, Mandrup S. 2011.** Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. *EMBO Journal.* 2011, Vol. 30, 8, pp. 1459-1472.

**Simon JA, Kingston RE. 2009.** Mechanisms of Polycomb gene silencing: knowns and unknowns. *Nature Reviews Molecular Cell Biology.* 2009, Vol. 10, 10, pp. 697-708.

**Sizemore GM, Pitarresi JR, Balakrishnan S, Ostrowski MC. 2017.** The ETS family of oncogenic transcription factors in solid tumours. *Nature Reviews Cancer.* 2017, Vol. 17, 6, pp. 337-3351.

**Stachura DL, Chou ST, Weiss MJ. 2006.** Early block to erythromegakaryocytic development conferred by loss of transcription factor GATA-1. *Blood.* 2006, Vol. 107, 1, pp. 87-97.

**Stephens DC, Poon GM. 2016.** Differential sensitivity to methylated DNA by ETS-family transcription factors is intrinsically encoded in their DNA-binding domains. *Nucleic Acids Research.* 2016, Vol. 44, 18, pp. 8671-8681.

**Sterner DE, Berger SL. 2000.** Acetylation of histones and transcription-related factors. *Microbiology and Molecular Biology Reviews.* 2000, Vol. 64, 2, pp. 435-459.

**Suzuki T, Shimizu Y, Furuhata E, Maeda S, Kishima M, Nishimura H, Enomoto S, Hayashizaki Y, Suzuki H. 2017.** RUNX1 regulates site specificity of DNA demethylation by recruitment of DNA demethylation machineries in hematopoietic cells. *Blood advances.* 2017, Vol. 1, 20, pp. 1699-1711.

**Swinstead EE, Miranda TB, Paakinaho V, Baek S, Goldstein I, Hawkins M, Karpova TS, Ball D, Mazza D, Lavis LD, Grimm JB, Morisaki T, Grøntved L, Presman DM, Hager GL. 2016.** Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. *Cell.* 2016, Vol. 165, 3, pp. 593-605.

**Swinstead EE, Paakinaho V, Presman DM, Hager GL. 2016.** Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: A new perspective: Multiple transcription factors can effect chromatin pioneer functions through dynamic interactions with ATP-dependent chromatin remodeling factors. *BioEssays: news and reviews in molecular, cellular and developmental biology.* 2016, Vol. 38, 11.

**Tenen DG. 2003.** Disruption of differentiation in human cancer: AML shows the way. *Nature Reviews Cancer.* 2003, Vol. 3, 2, pp. 89-101.

**Towbin H, Staehelin T, Gordon J. 1979.** Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the National Academy of Sciences of the United States of America.* 1979, Vol. 76, 9, pp. 4350-4354.

**Tsompana M, Buck MJ. 2014.** Chromatin accessibility: a window into the genome. *Epigenetics and Chromatin.* 2014, Vol. 7, 33.

**Ungerbäck J, Hosokawa H, Wang X, Strid T, Williams BA, Sigvardsson M, Rothenberg EV. 2018.** Pioneering, chromatin remodeling, and epigenetic constraint in early T-cell gene regulation by SPI1 (PU.1). *Genome research.* 2018.

**Van Tendeloo VF, Ponsaerts P, Lardon F, Nijs G, Lenjou M, Van Broeckhoven C, Van Bockstaele DR, Berneman ZN. 2001.** Highly efficient gene delivery by mRNA electroporation in human hematopoietic cells: superiority to lipofection and passive pulsing of mRNA and to electroporation of plasmid cDNA for tumor antigen loading of dendritic cells. *Blood.* 2001, Vol. 98, 1, pp. 49-56.

**Verhelst J, Parthoens E, Schepens B, Fiers W, Saelens X. 2012.** Interferon-inducible protein Mx1 inhibits influenza virus by interfering with functional viral ribonucleoprotein complex assembly. *Journal of virology.* 2012, Vol. 86, 24, pp. 13445-13455.

**Voss TC, Hager GL. 2014.** Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics.* 2014, Vol. 15, 2, pp. 69-81.

**Walsh JC, DeKoter RP, Lee HJ, Smith ED, Lancki DW, Gurish MF, Friend DS, Stevens RL, Anastasi J, Singh H. 2002.** Cooperative and antagonistic interplay between PU.1 and GATA-2 in the specification of myeloid cell fates. *Immunity.* 2002, Vol. 17, 5, pp. 665-676.

**Wei GH, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, Yan J, Talukder S, Turunen M, Taipale, M., Stunnenberg HG, Ukkonen E, Hughes TR, Bulyk ML, Taipale J. 2010.** Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO Journal.* 2010, Vol. 29, 13, pp. 2147-2160.

**Whitfield TW, Wang J, Collins PJ, Partridge EC, Aldred SF, Trinklein ND, Myers RM, Weng Z. 2012.** Functional analysis of transcription factor binding sites in human promoters. *Genome Biology.* 2012, Vol. 13, 9.

**Willis SN, Tellier J, Liao Y, Trezise S, Light A, O'Donnell K, Garrett-Sinha LA, Shi W, Tarlinton DM, Nutt SL. 2017.** Environmental sensing by mature B cells is controlled by the transcription factors PU.1 and SpiB. *Nature Communications.* 2017, Vol. 8, 1.

**Wilson BG, Roberts CW. 2011.** SWI/SNF nucleosome remodellers and cancer. *Nature Reviews Cancer.* 2011, Vol. 11, 7, pp. 481-492.

**Wilson NK, Foster SD, Wang X, Knezevic K, Schütte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E, Pimanda JE, de Bruijn MF, Göttgens B. 2010.** Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell stem cell.* 2010, Vol. 7, 4, pp. 532-544.

**Witzel M, Petersheim D, Fan Y, Bahrami E, Racek T, Rohlfs M, Puchałka J, Mertes C, Gagneur J, Ziegenhain C, Enard W, Stray-Pedersen A, Arkwright PD, Abboud MR, Pazhakh V, Lieschke GJ, Krawitz PM, Dahlhoff M, et al. and Klein C. 2017.** Chromatin-remodeling factor SMARCD2 regulates transcriptional networks controlling differentiation of neutrophil granulocytes. *Nature Genetics.* 2017, Vol. 49, 5, pp. 742-752.

**Yee AA, Yin P, Siderovski DP, Mak TW, Litchfield DW, Arrowsmith CH. 1998.** Cooperative interaction between the DNA-binding domains of PU.1 and IRF4. *Journal of cellular biochemistry.* 1998, Vol. 279, 5, pp. 1075-1083.

**Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, Nitta KR, Taipale M, Popov A, Ginno PA, Domcke S, Yan J, Schübeler D, Vinson C, Taipale J. 2017.** Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science.* 2017, Vol. 356, 6337.

**Zaret KS, Lerner J, Iwafuchi-Doi M. 2016.** Chromatin Scanning by Dynamic Binding of Pioneer Factors. *Molecular Cell.* 2016, Vol. 62, 5, pp. 665-667.

**Zarnegar MA, Chen J, Rothenberg EV. 2010.** Cell-type-specific activation and repression of PU.1 by a complex of discrete, functionally specialized cis-regulatory elements. *Molecular and cellular biology.* 2010, Vol. 30, 20, pp. 4922-4939.

**Zehavi Y, Kedmi A, Ideses D, Juven-Gershon T. 2015.** TRF2: TRansForming the view of general transcription factors. *Trancription.* 2015, Vol. 6, 1, pp. 1-6.

**Zenke M, Steinlein P, Wagner E, Cotten M, Beug H, Birnstiel ML. 1990.** Receptor-mediated endocytosis of transferrin-polycation conjugates: an efficient way to introduce DNA into hematopoietic cells. *Proceedings of the National Academy of Sciences of the USA.* 1990, Vol. 87, 10, pp. 3655-3659.

**Zhu H, Wang G, Qian J. 2016.** Transcription factors as readers and effectors of DNA methylation. *Nature Reviews Genetics.* 2016, Vol. 17, 9, pp. 551-565.

**Zillner, K, Jerabek-Willemsen, M, Duhr, S, Braun, D, Längst, G and Baaske, P. 2012.** Microscale Thermophoresis as a Sensitive Method to Quantify Protein: Nucleic Acid Interactions in Solution. *Functional Genomics: Methods in Molecular Biology.* 2012, Vol. 815, pp. 241-252.

# 9 ABBREVIATIONS

| | |
|---|---|
| 5mC | 5’-methyl cytosine |
| aa | amino acid |
| ac | acetyl |
| ALL | acute lymphoblastic leukemia |
| AML | acute myeloid leukemia |
| Amp | ampicillin |
| AP | alkaline phosphatase |
| approx. | approximately |
| APS | ammonium persulfate |
| ARID1A/B | AT-rich interactive domain-containing protein 1A/B |
| as | antisense |
| ATACseq | assay for transposase-accessible chromatin sequencing |
| ATP | adenosine triphosphate |
| ATPase | ATP triphosphatase |
| BAF | BRG1-associated factor |
| BER | base excision repair |
| BioID | proximity-dependent biotin identification |
| BM | bone marrow |
| bp | base pair |
| BRD7 | bromodomain-containing 7 |
| BRE | TFIIB recognition element |
| BRG1 | BRM/SWI2-related gene 1 |
| BRM | brahma homologue |
| BSA | bovine serum albumin |
| BSF | biomedical sequencing facility |
| BSS | balanced salt solution |
| bZIP | basic leucine zipper |
| C | cytosine |
| CAGE | cap analysis of gene expression |
| CD | cluster of differentiation |
| cDNA | complementary DNA |
| C/EBPα/β | CCCAT/ enhancer binding protein α/β |

| | |
|---|---|
| ChIP | chromatin immunoprecipitation |
| CGI | cytosine guanine dinucleotide island |
| c-Jun (AP-1) | Jun-proto-oncogene |
| CMP | common myeloid progenitor |
| CNV | copy number variation |
| CpG | cytosine guanine dinucleotide |
| CPM | counts per million |
| CTCF | CCCTC-binding factor |
| C-terminal | carboxy-terminal |
| DAC | decitabine |
| DBD | DNA binding domain |
| DC | dendritic cell |
| DCE | downstream core element |
| dd | double-destilled |
| DEG | differential expressed genes |
| DGE | differential gene expression |
| DMSO | dimethyl sulfoxide |
| DNA | deoxyribonucleic acid |
| DNase | deoxyribonuclease |
| DNMT3B | DNA methyltransferases 3B |
| dNTP | deoxyribonucleotide triphosphate |
| DPE | downstream promotor element |
| dsDNA | double stranded DNA |
| DSG | disuccinimidyl glutarate |
| DTT | dithiothreitol |
| EB | elution buffer |
| EDTA | ethylenediaminetetraacetic acid |
| ESC | embryonic stem cell |
| EtOH | ethanol |
| ETS | E-26 transformation specific |
| FACS | fluorescence activated cell sorting |
| FCS | fetal calf serum |
| FDR | false discovery rate |
| FLI-1 | friend leukemia integration 1 transcription factor |
| FoxA1 | forkhead box A1 |

| | |
|---|---|
| G | guanine |
| GATA1/2/4 | GATA binding protein 1/2/4 |
| gDNA | genomic DNA |
| GFP | green fluorescent protein |
| GMP | granulocyte/monocyte progenitor |
| GO | gene ontology |
| h | hour |
| H1 | histone 1 |
| H2A/B | histone 2A/B |
| H3 | histone 3 |
| H4 | histone 4 |
| H3K27ac | acetylation of lysine 27 of histone 3 |
| H3K4me1/2/3 | mono-/di-/tri-methylation of lysine 27 of histone 3 |
| H3K27me3 | tri-methylation of lysine 27 of histone 3 |
| HAT | histone acetyl transferase |
| HDAC | histone deacetylase |
| Hepes | 4-(2-hydroxyethyl)-1-piperazine ethane sulphonic acid |
| HMT | histone methyltransferase |
| HOMER | hypergeometric optimization of motif enrichment |
| HRP | horseradish peroxidase |
| HSC | hematopoietic stem cell |
| HSF1 | heat-shock factor 1 |
| iBAQ | intensity-based absolute quantitation |
| ICSBP | interferon consensus sequence binding protein |
| IgG | immunoglobulin G |
| IGV | Integrative Genomics Viewer |
| IL | interleukin |
| Inr | initiator |
| INF-γ | interferon-gamma |
| IP | immunoprecipitation |
| IRF4/8 | interferon regulatory factor 4/8 |
| LMU | Ludwig-Maximilians-Universitaet München |
| lncRNA | long non-coding RNA |
| logFC | log-fold change |
| MAC | macrophage |

| | |
|---|---|
| MBD | methyl-CpG-binding domain |
| MBD4 | methyl-CpG-binding domain protein 4 |
| MCIp-seq | methyl-CpG-immunoprecipitation coupled to sequencing |
| mCpG | methyl-CpG |
| MDS | multidimensional scaling |
| me | methyl |
| MEP | megakaryocyte/erythroid progenitor |
| min | minute |
| MO | monocyte |
| mRNA | messenger RNA |
| MST | microscale thermophoresis |
| mut | mutant/mutated |
| NFAT | nuclear factor of activated T cells |
| NGS | next generation sequencing |
| NMR | nuclear magnetic resonance |
| NP-40 | Nonident P-40 |
| N-terminal | amino-terminal |
| OC | osteoclast |
| OCT4 | octamer-binding transcription factor 4 |
| o/n | over night |
| ORF | open reading frame |
| p | phosphate |
| PAA | poly acrylamide |
| PBAF | polybromo BRG1-associated factor |
| PB-MNC | peripheral blood mononuclear cell |
| PBS | phosphate buffered saline |
| PCA | prinicipal component analysis |
| PCR | polymerase chain reaction |
| PcG | Polycomb group |
| PEG | polyethylene glycol |
| PEST | peptide sequence rich in proline, glutamic acid, serine and threonine |
| PMA | phorbol 12-myristate 13-acetate |
| PMSF | phenylmethylsulfonyl fluoride |
| PTM | post-translational modification |

| | |
|---|---|
| PWM | position weight matrix |
| qPCR | quantitative PCR |
| rlog | regularized log |
| RNA | ribonucleic acid |
| RNA Pol II | RNA polymerase II |
| RNase | ribonuclease |
| RNAseq | RNA-sequencing |
| rpm | rounds per minute |
| RRE | Ras-responsive element |
| rRNA | ribosomal RNA |
| RT | room temperature |
| RUNX1 | runt-related transcription factor 1 |
| rxn | reaction |
| s | second |
| _s | sense |
| SAM | S-adenosylmethionine |
| SD | standard deviation |
| SDS | sodium dodecyl sulfate |
| seq | sequencing |
| SOC | super optimal broth with catabolite repression |
| Sp1 | specificity protein 1 |
| ssDNA | single stranded DNA |
| TAD | topologically associated domain or transactivation domain |
| TAE | tris-acetate-EDTA |
| TagDir | tag directory |
| TBP | TATA-box binding protein |
| TBSAseq | targeted bisulfite amplicon sequencing |
| TDG | thymine DNA glycosylase |
| TE | tris-EDTA |
| TEMED | N,N,N′,N′-Tetramethylethylenediamine |
| TET2 | ten-eleven translocation methyl cytosine dioxygenase 2 |
| TF | transcription factor |
| TFBS | transcription factor binding site |
| TM | Trademark |
| t-SNE | t-Distributed Stochastic Neighbor Embedding |

| | |
|---|---|
| TSS | transcription start site |
| U | unit |
| UCSC | University of California, Santa Cruz |
| UCSD | University of California, Santa Diego |
| VD3 | Vitamin $D_3$ |
| vs. | versus |
| WB | western blot |
| WT | wild type |
| ZfP | Zentrallabor für Proteinanalytik |

# 10 APPENDIX

## 10.1 Computational Analysis of NGS Data Sets

Computational analyses of NGS data were performed on a Linux Debian server with the indicated programs and versions (see section 3.13). Raw read files were delivered in fastq or bam format from the sequencing facilities and obtained as bcl files from the onsite NextSeq 550 (Illumnia). The general workflow for generating DNA libraries for NGS is outlined in section 4.2.7. Basic scripts used for mapping purposes were generated by Prof. Dr. Michael Rehli. Scripts regarding own NGS data sets were generated in collaboration with Prof. Dr. Michael Rehli.

### 10.1.1 Mapping of NGS Data

#### 10.1.1.1 Mapping of ChIPseq Data

ChIPseq data was generated by indexed (maximum 16 samples/lane) single 50 bp sequencing on a HiSeq 3000/4000 sequencer (Illumina) using the latest four-channel sequencing chemistry SBSv4 (Illumina). The general ChIP workflow is explained in section 4.2.6. Obtained raw read bam files were first copied to the Linux server, automatically converted to fastq and compressed to fastq.gz files using the following script. FastQC was also performed within this script.

```
#!/bin/bash
# script from Michael Rehli
# bash script to download and check md5 sum for bam-files produced at the BSF in Vienna
args=("$@")
if [ ${#args[@]} != 7 ] ; then
echo -e "\nMissing or too many arguments! \n"
echo "Correct usage:"
echo "getBSFconvertSE.sh <user name> <password> <flowcell ID> <lane number> <sample
name> <output subdirectory in misc/data/rawdata> <new sample name>"
echo "Example:"
echo -e "getBSFconvertSE.sh mrehli xyz BSF_0411_HNH5NBBXX 4 S_10_ChIP_DD_S33767
chromatin/ChIP/MOMACDC 3dDCs_DSG+FA_R2_EGR2_A \n"
exit
fi

# defining variables per default and operator input
WEBFOLDER="https://biomedical-sequencing.at/downloads/download/BSF_downloads/group_Michael_Reh
li"
USER=${args[0]}
PASSW=${args[1]}
EXPERIMENT=${args[2]}
LANE="${args[3]}"
SAMPLE=${args[4]}
RUN=$(echo $EXPERIMENT| cut -d'_' -f 2)
PREFIX=${args[2]}_${args[3]}_
BSFFOLDER="/misc/data/rawData/BSF/run${RUN}"
echo -e "\ngetBSFconvert.sh will download data from the Biomedical Sequencing Facility in
Vienna,"
echo -e "convert it into FASTQ format, run FASTQC.\n"
BSFBAM=$WEBFOLDER/${args[0]}/${args[2]}/$PREFIX${args[4]}.bam
BSFMD5=$WEBFOLDER/${args[0]}/${args[2]}/$PREFIX${args[4]}.bam.md5
OWNBAM=$BSFFOLDER/$PREFIX${args[4]}.bam
```

```
OWNMD5=$BSFFOLDER/$PREFIX${args[4]}.bam.md5
if [ ! -d "$BSFFOLDER" ]; then
mkdir "$BSFFOLDER"
echo -e "generating the folder $BSFFOLDER \n"
fi
TARGETFOLDER="/misc/data/rawData/${args[5]}"
TARGET=$TARGETFOLDER/${args[6]}.fastq
if [ ! -d "$TARGETFOLDER" ]; then
mkdir -p "${TARGETFOLDER}/FastQC"
echo -e "generating the folder $TARGETFOLDER \n"
fi

# downloading the .bam and .bam.md5 files
echo -e "downloading the bam-file for $SAMPLE\n"
curl -u $USER:$PASSW $BSFBAM -o $OWNBAM --insecure
echo -e "downloading the md5-file for $SAMPLE\n"
curl -u $USER:$PASSW $BSFMD5 -o $OWNMD5 --insecure
#comparing checksums created in Vienna and here
if [ $(md5sum "$OWNBAM" | cut -b-32) == $(cat "$OWNMD5") ]; then
echo "MD5 checksum matches"

# provided that the checksums match, proceed with converting into fastq, FastQC, and gzip
echo -e "\nconverting the bam-file into fastq format \n"
/usr/bin/java -jar /misc/software/ngs/picard/src/v2.17.3/picard.jar SamToFastq I=$OWNBAM
FASTQ=$TARGET
echo -e "running fastqc \n"
fastqc-0.11.7 -o "${TARGETFOLDER}/FastQC" $TARGET
echo -e "zipping the fastq file \n"
gzip $TARGET
else
echo "MD5 checksum different! Try downloading again!"
fi
```

Generated fastq data sets were mapped to reference genome (human, version hg19) using the following bash script, which implements parts of the HOMER suite (Heinz et al. 2010).

```
#!/bin/bash
# script from Michael Rehli
# bash script to map ChIPseq-like data (SE- only), generate tagDirectories, coverage bigWigs
and TDF files
# optional for input data: copy number alterations

#setting homer environment
DIR_PKG="/misc/software/ngs"
PATH_PERL=/misc/software/package/perl/perl-5.26.1/bin
PATH_SAMTOOLS=${DIR_PKG}/samtools/samtools-1.6/bin
PATH_HOMER=${DIR_PKG}/homer/v4.9/bin
PATH_R=/misc/software/package/RBioC/3.4.3/bin
export PATH=${PATH_R}:${PATH_PERL}:${PATH_SAMTOOLS}:${PATH_HOMER}:${PATH}
export PATH

# Defining the program versions (needs to be adjusted to the server/workstation used)
# rhskl17 version
BOWTIE="/misc/software/ngs/bowtie/bowtie2-2.3.4-linux-x86_64/bowtie2"
SAMTOOLS="/misc/software/ngs/samtools/samtools-1.6/bin/samtools"
SAMBAMBA="/misc/software/ngs/sambamba/v0.6.7/sambamba"
BEDTOOLS="/misc/software/ngs/bedtools/bedtools2-2.27.1/bin/bedtools"
PICARD="/usr/bin/java -jar /misc/software/ngs/picard/src/v2.17.3/picard.jar"
IGVTOOLS="/usr/bin/java -Xmx4g -Djava.awt.headless=true -jar
/misc/software/viewer/IGV/IGVTools_2.3.98/igvtools.jar"

# required input:
# -f <fastq or fastq.gz>
# -g <genome> available genomes include hg19, GRCh38, mm9, mm10
# -o <output subdirectory> only subdirectories above ChIP- or Input-folder are given
# -n <sample name> used for sam, bigwig and tag directory
# optional:
# -i indicates that the sample is input
# -m indicates that the sample is MCIp
# -t <number of threads>
# options specific for input samples:
# -c <read length for mappability> generates copy number profiles and annotates CNVs with
Cancer Genes from Cosmic Database (Available mappability data for 50, 75, and 100 bp)
# -s <XX/XY> gender XX female, XY male (default) - only necessary if -c is given
# -a <tissue> E (epithelial), L (leukaemia/lymphoma), M (mesenchymal), O (other) will output
Cancer Genes only for given tissues (only human!)

# Set Script Name variable
SCRIPT=`basename ${BASH_SOURCE[0]}`

# Set fonts for Help.
NORM=`tput sgr0`
BOLD=`tput bold`
REV=`tput smso`

# USAGE function
function USAGE {
echo -e \\n"Help documentation for ${BOLD}${SCRIPT}.${NORM}"\\n
echo -e "${REV}Basic usage:${NORM}\n${BOLD}$SCRIPT -g <genome> -f <fastq> -o
<subdirectory> -n <name> -t <threads>${NORM} "\\n
# -c <read length> -s <gender>"\\n
echo "Required:"
echo "${BOLD}-g${NORM} <genome> available genomes include hg19, GRCh38, mm9, mm10"
echo "${BOLD}-f${NORM} <fastq or fastq.gz> forward read"
echo "${BOLD}-d${NORM} <output subdirectory> only subdirectories above ChIP- or
Input-folder are given"
echo "${BOLD}-n${NORM} <sample name> used for bam, bigwig and tag directory, etc."
echo -e \\n"Optional:"
echo "${BOLD}-t${NORM} <number of threads> (optional/default 16)"
echo "${BOLD}-i${NORM} treat sample as genomic input (stored in DNA directory)"
echo "${BOLD}-m${NORM} treat sample as MCIp (stored in DNA directory, TagDir with -keepOne)"
echo "${BOLD}-c${NORM} <read length> will use FreeC to generate CNV profiles of Input
files. Read length refers to the mappability files available for a given genome,"
echo " which are found in /misc/software/ngs/genome/sequence/XXXX (XXXX = genome)"
echo "${BOLD}-s${NORM} <gender> accepts either XX or XY (default) - only necessary if -c
is given"
echo "${BOLD}-a${NORM} <tissue> E (epithelial), L (leukaemia/lymphoma), M (mesenchymal), O
(other) will output COSMIC Cancer Genes for given tissues ${BOLD}HUMAN ONLY!${NORM}"
echo -e \\n"Example: ${BOLD}$SCRIPT -g hg19 -f TFread.fastq.gz -d MOMACDC -n TF -t 32
${NORM}"\\n
exit 1
}

#Check the number of arguments. If none are passed, print help and exit.
NUMARGS=$#
if [ $NUMARGS -eq 0 ]; then
USAGE
fi

# Defining available genomes
GENOMES=(hg19 GRCh38 hg38 mm9 mm10)
TISSUETYPES=(E L M O)
KEEP=""

# Set defaults
GENDER="XY"
REF="ref"
THREADS=16
RUNFREEC=0
ANNOTATE=0
INPUT=0
MCIP=0
FOLDER="ChIP"
TOPFOLDER="chromatin"

# Parse command line options
while getopts :g:f:d:n:t:c:s:a:imh OPTIONS; do
case $OPTIONS in
g) #set option "g" - Bowtie will will only accept GRCh38 while HOMER needs hg38
GENOME=$OPTARG
HOMERGENOME=$OPTARG
;;
f) #set option "f"
FASTQR1=$OPTARG
;;
d) #set option "d"
DIRECTORY=$OPTARG
;;
n) #set option "n"
SAMPLENAME=$OPTARG
;;
t) #set option "t"
```

```
THREADS=$OPTARG
SAMTHREADS=$((THREADS-1))
;;
c) #set option "c"
MAPPAB=$OPTARG
RUNFREEC=1
;;
s) #set option "s"
GENDER=$OPTARG
;;
h) #show help
USAGE
;;
i) #set option "i"
echo -e \\n"Sample is genomic input."\\n
INPUT=1
FOLDER="Input"
TOPFOLDER="DNA"
;;
m) #set option "m"
echo -e \\n"Sample is MCIp."\\n
MCIP=1
FOLDER="MCIp"
TOPFOLDER="DNA"
KEEP="-keepOne "
;;
a) #set option "a"
ANNOTATE=1
TISSUE=$OPTARG
;;
\?) #unrecognized option - show USAGE
echo -e \\n"Option -${BOLD}$OPTARG${NORM} not allowed."
USAGE
;;
esac
done
shift $((OPTIND-1))

# Sanity checks
# excude using -m and -i together
if [ $INPUT == 1 ] && [ $MCIP == 1 ]; then
echo -e \\n"${BOLD}Use either -i or -m.{NORM}"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 2
fi

# Check whether Tissues are selected and correct
if [ $ANNOTATE == 1 ] && [ $INPUT == 1 ]; then
INTISSUE=$(echo ${TISSUETYPES[@]} | grep -o "$TISSUE" | wc -w)
if [ $INTISSUE == 0 ] ; then
echo -e \\n"Tissue -${BOLD}$TISSUE${NORM} not available (Options include E, M, L & O)."
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 2
fi
fi

# Check whether a reasonable number of cores is given
if [ $THREADS -lt 1 ] || [ "$THREADS" -gt 63 ]; then
echo "Number of threads should be in the range of 1 - 63!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 9
fi

# Check whether GENOME is available
INARRAY=$(echo ${GENOMES[@]} | grep -o "$GENOME" | wc -w)
if [ $INARRAY == 0 ] ; then
echo -e \\n"Genome -${BOLD}$GENOME${NORM} not available."
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 2
fi
GENOMEPATH=$GENOME
HOMERGENOME=$GENOME
BIGWIGGENOME=$GENOME

# path to chromosome sizes
CHROMSIZES="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/$GENOME.chrom.sizes"
# path to GEM files (mappability info), and chromosome sizes for FreeC
# needs two different one's due to a bug in FreeC
```

```
GEMPATH="/misc/software/ngs/genome/sequence/${GENOME}/mappability"
GEM="${GEMPATH}/${GENOME}_${MAPPAB}.mappability"
case $GENDER in
XY) REF="ref"
;;
XX) REF="ref-Y"
;;
esac
REDCHROMSIZES="/misc/software/ngs/freec/FREEC-11.0/data/${GENOME}.${REF}.chrom.sizes"
if [ $GENOME == "hg38" ] || [ $GENOME == "GRCh38" ] ; then
GENOME="hg38"
GENOMEPATH="GRCh38.PRI_p10"
BIGWIGGENOME="GRCh38"
HOMERGENOME="hg38"
CHROMSIZES="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/GRCh38.PRI_p10.chrom.sizes
"
GEMPATH="/misc/software/ngs/genome/sequence/GRCh38.PRI_p10/mappability"
GEM="${GEMPATH}/GRCh38_${MAPPAB}.mappability"
REDCHROMSIZES="/misc/software/ngs/freec/FREEC-11.0/data/${GENOME}.${REF}.chrom.sizes"
fi

#bowtie index
INDEX="$GENOMEPATH/$GENOME"

# if -c is given, check whether mappability track is available
if [ "$RUNFREEC" == 1 ] && [ ! -f "$GEM" ]; then
echo "Gem file for ${GENOME} not found!"
echo -e "Available gem files for ${GENOME} are:"
for entry in $GEMPATH/*
do
echo $entry
done
echo -e \\n"Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 3
fi

# Check whether fastq file is available and has the right extension
if [ ! -f "$FASTQR1" ]; then
echo "Fastq file (read 1) not found!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 3
fi
case "$FASTQR1" in
*.fastq.gz) ;;
*.fastq) ;;
*.fq.gz) ;;
*.fq) ;;
*) echo "Fastq file (read 1) has wrong extension (should be fastq, fq, fastq.gz
or fq.gz!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 4
;;
esac

# Check whether the subfolders are given and already there - if not ask whether path should
be made
if [ -z "$DIRECTORY" ]; then
echo "Subdirectory is missing!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 7
fi

# rhskl17 version
MAPPINGDIR="/misc/data/processedData/mapping/${TOPFOLDER}/${BIGWIGGENOME}/${FOLDER}/${DIRECTOR
Y}"
TAGDIRDIR="/misc/data/processedData/tagDir/${TOPFOLDER}/${BIGWIGGENOME}/${FOLDER}/${DIRECTORY}
"
BIGWIGDIR="/misc/data/processedData/bigWig/${TOPFOLDER}/${BIGWIGGENOME}/${FOLDER}/${DIRECTORY}
"
if [ ! -d "$MAPPINGDIR" ]; then #looks for the mapping output folder - if not
available, askes whether to make it
echo "$MAPPINGDIR"
read -p "Do you want to create this mapping folder (y/n)?" choice
case "$choice" in
y|Y ) mkdir -p "$MAPPINGDIR"
;;
n|N ) echo "Didn't create the folder and stopped!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
```

```
exit 8
;;
* ) echo "invalid response";;
esac
fi
if [ ! -d "$TAGDIRDIR" ]; then #looks for the tagDir output folder - if not
available, askes whether to make it
echo "$TAGDIRDIR"
read -p "Do you want to create this tagDirectory folder (y/n)?" choice
case "$choice" in
y|Y ) mkdir -p "$TAGDIRDIR"
;;
n|N ) echo "Didn't create the folder and stopped!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 8
;;
* ) echo "invalid response";;
esac
fi
if [ ! -d "$BIGWIGDIR" ]; then #looks for the bigWig output folder - if not
available, askes whether to make it
echo "$BIGWIGDIR"
read -p "Do you want to create this bigWig folder (y/n)?" choice
case "$choice" in
y|Y ) mkdir -p "$BIGWIGDIR"
;;
n|N ) echo "Didn't create the folder and stopped!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 8
;;
* ) echo "invalid response";;
esac
fi

# Check whether sample name is given
if [ -z "$SAMPLENAME" ]; then
echo "Sample name is missing!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 9
fi
SAMPLENAME="${SAMPLENAME// /_}" # replaces spaces with underscores in sample name, just in
case...
if [ ! -d "${MAPPINGDIR}/logs" ]; then #looks for the logs output folder - if not
available, makes it
mkdir "${MAPPINGDIR}/logs"
fi

# If everything passed, create a log file for all the commands that are executed during the
processing
LOGFILE="${MAPPINGDIR}/logs/${SAMPLENAME}.commands.txt"
touch $LOGFILE

# Function to append command lines into logfile and then execute
exe() { echo -e "\$ $@"\\n >> $LOGFILE; "$@" ; }

# Generating a random number file name for temporary files
TEMPFILES="/loctmp/$RANDOM.tmp"

# Running Bowtie2
echo -e \\n"Running Bowtie2 on $SAMPLENAME."\\n
exe $BOWTIE --very-sensitive -p $THREADS -x $INDEX --met-file
$MAPPINGDIR/logs/$SAMPLENAME.aln_metrics.txt -U $FASTQR1 2>
$MAPPINGDIR/logs/$SAMPLENAME.aln_rates.txt -S $TEMPFILES.sam
echo -e "Generating TagDir, BigWig, and TDF and finding Peaks."\\n

# Generating a HOMER tag directory
exe makeTagDirectory $TAGDIRDIR/$SAMPLENAME $TEMPFILES.sorted.sam $KEEP -genome $HOMERGENOME
-checkGC

# Making a UCSC browser track
exe makeUCSCfile $TAGDIRDIR/$SAMPLENAME -bigWig $CHROMSIZES -o $BIGWIGDIR/$SAMPLENAME.bigwig

# Making an IGV browser track
exe $IGVTOOLS count $TEMPFILES.sorted.sam $BIGWIGDIR/$SAMPLENAME.cov.tdf $CHROMSIZES

# Run FreeC if requested
if [ $RUNFREEC == 1 ]; then
if [ ! -d "${MAPPINGDIR}/CNVdata" ]; then #looks for the CNVdata output
```

```
folder - if not available, makes it
mkdir "${MAPPINGDIR}/CNVdata"
fi
if [ ! -d "${MAPPINGDIR}/CNVprofiles" ]; then #looks for the CNVprofile output
folder - if not available, makes it
mkdir "${MAPPINGDIR}/CNVprofiles"
fi
exe mv $TEMPFILES.sorted.sam "${MAPPINGDIR}/${SAMPLENAME}.sam"
case "$GENOME" in
hg19 )
COSMIC="/misc/software/ngs/freec/FREEC-11.0/data/COSMIC_Cancer_Gene_Census_GRCh37v84_2
018.tsv"
;;
hg38 )
COSMIC="/misc/software/ngs/freec/FREEC-11.0/data/COSMIC_Cancer_Gene_Census_GRCh38v84_2
018.tsv"
;;
esac
exe runFreeC.pl ${MAPPINGDIR}/CNVdata ${GENDER} ${GENOME} ${MAPPINGDIR} ${SAMPLENAME}
${REDCHROMSIZES} ${GEM}

# annotation of CNVs with COSMIC genes
exe FreeC2bed.pl ${MAPPINGDIR}/CNVdata/${SAMPLENAME}.sam_CNVs ${TEMPFILES}.CNV.bed
exe mv ${MAPPINGDIR}/CNVdata/${SAMPLENAME}.pdf ${MAPPINGDIR}/CNVprofiles
if [ $GENOME == "hg38" ] || [ $GENOME == "GRCh38" ] || [ $GENOME == "hg19" ]; then
case "$ANNOTATE" in
0 ) echo -e "\$ ConvertCOSMIC.pl ${COSMIC} >${TEMPFILES}.cosmic.bed"\\n >>
$LOGFILE
ConvertCOSMIC.pl ${COSMIC} > ${TEMPFILES}.cosmic.bed
echo -e "\$ ${BEDTOOLS} intersect -a ${TEMPFILES}.CNV.bed -b
${TEMPFILES}.cosmic.bed -wa -wb >${TEMPFILES}.inter.bed"\\n >> $LOGFILE
${BEDTOOLS} intersect -a ${TEMPFILES}.CNV.bed -b ${TEMPFILES}.cosmic.bed
-wa -wb >${TEMPFILES}.inter.bed
exe reformatCOSMICannotation.pl ${TEMPFILES}.inter.bed
${MAPPINGDIR}/CNVdata/${SAMPLENAME}.annotated_CNVs.txt
;;
1 ) echo -e "\$ ConvertCOSMIC.pl ${COSMIC} -tissues ${TISSUE}
>${TEMPFILES}.cosmic.bed"\\n >> $LOGFILE
ConvertCOSMIC.pl ${COSMIC} -tissues ${TISSUE} > ${TEMPFILES}.cosmic.bed
echo -e "\$ ${BEDTOOLS} intersect -a ${TEMPFILES}.CNV.bed -b
${TEMPFILES}.cosmic.bed -wa -wb >${TEMPFILES}.inter.bed"\\n >> $LOGFILE
${BEDTOOLS} intersect -a ${TEMPFILES}.CNV.bed -b ${TEMPFILES}.cosmic.bed
-wa -wb >${TEMPFILES}.inter.bed
exe reformatCOSMICannotation.pl ${TEMPFILES}.inter.bed
${MAPPINGDIR}/CNVdata/${SAMPLENAME}.${TISSUE}.annotated_CNVs.txt
;;
esac
fi

#convert sam to bam
exe $SAMTOOLS view -bS -@ $SAMTHREADS -o "${MAPPINGDIR}/${SAMPLENAME}.bam"
"${MAPPINGDIR}/${SAMPLENAME}.sam"
exe rm "${MAPPINGDIR}/${SAMPLENAME}.sam"
else

# just convert sam to bam
exe $SAMTOOLS view -bS -@ $SAMTHREADS -o "${MAPPINGDIR}/${SAMPLENAME}.bam"
$TEMPFILES.sorted.sam
fi

# removing tmp files
rm ${TEMPFILES}.*
```

### 10.1.1.2 Mapping of ATACseq Data

ATACseq data was generated by indexed (maximum 12 samples/lane) paired end 42 bp sequencing on the NextSeq 550 sequencer (Illumina) using the latest two-channel sequencing chemistry SBSv2 (Illumina). The general ATACseq workflow is outlined in section 4.2.8. Obtained raw read bcl files were converted to fastq format using bcl2fastq (Illumina). Sample sheets needed for demultiplexing were generated using the Illumina Experiment manager. Fastq data was mapped to the human reference genome (version hg19) using the following bash script, which implements parts of the HOMER suite (Heinz et al. 2010).

```
#!/bin/bash
# script from Michael Rehli
# bash script to map ATAC-seq data, generate tagDirectories, coverage bigWigs,
# and to find ATAC-seq peaks

#setting homer environment
DIR_PKG="/misc/software/ngs"
PATH_PERL=/misc/software/package/perl/perl-5.26.1/bin
PATH_SAMTOOLS=${DIR_PKG}/samtools/samtools-1.6/bin
PATH_HOMER=${DIR_PKG}/homer/v4.9/bin
PATH_R=/misc/software/package/RBioC/3.4.3/bin
export PATH=${PATH_R}:${PATH_PERL}:${PATH_SAMTOOLS}:${PATH_HOMER}:${PATH}
export PATH

# Defining the program versions (needs to be adjusted to the server/workstation used)
# rhskl17 version
SKEWER="/misc/software/ngs/skewer/skewer-0.2.2/skewer"
BOWTIE="/misc/software/ngs/bowtie/bowtie2-2.3.4-linux-x86_64/bowtie2"
SAMTOOLS="/misc/software/ngs/samtools/samtools-1.6/bin/samtools"
PICARD="/usr/bin/java -jar /misc/software/ngs/picard/src/v2.17.3/picard.jar"
IGVTOOLS="/usr/bin/java -Xmx4g -Djava.awt.headless=true -jar
/misc/software/viewer/IGV/IGVTools_2.3.98/igvtools.jar"

# required input:
# -f <fastq or fastq.gz> forward read
# -r <fastq or fastq.gz> reverse read (if not available, will run as single end)
# -g <genome> available genomes include hg19, GRCh38, mm9, mm10
# -o <output subdirectory> only subdirectories above ATAC-folder are given
# -n <sample name> used for sam, bigwig and tag directory
# optional:
# -t <number of threads>
# -s <path to adapter sequences> (or type NEXTERA to use Nextera-adapters)

# Set Script Name variable
SCRIPT=`basename ${BASH_SOURCE[0]}`

# Set fonts for Help.
NORM=`tput sgr0`
BOLD=`tput bold`
REV=`tput smso`

# USAGE function
function USAGE {
echo -e \\n"Help documentation for ${BOLD}${SCRIPT}.${NORM}"\\n
echo -e "${REV}Basic usage:${NORM}\n${BOLD}$SCRIPT -g <genome> -f <fastq> -r <fastq> -o
<subdirectory> -n <name> -t <threads>${NORM}"\\n
echo "Required:"
echo "${BOLD}-g${NORM} <genome> available genomes include hg19, GRCh38, mm9, mm10"
echo "${BOLD}-f${NORM} <fastq or fastq.gz> forward read"
echo "${BOLD}-r${NORM} <fastq or fastq.gz> reverse read (if not available, will run as
single end)"
echo "${BOLD}-d${NORM} <output subdirectory> only subdirectories above ATAC-folder are
given"
echo "${BOLD}-n${NORM} <sample name> used for sam, bigwig and tag directory"
echo -e "${BOLD}-t${NORM} <number of threads> (optional/default 16)"\\n
echo -e "${BOLD}-s${NORM} <path to adapter sequences> (or type NEXTERA to use
Nextera-adapters)"\\n
echo -e "Example: ${BOLD}$SCRIPT -g hg19 -f readA.fastq -r readB.fastq -d MOMACDC -n test
-t 32 -s NEXTERA${NORM}"\\n
exit 1
}

#Check the number of arguments. If none are passed, print help and exit.
NUMARGS=$#
if [ $NUMARGS -eq 0 ]; then
USAGE
fi

# Defining available genomes
GENOMES=(hg19 GRCh38 hg38 mm9 mm10)

# Set default number of threads to 16
THREADS=16
TRIM=0

# Parse command line options
while getopts :g:f:r:d:n:t:s:h OPTIONS; do
case $OPTIONS in
g) #set option "g" - Bowtie will will only accept GRCh38 while HOMER needs hg38
GENOME=$OPTARG
HOMERGENOME=$OPTARG
;;
f) #set option "f"
FASTQR1=$OPTARG
;;
r) #set option "r"
FASTQR2=$OPTARG
;;
d) #set option "d"
DIRECTORY=$OPTARG
;;
n) #set option "n"
SAMPLENAME=$OPTARG
;;
t) #set option "t"
THREADS=$OPTARG
SAMTHREADS=$((THREADS-1))
;;
s) #set option "s"
ADAPTER=$OPTARG
TRIM=1
;;
h) #show help
USAGE
;;
\?) #unrecognized option - show USAGE
echo -e \\n"Option -${BOLD}$OPTARG${NORM} not allowed."
USAGE
;;
esac
done
shift $((OPTIND-1))

# Sanity checks
# Check whether a reasonable number of cores is given
if [ $THREADS -lt 1 ] || [ "$THREADS" -gt 63 ]; then
echo "Number of threads should be in the range of 1 - 63!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 9
fi

# Check whether GENOME is available
INARRAY=$(echo ${GENOMES[@]} | grep -o "$GENOME" | wc -w)
if [ $INARRAY == 0 ] ; then
echo -e \\n"Genome -${BOLD}$GENOME${NORM} not available."
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 2
fi
GENOMEPATH=$GENOME
HOMERGENOME=$GENOME
BIGWIGGENOME=$GENOME

# path to chromosome sizes
CHROMSIZES="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/$GENOME.chrom.sizes"
if [ $GENOME == "hg38" ] || [ $GENOME == "GRCh38" ] ; then
GENOME="hg38"
```

```
GENOMEPATH="GRCh38.PRI_p10"
BIGWIGGENOME="GRCh38"
HOMERGENOME="hg38"
CHROMSIZES="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/GRCh38.PRI_p10.chrom.sizes
"
fi

#bowtie index
INDEX="$GENOMEPATH/$GENOME"

# Check whether one or two reads are given, whether files are available, non-identical and
have the right extensions
if [ ! -f "$FASTQR1" ]; then
echo "Fastq file (read 1) not found!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 3
fi
case "$FASTQR1" in
*.fastq.gz) ;;
*.fastq) ;;
*.fq.gz) ;;
*.fq) ;;
*) echo "Fastq file (read 1) has wrong extension (should be fastq, fq, fastq.gz
or fq.gz!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 4
;;
esac
if [ -z "$FASTQR2" ]; then
SEQTYPE=1
TYPE="SE"
SKEWTYPE="any"
echo -e \\n"Data processing for ${BOLD}single-end${NORM} reads"
elif [ ! -f "$FASTQR2" ]; then
echo "Fastq file (read 2) not found!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 5
elif [ "$FASTQR1" == "$FASTQR2" ]; then
echo "Fastq files for read 1 and 2 are identical!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 6
else
SEQTYPE=2
TYPE="PE"
SKEWTYPE="pe"
echo -e \\n"Data Processing for ${BOLD}paired-end${NORM} reads"
case "$FASTQR2" in
*.fastq.gz) ;;
*.fastq) ;;
*.fq.gz) ;;
*.fq) ;;
*) echo "Fastq file (read 2) has wrong extension (should be fastq, fq,
fastq.gz or fq.gz!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 7
;;
esac
fi

# Check wether the subfolders are already there - if not ask whether path should be made
if [ -z "$DIRECTORY" ]; then
echo "Subdirectory is missing!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 7
fi

# rhskl17 version
MAPPINGDIR="/misc/data/processedData/mapping/chromatin/${BIGWIGGENOME}/ATAC/${DIRECTORY}"
TAGDIRDIR="/misc/data/processedData/tagDir/chromatin/${BIGWIGGENOME}/ATAC/${DIRECTORY}"
BIGWIGDIR="/misc/data/processedData/bigWig/chromatin/${BIGWIGGENOME}/ATAC/${DIRECTORY}"
if [ ! -d "$MAPPINGDIR" ]; then #looks for the mapping output folder - if not
available, askes whether to make it
echo "$MAPPINGDIR"
read -p "Do you want to create this mapping folder (y/n)?" choice
case "$choice" in
y|Y ) mkdir -p "$MAPPINGDIR"
;;
n|N ) echo "Didn't create the folder and stopped!"
```

```
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 8
;;
* ) echo "invalid response";;
esac
fi
if [ ! -d "$TAGDIRDIR" ]; then #looks for the tagDir output folder - if not
available, askes whether to make it
echo "$TAGDIRDIR"
read -p "Do you want to create this tagDirectory folder (y/n)?" choice
case "$choice" in
y|Y ) mkdir -p "$TAGDIRDIR"
;;
n|N ) echo "Didn't create the folder and stopped!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 8
;;
* ) echo "invalid response";;
esac
fi
if [ ! -d "$BIGWIGDIR" ]; then #looks for the bigWig output folder - if not
available, askes whether to make it
echo "$BIGWIGDIR"
read -p "Do you want to create this bigWig folder (y/n)?" choice
case "$choice" in
y|Y ) mkdir -p "$BIGWIGDIR"
;;
n|N ) echo "Didn't create the folder and stopped!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 8
;;
* ) echo "invalid response";;
esac
fi
# Check whether sample name is given
if [ -z "$SAMPLENAME" ]; then
echo "Sample name is missing!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 9
fi
SAMPLENAME="${SAMPLENAME// /_}" # replaces spaces with underscores in sample name, just in
case...

# Check if path for trimming is given or set for NEXTERA
if [ $TRIM == 1 ] && [ $ADAPTER == "NEXTERA" ]; then
ADAPTER="/misc/software/ngs/skewer/adapterFiles/nextera_adapters.fa"
elif [ $TRIM == 1 ] && [ ! -f "$ADAPTER" ]; then
echo "Adapter file not found!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 11
fi
if [ ! -d "${MAPPINGDIR}/logs" ]; then #looks for the logs output folder - if not
available, makes it
mkdir "${MAPPINGDIR}/logs"
fi

# If everything passed, create a log file for all the commands that are executed during the
processing
LOGFILE="${MAPPINGDIR}/logs/${SAMPLENAME}.commands.txt"
touch $LOGFILE

# Function to append command lines into logfile and then execute
exe() { echo -e "\$ $@"\\n >> $LOGFILE; "$@" ; }

# Generating a random number file name for temporary files
TEMPFILES="/loctmp/$RANDOM.tmp"

# Optionally starting the pipeline by trimming the reads with skewer
if [ $TRIM == 1 ] ; then
case "$SEQTYPE" in
1 ) exe $SKEWER -f sanger -t $THREADS -m any -x $ADAPTER -o $TEMPFILES $FASTQR1
FASTQR1="$TEMPFILES-trimmed.fastq"
mv "$TEMPFILES-trimmed.log" "${MAPPINGDIR}/logs/${SAMPLENAME}.trimming.log"
;;
2 ) exe $SKEWER -f sanger -t $THREADS -m pe -x $ADAPTER -o $TEMPFILES $FASTQR1 $FASTQR2
FASTQR1="$TEMPFILES-trimmed-pair1.fastq"
FASTQR2="$TEMPFILES-trimmed-pair2.fastq"
mv "$TEMPFILES-trimmed.log" "${MAPPINGDIR}/logs/${SAMPLENAME}.trimming.log"
```

```
;;
esac
fi

# Running Bowtie2
echo -e \\n"Running Bowtie2 on $SAMPLENAME."\\n
case "$SEQTYPE" in
1 ) exe $BOWTIE --very-sensitive -p $THREADS -x $INDEX --met-file
$MAPPINGDIR/logs/$SAMPLENAME.aln_metrics.txt -U $FASTQR1 2>
$MAPPINGDIR/logs/$SAMPLENAME.aln_rates.txt -S $TEMPFILES.sam
;;
2 ) exe $BOWTIE --very-sensitive --no-discordant -p $THREADS -x $INDEX --met-file
$MAPPINGDIR/logs/$SAMPLENAME.aln_metrics.txt -1 $FASTQR1 -2 $FASTQR2 2>
$MAPPINGDIR/logs/$SAMPLENAME.aln_rates.txt -S $TEMPFILES.sam
;;
esac

# Sorting and reducing the SAM-file to q>10
echo -e "Sorting and reducing $SAMPLENAME to q>10."\\n
exe $SAMTOOLS view -h -S -q 10 $TEMPFILES.sam -@ $SAMTHREADS -o $TEMPFILES.red.sam
exe $SAMTOOLS sort -O sam -@ $SAMTHREADS -T $TEMPFILES -o $TEMPFILES.sorted.sam
$TEMPFILES.red.sam

# Shifting the read positions for ATAC
echo -e "Shifting the read positions."\\n
exe myATACshiftSAM.pl "$TEMPFILES.sorted.sam" "$TEMPFILES.shifted" "$TYPE"
echo -e "Generating TagDir, BigWig, and TDF and finding Peaks."\\n

# Generating a HOMER tag directory
exe makeTagDirectory $TAGDIRDIR/$SAMPLENAME $TEMPFILES.shifted.sam -genome $HOMERGENOME
-checkGC

# Making a UCSC browser track
exe makeUCSCfile $TAGDIRDIR/$SAMPLENAME -bigWig $CHROMSIZES -o $BIGWIGDIR/$SAMPLENAME.bigwig

# Making an IGV browser track
exe $IGVTOOLS count -e 150 $TEMPFILES.shifted.sam $BIGWIGDIR/$SAMPLENAME.cov.tdf $CHROMSIZES

# Finding peaks
exe findPeaks $TAGDIRDIR/$SAMPLENAME -region -size 150 -o auto
if [ "$SEQTYPE" == 2 ]; then
exe $PICARD CollectInsertSizeMetrics \
I="$TEMPFILES.shifted.sam" \
O="$TAGDIRDIR/$SAMPLENAME/insert_size_metrics.txt" \
H="$TAGDIRDIR/$SAMPLENAME/insert_size_histogram.pdf" \
W=500
fi
exe $SAMTOOLS view -bS -@ $SAMTHREADS -o "${MAPPINGDIR}/${SAMPLENAME}.shifted.bam"
$TEMPFILES.shifted.sam

# removing tmp files
rm ${TEMPFILES}.*
```

### 10.1.1.3 Mapping of RNAseq Data

RNAseq data was generated by indexed (maximum 10 samples/lane) paired end 42 or 78 bp sequencing on a HiSeq 3000/4000 sequencer (Illumina) using the latest four-channel sequencing chemistry SBSv4 (Illumina) or a Nextseq 550 (Illumina) using the latest two-channel sequencing chemistry SBSv2 (Illumina). The general RNAseq workflow is explained in section 4.2.9. Obtained raw read bam files from the BSF were processed using a paired end adapted version of the getBSFconvertSE.sh script (see section 10.1.1.1). Onsite sequenced RNAseq samples on the other hand, were converted using the bcl2fastq script (Illumina, see section 10.1.1.2).

```
#!/bin/bash
# script from Michael Rehli
# bash script to download and check md5 sum for bam-files produced at the BSF in Vienna
args=("$@")
if [ ${#args[@]} != 7 ] ; then
echo -e "\nMissing or too many arguments! \n"
echo "Correct usage:"
echo "getBSFconvertPE.sh <user name> <password> <flowcell ID> <lane number> <sample
name> <output subdirectory in misc/data/rawdata> <new sample name>"
echo "Example:"
echo -e "getBSFconvertPE.sh mrehli xyz BSF_0411_HNH5NBBXX 4 S_10_ChIP_DD_S33767
chromatin/ChIP/MOMACDC 3dDCs_DSG+FA_R2_EGR2_A \n"
exit
fi

# defining variables per default and operator input
WEBFOLDER="https://biomedical-sequencing.at/downloads/download/BSF_downloads/group_Michael_Reh
li"
USER=${args[0]}
PASSW=${args[1]}
EXPERIMENT=${args[2]}
LANE="${args[3]}"
SAMPLE=${args[4]}
RUN=$(echo $EXPERIMENT| cut -d'_' -f 2)
PREFIX=${args[2]}_${args[3]}_
BSFFOLDER="/misc/data/rawData/BSF/run${RUN}"
echo -e "\ngetBSFconvert.sh will download data from the Biomedical Sequencing Facility in
Vienna,"
echo -e "convert it into FASTQ format, run FASTQC.\n"
BSFBAM=$WEBFOLDER/${args[0]}/${args[2]}/$PREFIX${args[4]}.bam
BSFMD5=$WEBFOLDER/${args[0]}/${args[2]}/$PREFIX${args[4]}.bam.md5
OWNBAM=$BSFFOLDER/$PREFIX${args[4]}.bam
OWNMD5=$BSFFOLDER/$PREFIX${args[4]}.bam.md5
if [ ! -d "$BSFFOLDER" ]; then
mkdir "$BSFFOLDER"
echo -e "generating the folder $BSFFOLDER \n"
fi
TARGETFOLDER="/misc/data/rawData/${args[5]}"
TARGETR1="${TARGETFOLDER}/${args[6]}.R1.fastq"
TARGETR2="${TARGETFOLDER}/${args[6]}.R2.fastq"
if [ ! -d "$TARGETFOLDER" ]; then
mkdir -p "${TARGETFOLDER}/FastQC"
echo -e "generating the folder $TARGETFOLDER \n"
fi

# downloading the .bam and .bam.md5 files
echo -e "downloading the bam-file for $SAMPLE\n"
curl -u $USER:$PASSW $BSFBAM -o $OWNBAM --insecure
echo -e "downloading the md5-file for $SAMPLE\n"
curl -u $USER:$PASSW $BSFMD5 -o $OWNMD5 –insecure

#comparing checksums created in Vienna and here
if [ $(md5sum "$OWNBAM" | cut -b-32) == $(cat "$OWNMD5") ]; then
echo "MD5 checksum matches"

# provided that the checksums match, proceed with converting into fastq, FastQC, and gzip
echo -e "\nconverting the bam-file into fastq format \n"
/usr/bin/java -jar /misc/software/ngs/picard/src/v2.17.3/picard.jar SamToFastq I=$OWNBAM
FASTQ=$TARGETR1 SECOND_END_FASTQ=$TARGETR2
echo -e "running fastqc \n"
fastqc-0.11.7 -o "${TARGETFOLDER}/FastQC" $TARGETR1
fastqc-0.11.7 -o "${TARGETFOLDER}/FastQC" $TARGETR2
echo -e "zipping the fastq file \n"
gzip $TARGETR1
gzip $TARGETR2
else
echo "MD5 checksum different! Try downloading again!"
fi
```

Generated fastq data sets were mapped to the human reference genome (version hg19) using the following bash script, which implements parts of the HOMER suite (Heinz et al. 2010) and uses STAR for mapping.

```
#!/bin/bash
# script from Michael Rehli
# bash script to map RNA-seq data, generate a count table, and coverage bigWigs
# will only keep unique.aligned.bam, bigwig and tdf files as well as gene count data
# IF YOU REQUIRE MULTIMAPPERS, OR ANY OTHER ORIGINAL FILES, PLEASE RUN STAR MANUALLY

#setting homer environment
DIR_PKG="/misc/software/ngs"
PATH_PERL=/misc/software/package/perl/perl-5.26.1/bin
PATH_SAMTOOLS=${DIR_PKG}/samtools/samtools-1.6/bin
PATH_HOMER=${DIR_PKG}/homer/v4.9/bin
PATH_R=/misc/software/package/RBioC/3.4.3/bin
export PATH=${PATH_R}:${PATH_PERL}:${PATH_SAMTOOLS}:${PATH_HOMER}:${PATH}
export PATH

# Defining the program versions (needs to be adjusted to the server/workstation used)
# rhskl17 version
STAR="/misc/software/ngs/STAR/v2.5.3a/STAR/source/STAR"
SAMTOOLS="/misc/software/ngs/samtools/samtools-1.6/bin/samtools"
PICARD="/usr/bin/java -jar /misc/software/ngs/picard/src/v2.17.3/picard.jar"
IGVTOOLS="/usr/bin/java -Xmx4g -Djava.awt.headless=true -jar
/misc/software/viewer/IGV/IGVTools_2.3.98/igvtools.jar"
INDEXSDIR="/misc/software/ngs/genome/index/STAR"

# required input:
# -f <fastq or fastq.gz> forward read
# -r <fastq or fastq.gz> reverse read (if not available, or set "NA" will run as single end)
# -g <genome> available genomes include hg19, GRCh38, GRCm38
# -l <read length>
# -o <output subdirectory> only subdirectories above RNA-folder are given
# -n <sample name> used for sam, bigwig and tag directory
# optional:
# -t <number of threads>
# -c <read length> trim reads to the given length

# Set Script Name variable
SCRIPT=`basename ${BASH_SOURCE[0]}`

# Set fonts for Help.
NORM=`tput sgr0`
BOLD=`tput bold`
REV=`tput smso`

# USAGE function
function USAGE {
echo -e \\n"Help documentation for ${BOLD}${SCRIPT}.${NORM}"\\n
echo -e "${REV}Basic usage:${NORM}\n${BOLD}$SCRIPT -g <genome> -f <fastq> -r <fastq> -o
<subdirectory> -n <name> -t <threads> -c <trim length>${NORM}"\\n
echo "Required:"
echo "${BOLD}-g${NORM} <genome> available genomes include hg19, GRCh38, GRCm38"
echo "${BOLD}-f${NORM} <fastq or fastq.gz> forward read"
echo "${BOLD}-r${NORM} <fastq or fastq.gz> reverse read (if not available, or set \"NA\"
will run as single end)"
echo "${BOLD}-d${NORM} <output subdirectory> only subdirectories above RNA-folder are given"
echo "${BOLD}-n${NORM} <sample name> used for sam, bigwig and tag directory"
echo "${BOLD}-t${NORM} <number of threads> (optional/default 16)"
echo "${BOLD}-l${NORM} <read length> used for index"
echo "${BOLD}-c${NORM} <trim length> trims Fastq file to the given length before
processing (optional)"
echo "Available indices:"
for entry in $INDEXSDIR/*
do
echo $entry
done
echo -e \\n"Example: ${BOLD}$SCRIPT -g hg19 -l 50 -f read1.fastq.gz -r read2.fastq.gz -d
test -n testrun -t 24${NORM} -c 43"\\n
exit 1
}

#Check the number of arguments. If none are passed, print help and exit.
NUMARGS=$#
if [ $NUMARGS -eq 0 ]; then
USAGE
fi
```

```
# Defining available genomes
GENOMES=(hg19 GRCh38 hg38 GRCm38)

# Set default number of threads to 16
THREADS=16
TRIM=0

# Parse command line options
while getopts :g:f:r:d:n:t:l:c:h OPTIONS; do
case $OPTIONS in
g) #set option "g" - Bowtie will will only accept GRCh38 while HOMER needs hg38
GENOME=$OPTARG
HOMERGENOME=$OPTARG
;;
f) #set option "f"
FASTQR1=$OPTARG
;;
r) #set option "r"
FASTQR2=$OPTARG
;;
d) #set option "d"
DIRECTORY=$OPTARG
;;
n) #set option "n"
SAMPLENAME=$OPTARG
;;
t) #set option "t"
THREADS=$OPTARG
SAMTHREADS=$((THREADS-1))
;;
l) #set option "l"
LENGTH=$OPTARG
;;
c) #set option "c"
TRIMLENGTH=$OPTARG
TRIM=1
;;
h) #show help
USAGE
;;
\?) #unrecognized option - show USAGE
echo -e \\n"Option -${BOLD}$OPTARG${NORM} not allowed."
USAGE
;;
esac
done
shift $((OPTIND-1))

# Sanity checks
# Check whether a reasonable number of cores is given
if [ $THREADS -lt 1 ] || [ "$THREADS" -gt 63 ]; then
echo "Number of threads should be in the range of 1 - 63!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 9
fi

# Check whether GENOME is available
INARRAY=$(echo ${GENOMES[@]} | grep -o "$GENOME" | wc -w)
if [ $INARRAY == 0 ] ; then
echo -e \\n"Genome -${BOLD}$GENOME${NORM} not available."
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 2
fi
GENOMEPATH=$GENOME
BIGWIGGENOME=$GENOME

# path to chromosome sizes
CHROMSIZES="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/$GENOME.chrom.sizes"


# STAR index
INDEX="$INDEXSDIR/index_${GENOME}_${LENGTH}"
if [ $GENOME == "hg38" ] || [ $GENOME == "GRCh38" ] ; then
GENOME="hg38"
INDEX="$INDEXSDIR/index_GRCh38.PRI.p10_$LENGTH"
BIGWIGGENOME="GRCh38"
CHROMSIZES="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/GRCh38.PRI_p10.chrom.sizes
"
```

```
fi
if [ $GENOME == "mm10" ] || [ $GENOME == "GRCm38" ] ; then
GENOME="GRCm38"
INDEX="$INDEXSDIR/index_GRCm38.PRI.p5_$LENGTH"
BIGWIGGENOME="GRCm38"
CHROMSIZES="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/GRCm38.PRI_p5.chrom.sizes"
fi

#check if index is available
if [ -z "$INDEX" ]; then
echo "Index is missing!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 7
fi

# Check whether one or two reads are given, whether files are available, non-identical and
have the right extensions
if [ ! -f "$FASTQR1" ]; then
echo "Fastq file (read 1) not found!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 3
fi
case "$FASTQR1" in
*.fastq.gz) ZIP="--readFilesCommand gunzip -c "
;;
*.fastq) ZIP=""
;;
*.fq.gz) ZIP="--readFilesCommand gunzip -c "
;;
*.fq) ZIP=""
;;
*) echo "Fastq file (read 1) has wrong extension (should be fastq, fq, fastq.gz
or fq.gz!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 4
;;
esac
if [ -z "$FASTQR2" ] || [ "$FASTQR2" == "NA" ] || [ "$FASTQR2" == "na" ] ; then
SEQTYPE=1
TYPE="SE"
SKEWTYPE="any"
echo -e \\n"Data processing for ${BOLD}single-end${NORM} reads"
elif [ ! -f "$FASTQR2" ]; then
echo "Fastq file (read 2) not found!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 5
elif [ "$FASTQR1" == "$FASTQR2" ]; then
echo "Fastq files for read 1 and 2 are identical!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 6
else
SEQTYPE=2
TYPE="PE"
SKEWTYPE="pe"
echo -e \\n"Data Processing for ${BOLD}paired-end${NORM} reads"
case "$FASTQR2" in
*.fastq.gz) ZIP="--readFilesCommand gunzip -c "
;;
*.fastq) ZIP=""
;;
*.fq.gz) ZIP="--readFilesCommand gunzip -c "
;;
*.fq) ZIP=""
;;
*) echo "Fastq file (read 2) has wrong extension (should be fastq, fq,
fastq.gz or fq.gz!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 7
;;
esac
fi

# Check whether the subfolders are already there - if not ask whether path should be made
if [ -z "$DIRECTORY" ]; then
echo "Subdirectory is missing!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 7
fi
```

```
# rhskl17 version
MAPPINGDIR="/misc/data/processedData/mapping/RNA/${BIGWIGGENOME}/RNAseq/${DIRECTORY}"
BIGWIGDIR="/misc/data/processedData/bigWig/RNA/${BIGWIGGENOME}/RNAseq/${DIRECTORY}"
if [ ! -d "$MAPPINGDIR" ]; then #looks for the mapping output folder - if not available, askes whether to make it
echo "$MAPPINGDIR"
read -p "Do you want to create this mapping folder (y/n)?" choice
case "$choice" in
y|Y ) mkdir -p "$MAPPINGDIR"
;;
n|N ) echo "Didn't create the folder and stopped!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 8
;;
* ) echo "invalid response";;
esac
fi
if [ ! -d "$BIGWIGDIR" ]; then #looks for the bigWig output folder - if not available, askes whether to make it
echo "$BIGWIGDIR"
read -p "Do you want to create this bigWig folder (y/n)?" choice
case "$choice" in
y|Y ) mkdir -p "$BIGWIGDIR"
;;
n|N ) echo "Didn't create the folder and stopped!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 8
;;
* ) echo "invalid response";;
esac
fi

# Check whether sample name is given
if [ -z "$SAMPLENAME" ]; then
echo "Sample name is missing!"
echo -e "Use ${BOLD}$SCRIPT -h${NORM} to see the help documentation."\\n
exit 9
fi
SAMPLENAME="${SAMPLENAME// /_}" # replaces spaces with underscores in sample name, just in case...

# Creating the folder for STAR
if [ ! -d "${MAPPINGDIR}/${SAMPLENAME}" ]; then #looks for the mapping output folder - if not available, askes whether to make it
mkdir "${MAPPINGDIR}/${SAMPLENAME}"
fi
if [ ! -d "${MAPPINGDIR}/logs" ]; then #looks for the logs output folder - if not available, makes it
mkdir "${MAPPINGDIR}/logs"
fi

# If everything passed, create a log file for all the commands that are executed during the processing
LOGFILE="${MAPPINGDIR}/logs/${SAMPLENAME}.commands.txt"
touch $LOGFILE

# Function to append command lines into logfile and then execute
exe() { echo -e "\$ $@"\\n >> $LOGFILE; "$@" ; }

# Generating a random number file name for temporary files
TEMPFILES="/loctmp/$RANDOM.tmp"

# Optionally starting the pipeline by trimming the reads with homerTools
if [ $TRIM == 1 ] ; then
ZIP=""
case "$SEQTYPE" in
1 ) exe homerTools trim -len $TRIMLENGTH $FASTQR1
exe mv "$FASTQR1.trimmed" "$TEMPFILES-trimmed.fastq"
FASTQR1="$TEMPFILES-trimmed.fastq"
# mv "$FASTQR1.lengths" "${MAPPINGDIR}/logs/${SAMPLENAME}.trimming.log"
;;
2 ) exe homerTools trim -len $TRIMLENGTH $FASTQR1
exe mv "$FASTQR1.trimmed" "$TEMPFILES-trimmed-pair1.fastq"
FASTQR1="$TEMPFILES-trimmed-pair1.fastq"
# mv "$FASTQR1.lengths" "${MAPPINGDIR}/logs/${SAMPLENAME}.R1.trimming.log"
exe homerTools trim -len $TRIMLENGTH $FASTQR2
exe mv "$FASTQR2.trimmed" "$TEMPFILES-trimmed-pair2.fastq"
FASTQR2="$TEMPFILES-trimmed-pair2.fastq"
```

```
# mv "$FASTQR2.lengths" "${MAPPINGDIR}/logs/${SAMPLENAME}.R2.trimming.log"
;;
esac
fi

# Running STAR
echo -e \\n"Running STAR on $SAMPLENAME."\\n
cd "${MAPPINGDIR}/${SAMPLENAME}"
case "$SEQTYPE" in
1 ) exe $STAR --runThreadN $THREADS \
--genomeDir $INDEX \
--readFilesIn $FASTQR1 \
$ZIP\
--outFilterIntronMotifs RemoveNoncanonicalUnannotated \
--outReadsUnmapped Fastq \
--alignSJoverhangMin 8 \
--alignSJDBoverhangMin 1 \
--alignMatesGapMax 1000000 \
--alignIntronMax 1000000 \
--outSAMtype BAM SortedByCoordinate \
--quantMode GeneCounts \
--outWigType bedGraph \
--outWigStrand Stranded
;;
2 ) exe $STAR --runThreadN $THREADS \
--genomeDir $INDEX \
--readFilesIn $FASTQR1 $FASTQR2 \
$ZIP\
--outFilterIntronMotifs RemoveNoncanonicalUnannotated \
--outReadsUnmapped Fastq \
--alignSJoverhangMin 8 \
--alignSJDBoverhangMin 1 \
--alignMatesGapMax 1000000 \
--alignIntronMax 1000000 \
--outSAMtype BAM SortedByCoordinate \
--quantMode GeneCounts \
--outWigType bedGraph \
--outWigStrand Stranded
;;
esac

# Sorting bedgraph files
echo -e \\n"Sorting the bedGraph files for $SAMPLENAME."\\n
LC_COLLATE=C sort -k1,1 -k2,2n $MAPPINGDIR/$SAMPLENAME/Signal.Unique.str1.out.bg -o
$TEMPFILES.str1.sorted.bg
LC_COLLATE=C sort -k1,1 -k2,2n $MAPPINGDIR/$SAMPLENAME/Signal.Unique.str2.out.bg -o
$TEMPFILES.str2.sorted.bg
echo -e "\$ LC_COLLATE=C sort -k1,1 -k2,2n $MAPPINGDIR/$SAMPLENAME/Signal.Unique.str1.out.bg
-o $TEMPFILES.str1.sorted.bg"\\n >> $LOGFILE
echo -e "\$ LC_COLLATE=C sort -k1,1 -k2,2n $MAPPINGDIR/$SAMPLENAME/Signal.Unique.str2.out.bg
-o $TEMPFILES.str2.sorted.bg"\\n >> $LOGFILE

# inverting values for the second strand
exe awk 'BEGIN{FS=OFS="\t"}{print $1,$2,$3,$4*(-1)}' $TEMPFILES.str2.sorted.bg >
$TEMPFILES.str2.inv.bg

# generating bigWigs for both strands
echo -e \\n"Generating bigWig files for $SAMPLENAME."\\n
exe bedGraphToBigWig $TEMPFILES.str1.sorted.bg $CHROMSIZES
$BIGWIGDIR/$SAMPLENAME.Unique.for.bigwig
exe bedGraphToBigWig $TEMPFILES.str2.inv.bg $CHROMSIZES
$BIGWIGDIR/$SAMPLENAME.Unique.rev.bigwig

# generating corresponding tdf files for IGV browser
echo -e "Generating TDF files for $SAMPLENAME."\\n
exe bigWigToWig $BIGWIGDIR/$SAMPLENAME.Unique.for.bigwig $TEMPFILES.str1.sorted.wig
-udcDir=/loctmp
exe bigWigToWig $BIGWIGDIR/$SAMPLENAME.Unique.rev.bigwig $TEMPFILES.str2.inv.wig
-udcDir=/loctmp
exe $IGVTOOLS toTDF $TEMPFILES.str1.sorted.wig $BIGWIGDIR/$SAMPLENAME.Unique.for.cov.tdf
$CHROMSIZES
exe $IGVTOOLS toTDF $TEMPFILES.str2.inv.wig $BIGWIGDIR/$SAMPLENAME.Unique.rev.cov.tdf
$CHROMSIZES

# move and rename gene count table and bam file
case "$TRIM" in
1 ) exe mv ${MAPPINGDIR}/${SAMPLENAME}/Aligned.sortedByCoord.out.bam
${MAPPINGDIR}/${SAMPLENAME}.trimmed.sorted.bam
```

```
exe mv ${MAPPINGDIR}/${SAMPLENAME}/ReadsPerGene.out.tab
${MAPPINGDIR}/${SAMPLENAME}.trimmed.ReadsPerGene.txt
exe mv ${MAPPINGDIR}/${SAMPLENAME}/Log.final.out
${MAPPINGDIR}/logs/${SAMPLENAME}.trimmed.summary.log.txt
exe mv ${MAPPINGDIR}/${SAMPLENAME}/Log.out
${MAPPINGDIR}/logs/${SAMPLENAME}.trimmed.processing.log.txt
;;
0 ) exe mv ${MAPPINGDIR}/${SAMPLENAME}/Aligned.sortedByCoord.out.bam
${MAPPINGDIR}/${SAMPLENAME}.sorted.bam
exe mv ${MAPPINGDIR}/${SAMPLENAME}/ReadsPerGene.out.tab
${MAPPINGDIR}/${SAMPLENAME}.ReadsPerGene.txt
exe mv ${MAPPINGDIR}/${SAMPLENAME}/Log.final.out
${MAPPINGDIR}/logs/${SAMPLENAME}.summary.log.txt
exe mv ${MAPPINGDIR}/${SAMPLENAME}/Log.out
${MAPPINGDIR}/logs/${SAMPLENAME}.processing.log.txt
;;
esac

# removing tmp files
exe rm -r "${MAPPINGDIR}/${SAMPLENAME}"
exe rm ${TEMPFILES}.*
```

### 10.1.2 Analysis of PU.1 ChIPseq Data of Various Cell Types

Obtained data sets were mapped to the human reference genome (version hg19; see section 10.1.1.1) and further processed using the following bash script, which implements parts of the HOMER suite (Heinz et al. 2010) as well as parts of the R software (R Development Core Team 2008).

```
#!/bin/bash
#setting homer environment
DIR_PKG="/misc/software/ngs"
PATH_PERL=/misc/software/package/perl/perl-5.26.1/bin
PATH_SAMTOOLS=${DIR_PKG}/samtools/samtools-1.6/bin
PATH_HOMER=${DIR_PKG}/homer/v4.9/bin
PATH_R=/misc/software/package/RBioC/3.4.3/bin
export PATH=${PATH_R}:${PATH_PERL}:${PATH_SAMTOOLS}:${PATH_HOMER}:${PATH}
export PATH

BEDTOOLS="/misc/software/ngs/bedtools/bedtools2-2.27.1/bin/bedtools"
BLACKLIST_HG19="/misc/data/analysis/generalStuff/annotation/hg19/hg19.blacklist.bed"
WORKDIR="/misc/data/analysis/project_PU1/allCellTypes"
PEAKDIR="${WORKDIR}/peaks"
FIGURESDIR="${WORKDIR}/figures"
SRADIR="/misc/data/rawData/SRA"
MOTIFDIR="${WORKDIR}/motifs"
TMPDIR="/loctmp"
CHROMSIZES="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/hg19.chrom.sizes"
PU1MOTIF="/misc/data/analysis/project_PU1/PU1long.motif"
PHYLOP="/misc/data/analysis/generalStuff/conservation/hg19/hg19.100way.phyloP100way.bedGraph"
PHASTCONS="/misc/data/analysis/generalStuff/conservation/hg19/hg19.100way.phastCons.bedGraph"
HISTDIR="${WORKDIR}/peaks/hist"

# tagdirectories ChIP and Input
CHIPDIRBC="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/bloodCells"
INDIRBC="/misc/data/processedData/tagDir/DNA/hg19/Input/bloodCells"
CHIPDIRPD="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/publishedData"
INDIRPD="/misc/data/processedData/tagDir/DNA/hg19/Input/publishedData"
CHIPDIRMOMACDC="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/MOMACDC"
INDIRMOMACDC="/misc/data/processedData/tagDir/DNA/hg19/Input/MOMACDC"
CHIPDIRCL="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/CellLines"
INDIRCL="/misc/data/processedData/tagDir/DNA/hg19/Input/CellLines"
TMP="/misc/data/tmp"

# bigwigdirectories ChIP
BIGWIGDIRCL="/misc/data/processedData/bigWig/chromatin/hg19/ChIP/CellLines"
BIGWIGDIRPD="/misc/data/processedData/bigWig/chromatin/hg19/ChIP/publishedData"
BIGWIGDIRBC="/misc/data/processedData/bigWig/chromatin/hg19/ChIP/bloodCells"
BIGWIGDIRMOMACDC="/misc/data/processedData/bigWig/chromatin/hg19/ChIP/MOMACDC"
```

```
# directory for manually annotated motif lists
CUSTOMMOTIFS="/misc/data/analysis/project_PU1/allCellTypes/motifs/customAnnotation"

mkdir ${CUSTOMMOTIFS}
mkdir ${HISTDIR}

# BASIC ANALYSIS OF ALL Available PU1 ChIP SEQs

# CNV normalization for cell lines
# combined Input for RS411 cells
cat ${SRADIR}/RS411_Dex1h_Input.fastq.gz ${SRADIR}/RS411_NoStim_Input.fastq.gz >
${SRADIR}/RS411_comb_Input.fastq.gz
cat ${SRADIR}/K562_input_R2.1.fastq.gz ${SRADIR}/K562_input_R2.2.fastq.gz >
${SRADIR}/K562_input_R2comb.fastq.gz

# reran mapChIP.sh to include CNV calculations

# normalize & reduce tagDir
declare -a clTagDir=("${CHIPDIRCL}/other/EM3_PU1" "${CHIPDIRCL}/other/KG1_PU1"
"${CHIPDIRCL}/other/ML2_PU1" "${CHIPDIRCL}/other/U937_PU1" "${CHIPDIRCL}/other/NB4_PU1")
declare -a clinTagDir=("${INDIRCL}/EM3_genInput_621" "${INDIRCL}/KG1a_genInput_622"
"${INDIRCL}/ML2_genInput_623" "${INDIRCL}/U937_genInput_700" "${INDIRCL}/NB4_genInput_624")
CNVPATH="/misc/data/processedData/mapping/DNA/hg19/Input/CellLines/CNVdata"
COUNT=0
for SAMPLE in ${clTagDir[@]}; do
INPUT="${clinTagDir[${COUNT}]}"
INNAME=${INPUT##*/}
normalizeTagDirByCopyNumber.pl ${SAMPLE} -cnv "${CNVPATH}/${INNAME}.sam_CNVs" -remove
normalizeTagDirByCopyNumber.pl ${INPUT} -cnv "${CNVPATH}/${INNAME}.sam_CNVs" -remove
COUNT=$((COUNT+=1))
Done

declare -a clTagDir=("${CHIPDIRCL}/THP1/THP1_PMAVD3_PU1" "${CHIPDIRCL}/THP1/THP1_PU1")
declare -a clinTagDir=("${INDIRCL}/THP1/THP1_PMAVD3_chrInput" "${INDIRCL}/THP1/THP1_chrInput")
CNVPATH="/misc/data/processedData/mapping/DNA/hg19/Input/CellLines/CNVdata"
COUNT=0
for SAMPLE in ${clTagDir[@]}; do
INPUT="${clinTagDir[${COUNT}]}"
# normalizeTagDirByCopyNumber.pl ${SAMPLE} -cnv "${CNVPATH}/THP1_genInput_708.sam_CNVs"
-remove
normalizeTagDirByCopyNumber.pl ${INPUT} -cnv "${CNVPATH}/THP1_genInput_708.sam_CNVs"
-remove
COUNT=$((COUNT+=1))
done

declare -a clTagDir=("${CHIPDIRPD}/DOHH2_PU1" "${CHIPDIRPD}/OCILY7_PU1"
"${CHIPDIRPD}/H929_PU1")
declare -a clinTagDir=("${INDIRPD}/DOHH2_Input" "${INDIRPD}/OCILY7_Input"
"${INDIRPD}/H929_Input")
CNVPATH="/misc/data/processedData/mapping/DNA/hg19/Input/publishedData/CNVdata"
COUNT=0
for SAMPLE in ${clTagDir[@]}; do
INPUT="${clinTagDir[${COUNT}]}"
INNAME=${INPUT##*/}
normalizeTagDirByCopyNumber.pl ${SAMPLE} -cnv "${CNVPATH}/${INNAME}.sam_CNVs" -remove
normalizeTagDirByCopyNumber.pl ${INPUT} -cnv "${CNVPATH}/${INNAME}.sam_CNVs" -remove
COUNT=$((COUNT+=1))
done

declare -a clTagDir=("${CHIPDIRPD}/RS411_Dex1h_PU1" "${CHIPDIRPD}/RS411_NoStim_PU1")
declare -a clinTagDir=("${INDIRPD}/RS411_Dex1h_Input" "${INDIRPD}/RS411_NoStim_Input")
CNVPATH="/misc/data/processedData/mapping/DNA/hg19/Input/publishedData/CNVdata"
COUNT=0
for SAMPLE in ${clTagDir[@]}; do
INPUT="${clinTagDir[${COUNT}]}"
# normalizeTagDirByCopyNumber.pl ${SAMPLE} -cnv "${CNVPATH}/RS411_comb_chrInput.sam_CNVs"
-remove
normalizeTagDirByCopyNumber.pl ${INPUT} -cnv "${CNVPATH}/RS411_comb_chrInput.sam_CNVs"
-remove
COUNT=$((COUNT+=1))
done

declare -a clTagDir=("${CHIPDIRPD}/K562_PU1_R2.2" "${CHIPDIRPD}/K562_PU1_R2.1"
"${CHIPDIRPD}/K562_PU1_R1")
declare -a clinTagDir=("${INDIRPD}/K562_R2.2_chrIgG" "${INDIRPD}/K562_R2.1_chrIgG"
"${INDIRPD}/K562_R1_chrInput")
CNVPATH="/misc/data/processedData/mapping/DNA/hg19/Input/CellLines/CNVdata"
COUNT=0
```

```
for SAMPLE in ${clTagDir[@]}; do
INPUT="${clinTagDir[${COUNT}]}"
# normalizeTagDirByCopyNumber.pl ${SAMPLE} -cnv "${CNVPATH}/K562_genInput_619.sam_CNVs"
-remove
normalizeTagDirByCopyNumber.pl ${INPUT} -cnv "${CNVPATH}/K562_genInput_619.sam_CNVs"
-remove
COUNT=$((COUNT+=1))
Done

# remove scaffolds from tagDirs
declare -a normTagDir=("${CHIPDIRPD}/HPC_PU1" "${CHIPDIRPD}/GM12878_PU1"
"${CHIPDIRBC}/CD15_PU1_R2" "${CHIPDIRBC}/CD15_PU1_R1" "${CHIPDIRBC}/BDMC_PU1_R2"
"${CHIPDIRBC}/BDMC_PU1_R1" "${CHIPDIRBC}/CD19_PU1_R2" "${CHIPDIRBC}/CD19_PU1_R1"
"${CHIPDIRMOMACDC}/ncMO_PU1_CS_R2" "${CHIPDIRMOMACDC}/ncMO_PU1_CS_R1"
"${CHIPDIRMOMACDC}/cMO_PU1_CS_R2" "${CHIPDIRMOMACDC}/cMO_PU1_CS_R1"
"${CHIPDIRMOMACDC}/MO_PU1_SS_Lib61_R1" "${CHIPDIRMOMACDC}/MO_PU1_SS_Lib55_R2"
"${CHIPDIRMOMACDC}/DCd7_PU1_SS_S8_R2" "${CHIPDIRMOMACDC}/DCd7_PU1_SS_Lib60_R1"
"${CHIPDIRMOMACDC}/DC66h_PU1_SS_S7_R2" "${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib63_R2"
"${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib59_R1" "${CHIPDIRMOMACDC}/DC27h_PU1_SS_Lib62_R1"
"${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib56_R2" "${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib58_R1"
"${CHIPDIRMOMACDC}/ChIP_AK10_DC_PU1" "${CHIPDIRMOMACDC}/ChIP_AK19_DC_PU1"
"${CHIPDIRMOMACDC}/ChIP_AR07_MAC_PU1" "${CHIPDIRMOMACDC}/MO_PU.1_woDSG_Ch24"
"${CHIPDIRMOMACDC}/MAC_PU.1_woDSG_Ch28" "${CHIPDIRMOMACDC}/MAC_PU.1_Ch35"
"${CHIPDIRMOMACDC}/MO_4h_PU.1_Ch41" "${CHIPDIRMOMACDC}/MO_LPS_4h_PU.1_Ch43"
"${CHIPDIRMOMACDC}/MO18h_PU.1_Ch17" "${CHIPDIRPD}/MO_PU1" "${CHIPDIRPD}/MAC_PU1")
for DIR in ${normTagDir[@]}; do
normalizeTagDirByCopyNumber.pl ${DIR} -remove
done

declare -a inTagDir=("${INDIRPD}/HPC_chrInput" "${INDIRPD}/GM12878_Input"
"${INDIRBC}/CD15_R2_chrInput" "${INDIRBC}/CD15_R1_chrInput"
"${INDIRBC}/BDMC_PU1_R2_chrInput" "${INDIRBC}/BDMC_PU1_R1_chrInput"
"${INDIRBC}/CD8_R2_chrInput" "${INDIRBC}/CD19_R1_chrInput"
"${INDIRMOMACDC}/MOsub_R2_CS_chrInput" "${INDIRMOMACDC}/MOsub_R1_CS_chrInput"
"${INDIRMOMACDC}/MOsub_R2_CS_chrInput" "${INDIRMOMACDC}/MOsub_R1_CS_chrInput"
"${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput" "${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput"
"${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput" "${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput"
"${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput" "${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput"
"${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput" "${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput"
"${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput" "${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput"
"${INDIRMOMACDC}/DC_d4_AK20_chrFlag" "${INDIRMOMACDC}/DC_d4_AK18_chrFlag"
"${INDIRMOMACDC}/ChIP_AR04_MAC_chrInput" "${INDIRMOMACDC}/MO18h_Ch16_chrInput"
"${INDIRMOMACDC}/MO18h_Ch16_chrInput" "${INDIRMOMACDC}/MO18h_Ch16_chrInput"
"${INDIRMOMACDC}/MO18h_Ch16_chrInput" "${INDIRMOMACDC}/MO18h_Ch16_chrInput"
"${INDIRMOMACDC}/MO18h_Ch16_chrInput" "${INDIRPD}/MOMAC_chrIgG" "${INDIRPD}/MOMAC_chrIgG")
for IN in ${inTagDir[@]}; do
normalizeTagDirByCopyNumber.pl ${IN} -remove
done

# merge tag directories
makeTagDirectory ${CHIPDIRPD}/K562_PU1_CNVnormRefChr -d
"${CHIPDIRPD}/K562_PU1_R2.2_CNVnormRefChr" "${CHIPDIRPD}/K562_PU1_R2.1_CNVnormRefChr"
"${CHIPDIRPD}/K562_PU1_R1_CNVnormRefChr" -genome hg19 -checkGC
makeTagDirectory ${CHIPDIRPD}/RS411_PU1_CNVnormRefChr -d
"${CHIPDIRPD}/RS411_NoStim_PU1_CNVnormRefChr" "${CHIPDIRPD}/RS411_Dex1h_PU1_CNVnormRefChr"
-genome hg19 -checkGC
makeTagDirectory ${CHIPDIRBC}/BDMC_PU1_refChr -d "${CHIPDIRBC}/BDMC_PU1_R1_refChr"
"${CHIPDIRBC}/BDMC_PU1_R2_refChr" -genome hg19 -checkGC
makeTagDirectory ${CHIPDIRBC}/CD19_PU1_refChr -d "${CHIPDIRBC}/CD19_PU1_R1_refChr"
"${CHIPDIRBC}/CD19_PU1_R2_refChr" -genome hg19 -checkGC
makeTagDirectory ${CHIPDIRBC}/CD15_PU1_refChr -d "${CHIPDIRBC}/CD15_PU1_R1_refChr"
"${CHIPDIRBC}/CD15_PU1_R2_refChr" -genome hg19 -checkGC
makeTagDirectory ${CHIPDIRMOMACDC}/MAC_PU1_refChr -d "${CHIPDIRPD}/MAC_PU1_refChr"
"${CHIPDIRMOMACDC}/MAC_PU.1_woDSG_Ch28_refChr" "${CHIPDIRMOMACDC}/MAC_PU.1_Ch35_refChr"
"${CHIPDIRMOMACDC}/ChIP_AR07_MAC_PU1_refChr" -genome hg19 -checkGC
makeTagDirectory ${CHIPDIRMOMACDC}/DC_PU1_refChr -d
"${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib58_R1_refChr"
"${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib56_R2_refChr"
"${CHIPDIRMOMACDC}/DC27h_PU1_SS_Lib62_R1_refChr"
"${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib59_R1_refChr"
"${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib63_R2_refChr"
"${CHIPDIRMOMACDC}/DC66h_PU1_SS_S7_R2_refChr" "${CHIPDIRMOMACDC}/ChIP_AK10_DC_PU1_refChr"
"${CHIPDIRMOMACDC}/ChIP_AK19_DC_PU1_refChr" "${CHIPDIRMOMACDC}/DCd7_PU1_SS_Lib60_R1_refChr"
"${CHIPDIRMOMACDC}/DCd7_PU1_SS_S8_R2_refChr" -genome hg19 -checkGC
makeTagDirectory ${CHIPDIRMOMACDC}/MO_PU1_refChr -d "${CHIPDIRPD}/MO_PU1_refChr"
"${CHIPDIRMOMACDC}/MO_PU1_SS_Lib61_R1_refChr" "${CHIPDIRMOMACDC}/MO_PU1_SS_Lib55_R2_refChr"
"${CHIPDIRMOMACDC}/MO_PU.1_woDSG_Ch24_refChr" "${CHIPDIRMOMACDC}/cMO_PU1_CS_R1_refChr"
"${CHIPDIRMOMACDC}/cMO_PU1_CS_R2_refChr" "${CHIPDIRMOMACDC}/ncMO_PU1_CS_R1_refChr"
```

```
"${CHIPDIRMOMACDC}/ncMO_PU1_CS_R2_refChr" -genome hg19 -checkGC

# BigWigs for normalized TagDirs
# generating individual bigwigs
declare -a clTagDirN=("${CHIPDIRCL}/other/EM3_PU1_CNVnormRefChr"
"${CHIPDIRCL}/other/KG1_PU1_CNVnormRefChr" "${CHIPDIRCL}/other/ML2_PU1_CNVnormRefChr"
"${CHIPDIRCL}/other/U937_PU1_CNVnormRefChr" "${CHIPDIRCL}/other/NB4_PU1_CNVnormRefChr")
for SAMPLE in ${clTagDirN[@]}; do
SAMPLENAME=${SAMPLE##*/}
makeUCSCfile ${SAMPLE} -bigWig $CHROMSIZES -o $BIGWIGDIRCL/other/$SAMPLENAME.bigwig
done

declare -a clTagDirN=("${CHIPDIRCL}/THP1/THP1_PMAVD3_PU1_CNVnormRefChr"
"${CHIPDIRCL}/THP1/THP1_PU1_CNVnormRefChr")
for SAMPLE in ${clTagDirN[@]}; do
SAMPLENAME=${SAMPLE##*/}
makeUCSCfile ${SAMPLE} -bigWig $CHROMSIZES -o $BIGWIGDIRCL/THP1/$SAMPLENAME.bigwig
done

declare -a clTagDirN=("${CHIPDIRPD}/DOHH2_PU1_CNVnormRefChr"
"${CHIPDIRPD}/OCILY7_PU1_CNVnormRefChr" "${CHIPDIRPD}/H929_PU1_CNVnormRefChr"
"${CHIPDIRPD}/RS411_Dex1h_PU1_CNVnormRefChr" "${CHIPDIRPD}/RS411_NoStim_PU1_CNVnormRefChr"
"${CHIPDIRPD}/K562_PU1_R2.2_CNVnormRefChr" "${CHIPDIRPD}/K562_PU1_R2.1_CNVnormRefChr"
"${CHIPDIRPD}/K562_PU1_R1_CNVnormRefChr")
for SAMPLE in ${clTagDirN[@]}; do
SAMPLENAME=${SAMPLE##*/}
makeUCSCfile ${SAMPLE} -bigWig $CHROMSIZES -o $BIGWIGDIRPD/$SAMPLENAME.bigwig
done

declare -a normTagDir=("${CHIPDIRPD}/HPC_PU1" "${CHIPDIRPD}/GM12878_PU1"
"${CHIPDIRPD}/MO_PU1" "${CHIPDIRPD}/MAC_PU1")
for SAMPLE in ${normTagDir[@]}; do
SAMPLENAME=${SAMPLE##*/}
makeUCSCfile "${SAMPLE}_refChr" -bigWig $CHROMSIZES -o
"${BIGWIGDIRPD}/${SAMPLENAME}_refChr.bigwig"
done

declare -a normTagDir=("${CHIPDIRBC}/CD15_PU1_R2" "${CHIPDIRBC}/CD15_PU1_R1"
"${CHIPDIRBC}/BDMC_PU1_R2" "${CHIPDIRBC}/BDMC_PU1_R1" "${CHIPDIRBC}/CD19_PU1_R2"
"${CHIPDIRBC}/CD19_PU1_R1")
for SAMPLE in ${normTagDir[@]}; do
SAMPLENAME=${SAMPLE##*/}
makeUCSCfile "${SAMPLE}_refChr" -bigWig $CHROMSIZES -o
"${BIGWIGDIRBC}/${SAMPLENAME}_refChr.bigwig"
Done

declare -a normTagDir=("${CHIPDIRMOMACDC}/ncMO_PU1_CS_R2" "${CHIPDIRMOMACDC}/ncMO_PU1_CS_R1"
"${CHIPDIRMOMACDC}/cMO_PU1_CS_R2" "${CHIPDIRMOMACDC}/cMO_PU1_CS_R1"
"${CHIPDIRMOMACDC}/MO_PU1_SS_Lib61_R1" "${CHIPDIRMOMACDC}/MO_PU1_SS_Lib55_R2"
"${CHIPDIRMOMACDC}/DCd7_PU1_SS_S8_R2" "${CHIPDIRMOMACDC}/DCd7_PU1_SS_Lib60_R1"
"${CHIPDIRMOMACDC}/DC66h_PU1_SS_S7_R2" "${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib63_R2"
"${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib59_R1" "${CHIPDIRMOMACDC}/DC27h_PU1_SS_Lib62_R1"
"${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib56_R2" "${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib58_R1"
"${CHIPDIRMOMACDC}/ChIP_AK10_DC_PU1" "${CHIPDIRMOMACDC}/ChIP_AK19_DC_PU1"
"${CHIPDIRMOMACDC}/ChIP_AR07_MAC_PU1" "${CHIPDIRMOMACDC}/MO_PU.1_woDSG_Ch24"
"${CHIPDIRMOMACDC}/MAC_PU.1_woDSG_Ch28" "${CHIPDIRMOMACDC}/MAC_PU.1_Ch35"
"${CHIPDIRMOMACDC}/MO_4h_PU.1_Ch41" "${CHIPDIRMOMACDC}/MO_LPS_4h_PU.1_Ch43"
"${CHIPDIRMOMACDC}/MO18h_PU.1_Ch17")
for SAMPLE in ${normTagDir[@]}; do
SAMPLENAME=${SAMPLE##*/}
makeUCSCfile "${SAMPLE}_refChr" -bigWig $CHROMSIZES -o
"${BIGWIGDIRMOMACDC}/${SAMPLENAME}_refChr.bigwig"
done

# average bigwigs from replicates
myAverageBigWig.pl -bw $BIGWIGDIRBC/CD15_PU1_R1_refChr.bigwig \
$BIGWIGDIRBC/CD15_PU1_R1_refChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIRBC/CD15_PU1_ave_refChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIRBC/CD19_PU1_R1_refChr.bigwig \
$BIGWIGDIRBC/CD19_PU1_R1_refChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIRBC/CD19_PU1_ave_refChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIRBC/BDMC_PU1_R1_refChr.bigwig \
$BIGWIGDIRBC/BDMC_PU1_R1_refChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIRBC/BDMC_PU1_ave_refChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIRPD/K562_PU1_R2.2_CNVnormRefChr.bigwig \
$BIGWIGDIRPD/K562_PU1_R2.1_CNVnormRefChr.bigwig \
$BIGWIGDIRPD/K562_PU1_R1_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIRPD/K562_PU1_ave_refChr.bigwig
```

```
myAverageBigWig.pl -bw $BIGWIGDIRMOMACDC/ncMO_PU1_CS_R1_refChr.bigwig \
$BIGWIGDIRMOMACDC/ncMO_PU1_CS_R2_refChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIRMOMACDC/ncMO_PU1_ave_refChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIRMOMACDC/cMO_PU1_CS_R1_refChr.bigwig \
$BIGWIGDIRMOMACDC/cMO_PU1_CS_R2_refChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIRMOMACDC/cMO_PU1_ave_refChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIRMOMACDC/MO_PU1_SS_Lib61_R1_refChr.bigwig \
$BIGWIGDIRMOMACDC/MO_PU1_SS_Lib55_R2_refChr.bigwig \
$BIGWIGDIRMOMACDC/MO_PU.1_woDSG_Ch24_refChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIRMOMACDC/MO_PU1_ave_refChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIRMOMACDC/DCd7_PU1_SS_S8_R2_refChr.bigwig \
$BIGWIGDIRMOMACDC/DCd7_PU1_SS_Lib60_R1_refChr.bigwig \
$BIGWIGDIRMOMACDC/ChIP_AK10_DC_PU1_refChr.bigwig \
$BIGWIGDIRMOMACDC/ChIP_AK19_DC_PU1_refChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIRMOMACDC/DCd7_PU1_ave_refChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIRMOMACDC/DC42h_PU1_SS_Lib63_R2_refChr.bigwig \
$BIGWIGDIRMOMACDC/DC42h_PU1_SS_Lib59_R1_refChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIRMOMACDC/DC42h_PU1_ave_refChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIRMOMACDC/DC18h_PU1_SS_Lib56_R2_refChr.bigwig \
$BIGWIGDIRMOMACDC/DC18h_PU1_SS_Lib58_R1_refChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIRMOMACDC/DC18h_PU1_ave_refChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIRMOMACDC/ChIP_AR07_MAC_PU1_refChr.bigwig \
$BIGWIGDIRMOMACDC/MAC_PU.1_woDSG_Ch28_refChr.bigwig \
$BIGWIGDIRMOMACDC/MAC_PU.1_Ch35_refChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIRMOMACDC/MAC_PU1_ave_refChr.bigwig

# peak finding
mkdir -p ${PEAKDIR}
mkdir ${FIGURESDIR}

declare -a CHIPS=("${CHIPDIRPD}/HPC_PU1_refChr" "${CHIPDIRPD}/GM12878_PU1_refChr"
"${CHIPDIRBC}/CD15_PU1_R2_refChr" "${CHIPDIRBC}/CD15_PU1_R1_refChr"
"${CHIPDIRBC}/BDMC_PU1_R2_refChr" "${CHIPDIRBC}/BDMC_PU1_R1_refChr"
"${CHIPDIRBC}/CD19_PU1_R2_refChr" "${CHIPDIRBC}/CD19_PU1_R1_refChr"
"${CHIPDIRMOMACDC}/ncMO_PU1_CS_R2_refChr" "${CHIPDIRMOMACDC}/ncMO_PU1_CS_R1_refChr"
"${CHIPDIRMOMACDC}/cMO_PU1_CS_R2_refChr" "${CHIPDIRMOMACDC}/cMO_PU1_CS_R1_refChr"
"${CHIPDIRMOMACDC}/MO_PU1_SS_Lib61_R1_refChr" "${CHIPDIRMOMACDC}/MO_PU1_SS_Lib55_R2_refChr"
"${CHIPDIRMOMACDC}/DCd7_PU1_SS_S8_R2_refChr" "${CHIPDIRMOMACDC}/DCd7_PU1_SS_Lib60_R1_refChr"
"${CHIPDIRMOMACDC}/DC66h_PU1_SS_S7_R2_refChr"
"${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib63_R2_refChr"
"${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib59_R1_refChr"
"${CHIPDIRMOMACDC}/DC27h_PU1_SS_Lib62_R1_refChr"
"${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib56_R2_refChr"
"${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib58_R1_refChr" "${CHIPDIRMOMACDC}/ChIP_AK10_DC_PU1_refChr"
"${CHIPDIRMOMACDC}/ChIP_AK19_DC_PU1_refChr" "${CHIPDIRMOMACDC}/ChIP_AR07_MAC_PU1_refChr"
"${CHIPDIRMOMACDC}/MO_PU.1_woDSG_Ch24_refChr" "${CHIPDIRMOMACDC}/MAC_PU.1_woDSG_Ch28_refChr"
"${CHIPDIRMOMACDC}/MAC_PU.1_Ch35_refChr" "${CHIPDIRMOMACDC}/MO_4h_PU.1_Ch41_refChr"
"${CHIPDIRMOMACDC}/MO_LPS_4h_PU.1_Ch43_refChr" "${CHIPDIRMOMACDC}/MO18h_PU.1_Ch17_refChr"
"${CHIPDIRPD}/MO_PU1_refChr" "${CHIPDIRPD}/MAC_PU1_refChr"
"${CHIPDIRCL}/other/EM3_PU1_CNVnormRefChr" "${CHIPDIRCL}/other/KG1_PU1_CNVnormRefChr"
"${CHIPDIRCL}/other/ML2_PU1_CNVnormRefChr" "${CHIPDIRCL}/other/U937_PU1_CNVnormRefChr"
"${CHIPDIRCL}/other/NB4_PU1_CNVnormRefChr" "${CHIPDIRCL}/THP1/THP1_PMAVD3_PU1_CNVnormRefChr"
"${CHIPDIRCL}/THP1/THP1_PU1_CNVnormRefChr" "${CHIPDIRPD}/DOHH2_PU1_CNVnormRefChr"
"${CHIPDIRPD}/OCILY7_PU1_CNVnormRefChr" "${CHIPDIRPD}/H929_PU1_CNVnormRefChr"
"${CHIPDIRPD}/K562_PU1_R2.2_CNVnormRefChr" "${CHIPDIRPD}/K562_PU1_R2.1_CNVnormRefChr"
"${CHIPDIRPD}/K562_PU1_R1_CNVnormRefChr" "${CHIPDIRPD}/RS411_Dex1h_PU1_CNVnormRefChr"
"${CHIPDIRPD}/RS411_NoStim_PU1_CNVnormRefChr")

declare -a INPUTS=("${INDIRPD}/HPC_chrInput_refChr" "${INDIRPD}/GM12878_Input_refChr"
"${INDIRBC}/CD15_R2_chrInput_refChr" "${INDIRBC}/CD15_R1_chrInput_refChr"
"${INDIRBC}/BDMC_PU1_R2_chrInput_refChr" "${INDIRBC}/BDMC_PU1_R1_chrInput_refChr"
"${INDIRBC}/CD8_R2_chrInput_refChr" "${INDIRBC}/CD19_R1_chrInput_refChr"
"${INDIRMOMACDC}/MOsub_R2_CS_chrInput_refChr" "${INDIRMOMACDC}/MOsub_R1_CS_chrInput_refChr"
"${INDIRMOMACDC}/MOsub_R2_CS_chrInput_refChr" "${INDIRMOMACDC}/MOsub_R1_CS_chrInput_refChr"
"${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput_refChr"
"${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput_refChr"
"${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput_refChr"
"${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput_refChr"
"${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput_refChr"
"${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput_refChr"
"${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput_refChr"
"${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput_refChr"
"${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput_refChr"
"${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput_refChr"
"${INDIRMOMACDC}/DC_d4_AK20_chrFlag_refChr" "${INDIRMOMACDC}/DC_d4_AK18_chrFlag_refChr"
"${INDIRMOMACDC}/ChIP_AR04_MAC_chrInput_refChr" "${INDIRMOMACDC}/MO18h_Ch16_chrInput_refChr"
"${INDIRMOMACDC}/MO18h_Ch16_chrInput_refChr" "${INDIRMOMACDC}/MO18h_Ch16_chrInput_refChr"
"${INDIRMOMACDC}/MO18h_Ch16_chrInput_refChr" "${INDIRMOMACDC}/MO18h_Ch16_chrInput_refChr"
"${INDIRMOMACDC}/MO18h_Ch16_chrInput_refChr" "${INDIRPD}/MOMAC_chrIgG_refChr"
"${INDIRPD}/MOMAC_chrIgG_refChr" "${INDIRCL}/EM3_genInput_621_CNVnormRefChr"
"${INDIRCL}/KG1a_genInput_622_CNVnormRefChr" "${INDIRCL}/ML2_genInput_623_CNVnormRefChr"
"${INDIRCL}/U937_genInput_700_CNVnormRefChr" "${INDIRCL}/NB4_genInput_624_CNVnormRefChr"
"${INDIRCL}/THP1/THP1_PMAVD3_chrInput_CNVnormRefChr"
"${INDIRCL}/THP1/THP1_chrInput_CNVnormRefChr" "${INDIRPD}/DOHH2_Input_CNVnormRefChr"
"${INDIRPD}/OCILY7_Input_CNVnormRefChr" "${INDIRPD}/H929_Input_CNVnormRefChr"
"${INDIRPD}/K562_R2.2_chrIgG_CNVnormRefChr" "${INDIRPD}/K562_R2.1_chrIgG_CNVnormRefChr"
"${INDIRPD}/K562_R1_chrInput_CNVnormRefChr" "${INDIRPD}/RS411_Dex1h_Input_CNVnormRefChr"
"${INDIRPD}/RS411_NoStim_Input_CNVnormRefChr")

declare -a NAMES=("HPC" "GM12878" "CD15_R2" "CD15_R1" "BDMC_R2" "BDMC_R1" "CD19_R2"
"CD19_R1" "ncMO_R2" "ncMO_R1" "cMO_R2" "cMO_R1" "MO_R1" "MO_R2" "DCd7_R2" "DCd7_R1"
"DC66h_R2" "DC42h_R2" "DC42h_R1" "DC27h_R1" "DC18h_R2" "DC18h_R1" "DC96h_R3" "DC96h_R4"
"MAC_R3" "MO_R4" "MAC_R4.1" "MAC_R4.2" "MO4h_R4" "MO4h_LPS_R4" "MO18h_R5" "MO_R0" "MAC_R0"
"EM3" "KG1" "ML2" "U937" "NB4" "THP1_PMAVD3" "THP1" "DOHH2" "OCILY7" "H929" "K562_R2.2"
"K562_R2.1" "K562_R1" "RS411_Dex1h" "RS411_NoStim")

_DATE=$(date +%s)
COUNT=0
for CHIP in ${CHIPS[@]}; do
NAME="${NAMES[${COUNT}]}"
INPUT="${INPUTS[${COUNT}]}"
cat >"${TMPDIR}/peaks.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findPeaks ${CHIP} -i ${INPUT} -style factor -fdr 0.00001 -o
${PEAKDIR}/${NAME}.factor.fdr05.peaks.txt
myFilterFile.pl ${PEAKDIR}/${NAME}.factor.fdr05.peaks.txt -column 6 -lowerlimit 15 >
${PEAKDIR}/${NAME}.factor.fdr05.ntag15.peaks.txt
pos2bed.pl ${PEAKDIR}/${NAME}.factor.fdr05.ntag15.peaks.txt >
${TMPDIR}/tmp.${NAME}.factor.fdr05.ntag15.peaks.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.${NAME}.factor.fdr05.ntag15.peaks.bed -b
$BLACKLIST_HG19 -v > ${TMPDIR}/tmp.${NAME}.factor.fdr05.ntag15.peaks.black.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.${NAME}.factor.fdr05.ntag15.peaks.black.bed -g hg19
-f 0.8 -s 50
pos2bed.pl ${TMPDIR}/tmp.${NAME}.factor.fdr05.ntag15.peaks.black.mapScoreFiltered.txt >
${PEAKDIR}/${NAME}.factor.fdr05.ntag15.filtered.pos.bed
bed2pos.pl ${PEAKDIR}/${NAME}.factor.fdr05.ntag15.filtered.pos.bed >
${PEAKDIR}/${NAME}.factor.fdr05.ntag15.filtered.pos.txt
EOF
COUNT=$((COUNT+=1))
chmod 750 "${TMPDIR}/peaks.${NAME}.${_DATE}.sh"
echo "finding peaks for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/peaks.${NAME}.${_DATE}.sh"
done


# loop to check when screen sessions are done
#--------------------------------------------
for NAME in "${NAMES[@]}" ; do
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

# merging peaks
PEAKLIST=""
for NAME in ${NAMES[@]}; do
PEAKLIST="${PEAKLIST} ${PEAKDIR}/${NAME}.factor.fdr05.ntag15.filtered.pos.txt"
done
mergePeaks -d 100 ${PEAKLIST} -venn "${PEAKDIR}/allCelltypes.factor.venn.txt"
>"${PEAKDIR}/allCelltypes.factor.fdr05.ntag15.merged.txt"
sed -e
's%/misc/data/analysis/project_PU1/allCellTypes/peaks/%%g;s%.factor.fdr05.ntag15.filtered.pos.
txt%%g;s%|%.%g' \
"${PEAKDIR}/allCelltypes.factor.fdr05.ntag15.merged.txt" | cut -f1-8
>${PEAKDIR}/allCelltypes.factor.fdr05.ntag15.reformated.txt
wc -l ${PEAKDIR}/allCelltypes.factor.fdr05.ntag15.reformated.txt

# rlog normalization of tag counts in peaks (no background subtraction)
declare -a CHIPS=("${CHIPDIRPD}/K562_PU1_R2.2" "${CHIPDIRPD}/K562_PU1_R2.1"
"${CHIPDIRPD}/K562_PU1_R1" "${CHIPDIRPD}/HPC_PU1" "${CHIPDIRPD}/RS411_Dex1h_PU1"
"${CHIPDIRPD}/RS411_NoStim_PU1" "${CHIPDIRPD}/GM12878_PU1" "${CHIPDIRPD}/DOHH2_PU1"
"${CHIPDIRPD}/OCILY7_PU1" "${CHIPDIRPD}/H929_PU1" "${CHIPDIRBC}/CD15_PU1_R2"
"${CHIPDIRBC}/CD15_PU1_R1" "${CHIPDIRBC}/BDMC_PU1_R2" "${CHIPDIRBC}/BDMC_PU1_R1"
"${CHIPDIRBC}/CD19_PU1_R2" "${CHIPDIRBC}/CD19_PU1_R1" "${CHIPDIRCL}/THP1/THP1_PMAVD3_PU1"
"${CHIPDIRCL}/THP1/THP1_PU1" "${CHIPDIRCL}/other/EM3_PU1" "${CHIPDIRCL}/other/KG1_PU1"
"${CHIPDIRCL}/other/ML2_PU1" "${CHIPDIRCL}/other/U937_PU1" "${CHIPDIRCL}/other/NB4_PU1"
"${CHIPDIRMOMACDC}/ncMO_PU1_CS_R2" "${CHIPDIRMOMACDC}/ncMO_PU1_CS_R1"
"${CHIPDIRMOMACDC}/cMO_PU1_CS_R2" "${CHIPDIRMOMACDC}/cMO_PU1_CS_R1"
"${CHIPDIRMOMACDC}/MO_PU1_SS_Lib61_R1" "${CHIPDIRMOMACDC}/MO_PU1_SS_Lib55_R2"
"${CHIPDIRMOMACDC}/DCd7_PU1_SS_S8_R2" "${CHIPDIRMOMACDC}/DCd7_PU1_SS_Lib60_R1"
"${CHIPDIRMOMACDC}/DC66h_PU1_SS_S7_R2" "${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib63_R2"
"${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib59_R1" "${CHIPDIRMOMACDC}/DC27h_PU1_SS_Lib62_R1"
"${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib56_R2" "${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib58_R1"
"${CHIPDIRMOMACDC}/ChIP_AK10_DC_PU1" "${CHIPDIRMOMACDC}/ChIP_AK19_DC_PU1"
"${CHIPDIRMOMACDC}/ChIP_AR07_MAC_PU1" "${CHIPDIRMOMACDC}/MO_PU.1_woDSG_Ch24"
"${CHIPDIRMOMACDC}/MAC_PU.1_woDSG_Ch28" "${CHIPDIRMOMACDC}/MAC_PU.1_Ch35"
"${CHIPDIRMOMACDC}/MO_4h_PU.1_Ch41" "${CHIPDIRMOMACDC}/MO_LPS_4h_PU.1_Ch43"
"${CHIPDIRMOMACDC}/MO18h_PU.1_Ch17" "${CHIPDIRPD}/MO_PU1" "${CHIPDIRPD}/MAC_PU1")

declare -a INPUTS=("${INDIRPD}/K562_R2.2_chrIgG" "${INDIRPD}/K562_R2.1_chrIgG"
"${INDIRPD}/K562_R1_chrInput" "${INDIRPD}/HPC_chrInput" "${INDIRPD}/RS411_Dex1h_Input"
"${INDIRPD}/RS411_NoStim_Input" "${INDIRPD}/GM12878_Input" "${INDIRPD}/DOHH2_Input"
"${INDIRPD}/OCILY7_Input" "${INDIRPD}/H929_Input" "${INDIRBC}/CD15_R2_chrInput"
"${INDIRBC}/CD15_R1_chrInput" "${INDIRBC}/BDMC_PU1_R2_chrInput"
"${INDIRBC}/BDMC_PU1_R1_chrInput" "${INDIRBC}/CD8_R2_chrInput" "${INDIRBC}/CD19_R1_chrInput"
"${INDIRCL}/THP1/THP1_PMAVD3_chrInput" "${INDIRCL}/THP1/THP1_chrInput"
"${INDIRCL}/EM3_genInput_621" "${INDIRCL}/KG1a_genInput_622" "${INDIRCL}/ML2_genInput_623"
"${INDIRCL}/U937_genInput_700" "${INDIRCL}/NB4_genInput_624"
"${INDIRMOMACDC}/MOsub_R2_CS_chrInput" "${INDIRMOMACDC}/MOsub_R1_CS_chrInput"
"${INDIRMOMACDC}/MOsub_R2_CS_chrInput" "${INDIRMOMACDC}/MOsub_R1_CS_chrInput"
"${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput" "${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput"
"${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput" "${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput"
"${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput" "${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput"
"${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput" "${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput"
"${INDIRMOMACDC}/DC18h_SS_S9_R2_chrInput" "${INDIRMOMACDC}/DC18h_SS_Lib57_R1_chrInput"
"${INDIRMOMACDC}/DC_d4_AK20_chrFlag" "${INDIRMOMACDC}/DC_d4_AK18_chrFlag"
"${INDIRMOMACDC}/ChIP_AR04_MAC_chrInput" "${INDIRMOMACDC}/MO18h_Ch16_chrInput"
"${INDIRMOMACDC}/MO18h_Ch16_chrInput" "${INDIRMOMACDC}/MO18h_Ch16_chrInput"
"${INDIRMOMACDC}/MO18h_Ch16_chrInput" "${INDIRMOMACDC}/MO18h_Ch16_chrInput"
"${INDIRMOMACDC}/MO18h_Ch16_chrInput" "${INDIRPD}/MOMAC_chrIgG" "${INDIRPD}/MOMAC_chrIgG")

declare -a NAMES=("K562_R2.2" "K562_R2.1" "K562_R1" "HPC" "RS411_Dex1h" "RS411_NoStim"
"GM12878" "DOHH2" "OCILY7" "H929" "CD15_R2" "CD15_R1" "BDMC_R2" "BDMC_R1" "CD19_R2"
"CD19_R1" "THP1_PMAVD3" "THP1" "EM3" "KG1" "ML2" "U937" "NB4" "ncMO_R2" "ncMO_R1" "cMO_R2"
"cMO_R1" "MO_R1" "MO_R2" "DCd7_R2" "DCd7_R1" "DC66h_R2" "DC42h_R2" "DC42h_R1" "DC27h_R1"
"DC18h_R2" "DC18h_R1" "DC96h_R3" "DC96h_R4" "MAC_R3" "MO_R4" "MAC_R4.1" "MAC_R4.2" "MO4h_R4"
"MO4h_LPS_R4" "MO18h_R5" "MO_R0" "MAC_R0")

CHIPDIRS="${CHIPDIRPD}/K562_PU1_R2.2 ${CHIPDIRPD}/K562_PU1_R2.1 ${CHIPDIRPD}/K562_PU1_R1
${CHIPDIRPD}/HPC_PU1 ${CHIPDIRPD}/RS411_Dex1h_PU1 ${CHIPDIRPD}/RS411_NoStim_PU1
${CHIPDIRPD}/GM12878_PU1 ${CHIPDIRPD}/DOHH2_PU1 ${CHIPDIRPD}/OCILY7_PU1
${CHIPDIRPD}/H929_PU1 ${CHIPDIRBC}/CD15_PU1_R2 ${CHIPDIRBC}/CD15_PU1_R1
${CHIPDIRBC}/BDMC_PU1_R2 ${CHIPDIRBC}/BDMC_PU1_R1 ${CHIPDIRBC}/CD19_PU1_R2
${CHIPDIRBC}/CD19_PU1_R1 ${CHIPDIRCL}/THP1/THP1_PMAVD3_PU1 ${CHIPDIRCL}/THP1/THP1_PU1
${CHIPDIRCL}/other/EM3_PU1 ${CHIPDIRCL}/other/KG1_PU1 ${CHIPDIRCL}/other/ML2_PU1
${CHIPDIRCL}/other/U937_PU1 ${CHIPDIRCL}/other/NB4_PU1 ${CHIPDIRMOMACDC}/ncMO_PU1_CS_R2
${CHIPDIRMOMACDC}/ncMO_PU1_CS_R1 ${CHIPDIRMOMACDC}/cMO_PU1_CS_R2
${CHIPDIRMOMACDC}/cMO_PU1_CS_R1 ${CHIPDIRMOMACDC}/MO_PU1_SS_Lib61_R1
${CHIPDIRMOMACDC}/MO_PU1_SS_Lib55_R2 ${CHIPDIRMOMACDC}/DCd7_PU1_SS_S8_R2
${CHIPDIRMOMACDC}/DCd7_PU1_SS_Lib60_R1 ${CHIPDIRMOMACDC}/DC66h_PU1_SS_S7_R2
${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib63_R2 ${CHIPDIRMOMACDC}/DC42h_PU1_SS_Lib59_R1
${CHIPDIRMOMACDC}/DC27h_PU1_SS_Lib62_R1 ${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib56_R2
${CHIPDIRMOMACDC}/DC18h_PU1_SS_Lib58_R1 ${CHIPDIRMOMACDC}/ChIP_AK10_DC_PU1
${CHIPDIRMOMACDC}/ChIP_AK19_DC_PU1 ${CHIPDIRMOMACDC}/ChIP_AR07_MAC_PU1
${CHIPDIRMOMACDC}/MO_PU.1_woDSG_Ch24 ${CHIPDIRMOMACDC}/MAC_PU.1_woDSG_Ch28
${CHIPDIRMOMACDC}/MAC_PU.1_Ch35 ${CHIPDIRMOMACDC}/MO_4h_PU.1_Ch41
${CHIPDIRMOMACDC}/MO_LPS_4h_PU.1_Ch43 ${CHIPDIRMOMACDC}/MO18h_PU.1_Ch17 ${CHIPDIRPD}/MO_PU1
${CHIPDIRPD}/MAC_PU1"

annotatePeaks.pl ${PEAKDIR}/allCelltypes.factor.fdr05.ntag15.reformated.txt hg19 -size 200
-d ${CHIPDIRS} -rlog > ${PEAKDIR}/allCelltypes.ann.rlog.txt -cpu 48
```

```
cut -f 1,20-67 ${PEAKDIR}/allCelltypes.ann.rlog.txt | tail -n +2 >
${PEAKDIR}/tmp.allCelltypes.ann.txt

echo
$'ID\tK562_R2.2\tK562_R2.1\tK562_R1\tHPC\tRS411_Dex1h\tRS411_NoStim\tGM12878\tDOHH2\tOCILY7\tH
929\tCD15_R2\tCD15_R1\tBDMC_R2\tBDMC_R1\tCD19_R2\tCD19_R1\tTHP1_PMAVD3\tTHP1\tEM3\tKG1\tML2\tU
937\tNB4\tncMO_R2\tncMO_R1\tcMO_R2\tcMO_R1\tMO_R1\tMO_R2\tDCd7_R2\tDCd7_R1\tDC66h_R2\tDC42h_R2
\tDC42h_R1\tDC27h_R1\tDC18h_R2\tDC18h_R1\tDC96h_R3\tDC96h_R4\tMAC_R3\tMO_R4\tMAC_R4.1\tMAC_R4.
2\tMO4h_R4\tMO4h_LPS_R4\tMO18h_R5\tMO_R0\tMAC_R0' \
| cat - ${PEAKDIR}/tmp.allCelltypes.ann.txt > ${PEAKDIR}/allCelltypes.ann.rlog.txt
rm ${PEAKDIR}/tmp.allCelltypes.ann.txt

# t-SNE 2D Embedding
cat >"${TMPDIR}/R.PCA.${_DATE}.R" <<EOF
library(ggplot2)
library(ggrepel)
library(Rtsne)
#logcpm <- read.delim("${PEAKDIR}/allCelltypes.ann.rlog.txt", row.names="ID")
celltype <- factor(c("K562_R2.2", "K562_R2.1", "K562_R1", "HPC", "RS411_Dex1h",
"RS411_NoStim", "GM12878", "DOHH2", "OCILY7", "H929", "CD15_R2", "CD15_R1", "BDMC_R2",
"BDMC_R1", "CD19_R2", "CD19_R1", "THP1_PMAVD3", "THP1", "EM3", "KG1", "ML2", "U937", "NB4",
"ncMO_R2", "ncMO_R1", "cMO_R2", "cMO_R1", "MO_R1", "MO_R2", "DCd7_R2", "DCd7_R1",
"DC66h_R2", "DC42h_R2", "DC42h_R1", "DC27h_R1", "DC18h_R2", "DC18h_R1", "DC96h_R3",
"DC96h_R4", "MAC_R3", "MO_R4", "MAC_R4.1", "MAC_R4.2", "MO4h_R4", "MO4h_LPS_R4", "MO18h_R5",
"MO_R0", "MAC_R0"))
colors2 <- c("darkturquoise", "darkturquoise", "darkturquoise", "hotpink", "firebrick",
"firebrick", "firebrick", "firebrick", "firebrick", "firebrick", "darkcyan", "darkcyan",
"darkslategray4", "darkslategray4", "orange2", "orange2", "blue", "blue", "blue", "blue",
"blue", "blue", "blue", "blueviolet", "blueviolet", "blueviolet", "blueviolet",
"blueviolet", "blueviolet", "forestgreen", "forestgreen", "forestgreen", "forestgreen",
"forestgreen", "forestgreen", "forestgreen", "forestgreen", "forestgreen", "forestgreen",
"dodgerblue1", "blueviolet", "dodgerblue1", "dodgerblue1", "slateblue3", "slateblue3",
"slateblue3", "blueviolet", "dodgerblue1")
#mydata <- data.matrix(t(logcpm))
#rtsne_out <- Rtsne(mydata, check_duplicates = FALSE, pca = TRUE, perplexity=6, theta=0.225,
dims=2, max_iter = 10000)
#embedding <- as.data.frame(rtsne_out\$Y)
#write.table(embedding, file = "${PEAKDIR}/allCelltypes.ann.rlog.tSNEembedding.txt", sep =
"\t", col.names=NA, quote=FALSE)
embedding <- read.delim("${EMBEDDING}")
embedding\$Class <- as.factor(celltype)
embedding\$Color <- as.factor(colors2)
p <- ggplot(embedding, aes(x=V1, y=V2, label=celltype)) +
geom_point(size=2, col=embedding\$Color) +
geom_text_repel(aes(label=celltype), col=embedding\$Color, size=1.5,segment.size=0.2,
min.segment.length=0.2, point.padding=.15, segment.alpha=0.5) +
guides(colour = guide_legend(override.aes = list(size=5))) +
xlab("tSNE-X") + ylab("tSNE-Y") +
ggtitle("t-SNE 2D Embedding\nfor all Cell Types") +
#xlim(-200, 200) + ylim(-200, 200) +
theme_light(base_size=8) + theme(plot.title = element_text(size = 10, face = "bold"))
pdf(file="${FIGURESDIR}/tSNE_merged.allCellTypes.ann.rlog.pdf", height=3.5, width=3.3)
plot(p, labels=TRUE)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.PCA.${_DATE}.R"
R < ${TMPDIR}/R.PCA.${_DATE}.R --no-save
rm ${TMPDIR}/R.PCA.${_DATE}.R

# join rlog annotated file with peak calling info
tail -n +2 ${PEAKDIR}/allCelltypes.factor.fdr05.ntag15.reformated.txt > ${TMPDIR}/tmp.1.txt
tail -n +2 ${PEAKDIR}/allCelltypes.ann.rlog.txt > ${TMPDIR}/tmp.2.txt
join -1 1 -2 1 -t $'\t' <(sort -k1,1 ${TMPDIR}/tmp.1.txt) <(sort -k1,1 ${TMPDIR}/tmp.2.txt)
> ${TMPDIR}/tmp.3.txt
sort -k8,8n -k7,7 ${TMPDIR}/tmp.3.txt > ${TMPDIR}/tmp.4.txt
cut -f1,9-56 ${TMPDIR}/tmp.4.txt > ${TMPDIR}/tmp.5.txt
echo
$'ID\tK562_R2.2\tK562_R2.1\tK562_R1\tHPC\tRS411_Dex1h\tRS411_NoStim\tGM12878\tDOHH2\tOCILY7\tH
929\tCD15_R2\tCD15_R1\tBDMC_R2\tBDMC_R1\tCD19_R2\tCD19_R1\tTHP1_PMAVD3\tTHP1\tEM3\tKG1\tML2\tU
937\tNB4\tncMO_R2\tncMO_R1\tcMO_R2\tcMO_R1\tMO_R1\tMO_R2\tDCd7_R2\tDCd7_R1\tDC66h_R2\tDC42h_R2
\tDC42h_R1\tDC27h_R1\tDC18h_R2\tDC18h_R1\tDC96h_R3\tDC96h_R4\tMAC_R3\tMO_R4\tMAC_R4.1\tMAC_R4.
2\tMO4h_R4\tMO4h_LPS_R4\tMO18h_R5\tMO_R0\tMAC_R0' \
| cat - ${TMPDIR}/tmp.5.txt > ${PEAKDIR}/allCelltypes.ann.rlog.sorted.txt

# normalization to total tags with reduced peak sets
declare -a CHIPS=("${CHIPDIRPD}/RS411_PU1_CNVnormRefChr" "${CHIPDIRPD}/GM12878_PU1_refChr"
"${CHIPDIRPD}/DOHH2_PU1_CNVnormRefChr" "${CHIPDIRPD}/OCILY7_PU1_CNVnormRefChr"
"${CHIPDIRPD}/H929_PU1_CNVnormRefChr" "${CHIPDIRBC}/CD19_PU1_refChr"
"${CHIPDIRPD}/HPC_PU1_refChr" "${CHIPDIRPD}/K562_PU1_CNVnormRefChr"
"${CHIPDIRBC}/BDMC_PU1_refChr" "${CHIPDIRBC}/CD15_PU1_refChr"
"${CHIPDIRCL}/other/KG1_PU1_CNVnormRefChr" "${CHIPDIRCL}/other/EM3_PU1_CNVnormRefChr"
"${CHIPDIRCL}/other/ML2_PU1_CNVnormRefChr" "${CHIPDIRCL}/other/U937_PU1_CNVnormRefChr"
"${CHIPDIRCL}/other/NB4_PU1_CNVnormRefChr" "${CHIPDIRCL}/THP1/THP1_PMAVD3_PU1_CNVnormRefChr"
"${CHIPDIRCL}/THP1/THP1_PU1_CNVnormRefChr" "${CHIPDIRMOMACDC}/MO_PU1_refChr"
"${CHIPDIRMOMACDC}/MO_4h_PU.1_Ch41_refChr" "${CHIPDIRMOMACDC}/MO_LPS_4h_PU.1_Ch43_refChr"
"${CHIPDIRMOMACDC}/MO18h_PU.1_Ch17_refChr" "${CHIPDIRMOMACDC}/MAC_PU1_refChr"
"${CHIPDIRMOMACDC}/DC_PU1_refChr")

declare -a NAMES=("RS411" "GM12878" "DOHH2" "OCILY7" "H929" "CD19" "HPC" "K562" "BDMC"
"CD15" "KG1" "EM3" "ML2" "U937" "NB4" "THP1" "THP1_PMAVD3" "MO" "MO4h" "MO4h_LPS" "MO18h"
"MAC" "DC" )

CHIPDIRS="${CHIPDIRPD}/RS411_PU1_CNVnormRefChr ${CHIPDIRPD}/GM12878_PU1_refChr
${CHIPDIRPD}/DOHH2_PU1_CNVnormRefChr ${CHIPDIRPD}/OCILY7_PU1_CNVnormRefChr
${CHIPDIRPD}/H929_PU1_CNVnormRefChr ${CHIPDIRBC}/CD19_PU1_refChr ${CHIPDIRPD}/HPC_PU1_refChr
${CHIPDIRPD}/K562_PU1_CNVnormRefChr ${CHIPDIRBC}/BDMC_PU1_refChr
${CHIPDIRBC}/CD15_PU1_refChr ${CHIPDIRCL}/other/KG1_PU1_CNVnormRefChr
${CHIPDIRCL}/other/EM3_PU1_CNVnormRefChr ${CHIPDIRCL}/other/ML2_PU1_CNVnormRefChr
${CHIPDIRCL}/other/U937_PU1_CNVnormRefChr ${CHIPDIRCL}/other/NB4_PU1_CNVnormRefChr
${CHIPDIRCL}/THP1/THP1_PMAVD3_PU1_CNVnormRefChr ${CHIPDIRCL}/THP1/THP1_PU1_CNVnormRefChr
${CHIPDIRMOMACDC}/MO_PU1_refChr ${CHIPDIRMOMACDC}/MO_4h_PU.1_Ch41_refChr
${CHIPDIRMOMACDC}/MO_LPS_4h_PU.1_Ch43_refChr ${CHIPDIRMOMACDC}/MO18h_PU.1_Ch17_refChr
${CHIPDIRMOMACDC}/MAC_PU1_refChr ${CHIPDIRMOMACDC}/DC_PU1_refChr"

annotatePeaks.pl ${PEAKDIR}/reducedCelltypes.peaks.txt hg19 -size 200 -d ${CHIPDIRS} >
${PEAKDIR}/new.reducedCelltypes.ann.noback_normtotal.txt -cpu 48

tail -n +2 ${PEAKDIR}/new.reducedCelltypes.ann.noback_normtotal.txt > ${PEAKDIR}/tmp.3.txt
sort -k7,7 ${PEAKDIR}/tmp.3.txt > ${PEAKDIR}/tmp.4.txt
cut -f 1,20-42 ${PEAKDIR}/tmp.4.txt | tail -n +2 > ${PEAKDIR}/tmp.new.reducedCelltypes.ann.txt
echo
$'ID\tRS411\tGM12878\tDOHH2\tOCILY7\tH929\tCD19\tHPC\tK562\tBDMC\tCD15\tKG1\tEM3\tML2\tU937\tN
B4\tTHP1_PMAVD3\tTHP1\tMO\tMO4h\tMO4h_LPS\tMO18h\tMAC\tDC' \
| cat - ${PEAKDIR}/tmp.new.reducedCelltypes.ann.txt >
${PEAKDIR}/reducedCelltypes.ann.noback_normtotal.txt

# generating image file with z-score normalization
cat >"${TMP}/R.image.R" <<EOF
library(RColorBrewer)
png(filename="${FIGURESDIR}/ReducedCelltypes_Peaks_Zscores.png", height=8390, width=8390)
par(fig=c(0,1,0,1),mar = rep(2, 4))
data <- read.delim("${PEAKDIR}/reducedCelltypes.ann.noback_normtotal.txt",
row.names="ID")
#d <- data.matrix(data)
scaled_data <- round(t(scale(t(data))), digits=2)
d <- as.matrix(scaled_data)
mycol <- colorRampPalette(c("black","red"))(199)
o <- d[order(nrow(d):1),]
o2 <- apply(o, 2, function(x) ifelse(x > 3, 3 ,x))
q <- apply(o2, 2, function(x) ifelse(x < 0, 0 ,x))
image(t(q),col=mycol, zlim=range(c(0,3)))
dev.off()
EOF
chmod 750 "${TMP}/R.image.R"
R < ${TMP}/R.image.R --no-save
rm ${TMP}/R.image.R

# separating clusters for motif and other downstream annotation
declare -a NAMES=(K562 HPC RS411 GM12878 DOHH2 OCILY7 H929 CD15 BDMC CD19 THP1_PMAVD3 THP1
EM3 KG1 ML2 U937 NB4 MO MO4h MO4h_LPS MO18h MAC DC)
declare -a NAMES=(RS411 GM12878 DOHH2 OCILY7 H929 CD19 HPC K562 BDMC CD15 KG1 EM3 ML2 U937
NB4 THP1 THP1_PMAVD3 MO MO4h MO4h_LPS MO18h MAC DC)
cd ${PEAKDIR}
for NAME in ${NAMES[@]}; do
grep -w "$NAME" reducedCelltypes.peaks.txt >reducedCelltypes.${NAME}.peaks.txt
done

# motif analyses for all Celltypes
declare -a NAMES=(RS411 GM12878 DOHH2 OCILY7 H929 CD19 HPC K562 BDMC CD15 KG1 EM3 ML2 U937
NB4 THP1_PMAVD3 THP1 MO MO4h MO4h_LPS MO18h MAC DC)

for NAME in ${NAMES[@]}; do
_DATE=$(date +%s)
cat >"${TMPDIR}/motifs.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
```

```
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findMotifsGenome.pl ${PEAKDIR}/reducedCelltypes.${NAME}.peaks.txt hg19r
${MOTIFDIR}/reducedCelltypes.${NAME} -size 200 -len 7,8,9,10,11,12,13,14 -p 4 -h
EOF
chmod 750 "${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
echo "finding motifs for Celltype ${NAME}"
screen -dm -S motif${NAME} bash -c "bash ${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
done

# create a file containing all known motif enrichments
_DATE=$(date +%s)
for NAME in ${NAMES[@]}; do
cat >>"${TMPDIR}/knownMotifs.reducedCelltypes.${_DATE}.txt" <<EOF
${MOTIFDIR}/reducedCelltypes.${NAME}/knownResults.txt
EOF
done

for NAME in ${NAMES[@]}; do
cat >>"${TMPDIR}/knownMotifs.names.${_DATE}.txt" <<EOF
${NAME}
EOF
done

mkdir "${MOTIFDIR}/reducedCelltypes.summary"
mySummarizeMotifResults.pl "${TMPDIR}/knownMotifs.reducedCelltypes.${_DATE}.txt" -namesFile
"${TMPDIR}/knownMotifs.names.${_DATE}.txt" \
-outDir "${MOTIFDIR}/reducedCelltypes.summary" -hc -minp 0.0000000001 -minr 2 -limit 3

# clustering not so nice, since same motif found in many different variances
# create a combined motif file containing only top motif out of all Homer motifs in all
Celltypes
COMBINEDMOTIFS="${MOTIFDIR}/combinedHomerMotifs.all.motifs"
touch ${COMBINEDMOTIFS}
for NAME in ${NAMES[@]}; do
cat ${COMBINEDMOTIFS} ${MOTIFDIR}/reducedCelltypes.${NAME}/homerMotifs.all.motifs >
"${TMPDIR}/homerMotifs.names.${_DATE}.txt"
mv "${TMPDIR}/homerMotifs.names.${_DATE}.txt" ${COMBINEDMOTIFS}
done
compareMotifs.pl ${COMBINEDMOTIFS} ${MOTIFDIR}/combinedHomerMotifs/final -reduceThresh .75
-matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 12
rm ${FILTEREDMOTIFS}
FILTEREDMOTIFS="${MOTIFDIR}/combinedFilteredHomerMotifs.all.motifs"
touch ${FILTEREDMOTIFS}
for ((i=1;i<=235;i++)); do
MOTIFCOR=$(awk -F'[()]' 'NR==1 {print $2}'
"${MOTIFDIR}/combinedHomerMotifs/final/homerResults/motif${i}.motif")
if [[ ${MOTIFCOR} > 0.85 ]]; then
cat ${FILTEREDMOTIFS} ${MOTIFDIR}/combinedHomerMotifs/final/homerResults/motif${i}.motif
> "${TMPDIR}/homerMotifs.comb.${_DATE}.txt"
mv "${TMPDIR}/homerMotifs.comb.${_DATE}.txt" ${FILTEREDMOTIFS}
fi
done

# file filtered by hand to remove {MOTIFCOR} < 0.85 -->
combinedFilteredHomerMotifs.final.all.motifs
declare -a NAMES=(RS411 GM12878 DOHH2 OCILY7 H929 CD19 HPC K562 BDMC CD15 KG1 EM3 ML2 U937
NB4 THP1 THP1_PMAVD3 MO MO4h MO4h_LPS MO18h MAC DC)

for NAME in ${NAMES[@]}; do
_DATE=$(date +%s)
cat >"${TMPDIR}/motifs.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findMotifsGenome.pl ${PEAKDIR}/reducedCelltypes.${NAME}.peaks.txt hg19r
${MOTIFDIR}/filtered.reducedCelltypes.${NAME} -size 200 -mknown
/misc/data/analysis/project_PU1/allCellTypes/motifs/combinedFilteredHomerMotifs.final.all.moti
fs -nomotif -p 2 -h
EOF
chmod 750 "${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
```

```
echo "finding motifs for Celltype ${NAME}"
screen -dm -S motif${NAME} bash -c "bash ${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
done


# create a new file containing known motif enrichments
_DATE=$(date +%s)
for NAME in ${NAMES[@]}; do
cat >>"${TMPDIR}/filtered.knownMotifs.reducedCelltypes.${_DATE}.txt" <<EOF
${MOTIFDIR}/filtered.reducedCelltypes.${NAME}/knownResults.txt
EOF
done
for NAME in ${NAMES[@]}; do
cat >>"${TMPDIR}/filtered.knownMotifs.names.${_DATE}.txt" <<EOF
${NAME}
EOF
Done

# mkdir "${MOTIFDIR}/filtered.reducedCelltypes.summary"
mySummarizeMotifResults.pl "${TMPDIR}/filtered.knownMotifs.reducedCelltypes.${_DATE}.txt"
-namesFile "${TMPDIR}/filtered.knownMotifs.names.${_DATE}.txt" \
-outDir "${MOTIFDIR}/filtered.reducedCelltypes.summary" -hc -minp 0.0000000001 -minr 2
-limit 2

# ballonplot for motif enrichment
_DATE=$(date +%s)
cat >"${TMPDIR}/R.ballon.${_DATE}.R" <<EOF
library(ggplot2)
library(reshape2)
library(ggpubr)
r <- read.table("${MOTIFDIR}/filtered.reducedCelltypes.summary/cleanedRatioTable.txt",
header=T, sep="\t")
mr <- melt(r)
q <- read.table("${MOTIFDIR}/filtered.reducedCelltypes.summary/cleanedqValueTable.txt",
header=T, sep="\t")
mq <- melt(q)
s <- read.table("${MOTIFDIR}/filtered.reducedCelltypes.summary/short.motif.names.txt",
header=T, sep="\t")
rq <- merge(mr,mq,by=c("Motif.Name","variable"))
table <- merge(s,rq,by="Motif.Name")
table\$logq <- log10(table\$value.y+0.00005)
table\$enr <- (table\$value.x+0.01)
p <- ggballoonplot(table, x = "variable", y = "Short.Name", size = "enr", color="black", fill
= "logq", size.range = c(0, 10), ggtheme = theme_bw())
p <- p + scale_fill_gradient2(low = "mediumblue", mid = "gray90",  high = "gray90", midpoint =
-1)
p <- p + scale_color_gradient2(low = "mediumblue", mid = "gray90",  high = "gray90", midpoint
= -1)
pdf(file="${FIGURESDIR}/ballonplot.motifenrichment.allCelltypes.pdf", height=5, width=5)
plot(p)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.ballon.${_DATE}.R"
R < "${TMPDIR}/R.ballon.${_DATE}.R"  --no-save
rm "${TMPDIR}/R.ballon.${_DATE}.R"

# motifscore distribution across Celltypes
declare -a NAMES=(RS411 GM12878 DOHH2 OCILY7 H929 CD19 HPC K562 BDMC CD15 KG1 EM3 ML2 U937
NB4 THP1 THP1_PMAVD3 MO MO4h MO4h_LPS MO18h MAC DC)

_DATE=$(date +%s)
for NAME in ${NAMES[@]}; do
cat >"${TMPDIR}/motifs.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
annotatePeaks.pl ${PEAKDIR}/reducedCelltypes.${NAME}.peaks.txt hg19 -size 200 -m
/misc/data/analysis/project_PU1/PU1long.motif -mscore -nogene -noann >
${TMPDIR}/tmp.14.${NAME}.txt
EOF
chmod 750 "${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
echo "annotating PU1 motif for Celltype ${NAME}"
screen -dm -S motif${NAME} bash -c "bash ${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
done
```

```
# loop to check when screen sessions are done
#--------------------------------------------
for NAME in ${NAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S motif${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample motif${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

# combined bean- and box-plot
declare -a NAMES=(RS411 GM12878 DOHH2 OCILY7 H929 CD19 HPC K562 BDMC CD15 KG1 EM3 ML2 U937
NB4 THP1 THP1_PMAVD3 MO MO4h MO4h_LPS MO18h MAC DC)

_DATE=$(date +%s)
COUNT=1
cat >"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
EOF
for NAME in ${NAMES[@]}; do
cat >>"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
data${COUNT} <- read.table("${TMPDIR}/tmp.14.${NAME}.txt", header=T, sep="\t")
EOF
COUNT=$((COUNT+=1))
done
COUNT=1
cat >>"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
# combining the data columns (bit complicated, because each column has different length)
EOF
for NAME in ${NAMES[@]}; do
cat >>"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
d${COUNT} <- data${COUNT}[-1,10]
EOF
COUNT=$((COUNT+=1))
done
cat >>"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
z <- c("d1", "d2", "d3", "d4", "d5", "d6", "d7", "d8", "d9", "d10", "d11", "d12", "d13",
"d14", "d15", "d16", "d17", "d18", "d19", "d20", "d21", "d22", "d23")
x <- lapply(z, get, envir=environment())
names(x) <- z
#determining the length of each column and defining labels
EOF
COUNT=1
for NAME in ${NAMES[@]}; do
cat >>"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
d${COUNT}num <- nrow(as.matrix(d${COUNT}))
lab${COUNT} <- paste("${NAME} (",d${COUNT}num,")",sep="")
EOF
COUNT=$((COUNT+=1))
done
cat >>"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
#defining colors
#beancol <-
list("gold","goldenrod2","darkorange1","darkorange3","firebrick1","firebrick3","firebrick","de
eppink","darkmagenta","deepskyblue2","dodgerblue1","dodgerblue3","blue2","blue4",
"aquamarine", "aquamarine3", "aquamarine4", "darkseagreen1", "darkseagreen3",
"darkseagreen4", "darkslategray1", "darkslategray3", "darkslategray4")
#boxcol <-
c("gray25","gray30","gray35","gray40","gray45","gray50","gray55","gray60","gray65","gray70","g
ray75","gray80","gray85","gray90","gray92","gray93","gray94","gray95","gray96","gray97","gray9
8","gray99","gray100")
beancol <-
list("firebrick","firebrick","firebrick","firebrick","firebrick","orange2","hotpink","darkturq
uoise","darkslategray4","darkcyan","blue","blue","blue","blue","blue","blue","blue","blueviole
t","slateblue3","slateblue3","slateblue3","dodgerblue1","forestgreen")
boxcol <-
c("gray73","gray74","gray75","gray76","gray77","gray78","gray79","gray80","gray81","gray82","g
ray83","gray84","gray85","gray86","gray87","gray88","gray89","gray90","gray91","gray92","gray9
3","gray94","gray95")
pdf(file="${FIGURESDIR}/motifScores.allCelltypes.pdf", height=4, width=5)
par(mar=c(8.5,5,1,1))
#plotting the beans
```

```
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1.5, beanlinewd = .5, maxstripline =
0.8,ylim=c(0,12),lwd=0.5)
#adding box plot on top
par(mar=c(8.5,5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(0,12),lwd=0.6)
#axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:23,
labels=c(lab1,lab2,lab3,lab4,lab5,lab6,lab7,lab8,lab9,lab10,lab11,lab12,lab13,lab14,lab15,lab1
6,lab17,lab18,lab19,lab20,lab21,lab22,lab23),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext(" Motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
#mtext("Celltype",family="Helvetica",ps=12,side=1,line=7,cex=1.2,padj=2.0)
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.motifbean.${_DATE}.R"
R < ${TMPDIR}/R.motifbean.${_DATE}.R --no-save
rm ${TMPDIR}/R.motifbean.${_DATE}.R


# analysis of motif co-occurence
# redo known motif analysis with PU.1 motif masked
declare -a NAMES=(RS411 GM12878 DOHH2 OCILY7 H929 CD19 HPC K562 BDMC CD15 KG1 EM3 ML2 U937
NB4 THP1 THP1_PMAVD3 MO MO4h MO4h_LPS MO18h MAC DC)

for NAME in ${NAMES[@]}; do
_DATE=$(date +%s)
cat >"${TMPDIR}/motifs.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findMotifsGenome.pl ${PEAKDIR}/reducedCelltypes.${NAME}.peaks.txt hg19r
${MOTIFDIR}/filtered.PU1masked.${NAME} -size 200 -mknown
/misc/software/ngs/homer/v4.9/data/knownTFs/vertebrates/known.motifs -maskMotif ${PU1MOTIF}
-nomotif -p 3 -h
EOF
chmod 750 "${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
echo "finding motifs for Celltype ${NAME}"
screen -dm -S motif${NAME} bash -c "bash ${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#---------------------------------------------
for NAME in ${NAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S motif${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample motif${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------

# network of motif co-association for selected samples
declare -a NAMES=(CD15 CD19 BDMC THP1_PMAVD3)
_DATE=$(date +%s)
for NAME in ${NAMES[@]}; do
      cat >"${TMPDIR}/motifs.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-
5.26.1/bin:/misc/software/ngs/samtools/samtools-
1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
annotatePeaks.pl ${PEAKDIR}/reducedCelltypes.${NAME}.peaks.txt hg19 -size 200 -m
/misc/software/ngs/homer/v4.9/data/knownTFs/vertebrates/known.motifs -fm ${PU1MOTIF} -
matrixMinDist 4 -nogene -noann -nmotifs -matrix ${TMPDIR}/.${NAME}rm >
${TMPDIR}/tmp.m.14.${NAME}.rm.txt
```

```
annotatePeaks.pl ${PEAKDIR}/reducedCelltypes.${NAME}.peaks.txt hg19 -size 200 -m ${PU1MOTIF} -
matrixMinDist 6 -nogene -noann -nmotifs -matrix ${TMPDIR}/.${NAME}pu >
${TMPDIR}/tmp.m.14.${NAME}.pu.txt
(head -n 1 ${TMPDIR}/tmp.m.14.${NAME}.rm.txt && tail -n +2 ${TMPDIR}/tmp.m.14.${NAME}.rm.txt |
sort -k 1,1 )> ${TMPDIR}/tmp.m.14.${NAME}.rm.sorted.txt
cut -f11- ${TMPDIR}/tmp.m.14.${NAME}.rm.sorted.txt | sed -e s/Distance\ From\
Peak\(sequence,strand,conservation\)//g > ${TMPDIR}/tmp.m3.14.${NAME}.txt
myTP.pl ${TMPDIR}/tmp.m3.14.${NAME}.txt > ${TMPDIR}/tmp.m3.14.${NAME}.tp.txt
join -1 1 -2 1 -t $'\t' <(sort -k1,1 ${CUSTOMMOTIFS}/knownMotifs${NAME}.classes.txt ) <(sort -
k1,1 ${TMPDIR}/tmp.m3.14.${NAME}.tp.txt) > ${TMPDIR}/tmp.m4.14.${NAME}.tp.txt
sort -k2,2 ${TMPDIR}/tmp.m4.14.${NAME}.tp.txt > ${TMPDIR}/tmp.m5.14.${NAME}.tp.txt
redMotifCounts.pl ${TMPDIR}/tmp.m5.14.${NAME}.tp.txt > ${TMPDIR}/tmp.m6.14.${NAME}.tp.txt
myTP.pl ${TMPDIR}/tmp.m6.14.${NAME}.tp.txt >${TMPDIR}/tmp.m6.14.${NAME}.txt
(head -n 1 ${TMPDIR}/tmp.m.14.${NAME}.pu.txt && tail -n +2 ${TMPDIR}/tmp.m.14.${NAME}.pu.txt |
sort -k 1,1 )> ${TMPDIR}/tmp.m.14.${NAME}.pu.sorted.txt
cut -f10- ${TMPDIR}/tmp.m.14.${NAME}.pu.sorted.txt | sed -e s/Distance\ From\
Peak\(sequence,strand,conservation\)//g > ${TMPDIR}/tmp.m3.14.${NAME}.pu.txt
redAutoMotifCounts.pl ${TMPDIR}/tmp.m3.14.${NAME}.pu.txt PU1 >
${TMPDIR}/tmp.m6.14.${NAME}.pu.txt
paste ${TMPDIR}/tmp.m6.14.${NAME}.pu.txt ${TMPDIR}/tmp.m6.14.${NAME}.txt >
${TMPDIR}/tmp.m7.14.${NAME}.txt
getNetworkMotifFiles.pl ${TMPDIR}/tmp.m7.14.${NAME}.txt ${WORKDIR}/${NAME}
EOF
       chmod 750 "${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
       echo "annotating PU1 motif for Celltype ${NAME}"
       screen -dm -S motif${NAME} bash -c "bash ${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#--------------------------------------------
for NAME in ${NAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S motif${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample motif${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

#plotting the networks
declare -a NAMES=(CD15 CD19 BDMC THP1_PMAVD3)

_DATE=$(date +%s)
for NAME in ${NAMES[@]}; do
       cat >"${TMPDIR}/network.${NAME}.${_DATE}.R" <<EOF
library("igraph")
nodes <- read.delim("${WORKDIR}/${NAME}.nodes.txt", header=T, as.is=T)
links <- read.delim("${WORKDIR}/${NAME}.edges.txt", header=T, as.is=T)
net <- graph.data.frame(links, nodes, directed=F)
colrs <- c("blue","firebrick1")
V(net)\$color <- colrs[factor(V(net)\$type.label)]
V(net)\$size <- sqrt(V(net)\$fraction)*12
V(net)\$label <- V(net)\$tf.name
E(net)\$width <- E(net)\$weight/2
E(net)\$edge.color <- "black"
l <- layout.star(net)
pdf(file="${FIGURESDIR}/motif-coenrich.network.${NAME}.pdf", height=6, width=6)
plot(net, layout=l, vertex.label.family="Helvetica", vertex.frame.color="black")
blob.size <- c(80,40,20,5)
legend(x=1.6,y=1, blob.size, pch=21, col="black", pt.bg="white", pt.cex=sqrt(blob.size)/0.42,
cex=.8, bty="n")
dev.off()
EOF
       chmod 750 "${TMPDIR}/network.${NAME}.${_DATE}.R"
       R < "${TMPDIR}/network.${NAME}.${_DATE}.R"  --no-save
       rm "${TMPDIR}/network.${NAME}.${_DATE}.R"
done
```

### 10.1.3 Analysis of TBSAseq Data of CTV-1 Cells

Paired end 300 bp sequencing of indexed TBSAseq samples was carried out at the Faculty of Biochemistry of the University of Regensburg (Department of Prof. Dr. Meister) using an Illumina MiSeq sequencer (four-channel sequencing chemistry, SBSv3). The general TBSAseq workflow is explained in section 4.2.10. Obtained raw read fastq files were processed utilizing the CRISPResso software (Pinello et al. 2016). The analyzed amplicon is listed in section 3.7.1. In a first step, a list of all raw read files was generated. In addition, the amplicon sequence and a fake guide sequence needed to run the software were given.

```
#!/bin/bash

#setting homer environment
DIR_PKG="/misc/software/ngs"
PATH_PERL=/misc/software/package/perl/perl-5.26.1/bin
PATH_SAMTOOLS=${DIR_PKG}/samtools/samtools-1.6/bin
PATH_HOMER=${DIR_PKG}/homer/v4.9/bin
PATH_R=/misc/software/package/RBioC/3.4.3/bin
export PATH=${PATH_R}:${PATH_PERL}:${PATH_SAMTOOLS}:${PATH_HOMER}:${PATH}
export PATH

#setting paths
CRISPRESSO="/misc/software/ngs/crispresso/build/v1.0.10/CRISPResso-master/crispresso_pythonEnv
/bin/CRISPResso"
OUTPUTDIR="/misc/data/analysis/project_PU1/BSamplicons"
FASTQDIR="/misc/data/rawData/DNA/BSamplicons"
#move amplicon data to the right folder
mv /misc/data/rawData/RNA/RepSeq/MiSeq_042018/Julia*.fastq.gz ${FASTQDIR}
mkdir -p ${OUTPUTDIR}

# run CRISPResso on bisulfite amplicons
declare -a R1=("${FASTQDIR}/Julia1_S55_L001_R1_001.fastq.gz"
"${FASTQDIR}/Julia2_S56_L001_R1_001.fastq.gz" "${FASTQDIR}/Julia3_S57_L001_R1_001.fastq.gz"
"${FASTQDIR}/Julia4_S58_L001_R1_001.fastq.gz" "${FASTQDIR}/Julia5_S59_L001_R1_001.fastq.gz"
"${FASTQDIR}/Julia6_S60_L001_R1_001.fastq.gz" "${FASTQDIR}/Julia1_S95_L001_R1_001.fastq.gz"
"${FASTQDIR}/Julia2_S96_L001_R1_001.fastq.gz" "${FASTQDIR}/Julia3_S97_L001_R1_001.fastq.gz"
"${FASTQDIR}/Julia4_S98_L001_R1_001.fastq.gz" "${FASTQDIR}/Julia5_S99_L001_R1_001.fastq.gz"
"${FASTQDIR}/Julia6_S100_L001_R1_001.fastq.gz")

declare -a R2=("${FASTQDIR}/Julia1_S55_L001_R2_001.fastq.gz"
"${FASTQDIR}/Julia2_S56_L001_R2_001.fastq.gz" "${FASTQDIR}/Julia3_S57_L001_R2_001.fastq.gz"
"${FASTQDIR}/Julia4_S58_L001_R2_001.fastq.gz" "${FASTQDIR}/Julia5_S59_L001_R2_001.fastq.gz"
"${FASTQDIR}/Julia6_S60_L001_R2_001.fastq.gz" "${FASTQDIR}/Julia1_S95_L001_R2_001.fastq.gz"
"${FASTQDIR}/Julia2_S96_L001_R2_001.fastq.gz" "${FASTQDIR}/Julia3_S97_L001_R2_001.fastq.gz"
"${FASTQDIR}/Julia4_S98_L001_R2_001.fastq.gz" "${FASTQDIR}/Julia5_S99_L001_R2_001.fastq.gz"
"${FASTQDIR}/Julia6_S100_L001_R2_001.fastq.gz")

declare -a NAMES=("CTV1_PU1_100nM_R1" "CTV1_PU1mut_100nM_R1" "CTV1_PU1_100nM_R2"
"CTV1_PU1mut_100nM_R2" "CTV1_PU1_100nM_R3" "CTV1_PU1mut_100nM_R3" "CTV1_gDNA_Ctrl"
"CTV1_DMSO_Ctrl" "CTV1_10nM" "CTV1_100nM" "CTV1_300nM" "CTV1_1000nM")
SEQ="GTTTTAGGGATTAGGAAGGGATTTTCGTTTAGTGTTAGGTTATTTATTTCGGGTAGTTCGTTAGGTTGCGGCGTTTTGTTTAGTTTGTT
GTAGTTGAGATCGAGTATTTTGAGTTTGGTTGGTAGTTTTTTAGGTATTTGTTTTAGTTTAGCGAACGATAGATTGAGGGAGTTTAGGGCGTTG
GATTATATGTATTTCGGAGCGTTAGGGTTTACGGTGGCGCGTAGCGAGTTGTGGTTGAGGTTTAGGTTGTGGGGTTGTAT"
GUIDE="CGGAGCGTTAGGGTTTACGGTGG"

COUNT=0
for READ1 in ${R1[@]}; do
NAME="${NAMES[${COUNT}]}"
READ2="${R2[${COUNT}]}"
$CRISPRESSO -r1 ${READ1} -r2 ${READ2} -a ${SEQ} -g ${GUIDE} --trim_sequences
--keep_intermediate -o ${OUTPUTDIR}/${NAME}
COUNT=$((COUNT+=1))
Done
```

After running CRISPResso the output was further processed using the CRISPRessoBS.pl perl script, to adapt the software for bisulfite amplicon sequencing results. The script separates alleles carrying deletions, determines the positions of individual C's in the reference allele and the counts C or T respectively at these positions.

```
# loop to run CRISPRessoBS.pl over all samples:
COUNT=0
for NAME in ${NAMES[@]}; do
READ1="${R1[${COUNT}]}"
READ1NAMEBASE=${READ1##*/}
READ1NAME=${READ1NAMEBASE%%.*}
READ2="${R2[${COUNT}]}"
READ2NAMEBASE=${READ2##*/}
READ2NAME=${READ2NAMEBASE%%.*}
CrispressoBS.pl
${OUTPUTDIR}/${NAME}/CRISPResso_on_${READ1NAME}_${READ2NAME}/Alleles_frequency_table.txt
-outDir ${OUTPUTDIR}/${NAME} -minDel ---
COUNT=$((COUNT+=1))
done

# CRISPRessoBS.pl script:
my $directory = "/loctmp";
use List::Util qw(sum);

if (@ARGV < 2) {
    print STDERR "\nCrispressoBS.pl <Alleles_frequency_table> -outDir
    <outputDir>\n";
    print STDERR "required options (either one or the other):\n";
    print STDERR "other options :\n";
    print STDERR "\t-minDel <minimum number of deleted nucleotides given as \"---\"
    (default ---) >\n\n";
    exit;
}

# basic parameters analyzed within the script:
my $rand = rand();
my $freqTable = $ARGV[0];
my $CpGpos = '';
my @CpGpos = ();
my $minDel = "---";
my $del = "-";
my @sumsNo = ();
my @sumsYes = ();
my @CpGsNo = ();
my @TpGsNo = ();
my @CpGsYes = ();
my @TpGsYes = ();
my $factor = 1;
my $outDir = '';
my $counterYes = 0;
my $counterNo = 0;
my $cutOff = 3;
my $CrisprEdit = 0;
my $CrisprAllele = 0;
my $CrisprMet = 0;
my $noCrisprMet = 0;

for (my $i=0;$i<@ARGV;$i++) {
    if ($ARGV[$i] eq '-outDir') {
        $outDir = $ARGV[++$i] ;
        $outDir .= '/' unless $outDir =~ m(/$);
    } elsif ($ARGV[$i] eq '-minDel') {
        $minDel = $ARGV[++$i] ;
        $cutOff = length($minDel) ;
    }
}

# generate text file with analysis summary for each sample:
my $out = $outDir . "summary.txt";
open IN, $freqTable ;
open OUT, ">$out";
while (<IN>) {
    chomp;
    s/\r//g;
    my @line= split /\t/;

    if ($. == 1) {
    } else {

        $CpGpos=`echo $line[1] | grep -bio C | grep -oE '[0-9]+' | awk '{print
        \$1+1}' | tr -s \"\n\" \" \"`;
        @CpGpos= split / /, $CpGpos;
        if ($. == 2) {
            print STDERR "\nposition of CpGs in Reference\n";
            for (my $i=0;$i<@CpGpos;$i++) {
            print STDERR "$CpGpos[$i]\t";
            }
            print STDERR "\n\n";

            for (my $i=0;$i<@CpGpos;$i++) {
                $sumsYes[$i] = 0;
            }
            for (my $i=0;$i<@CpGpos;$i++) {
                $sumsNo[$i] = 0;
            }
            for (my $i=0;$i<@CpGpos;$i++) {
                $CpGsYes[$i] = 0;
            }
            for (my $i=0;$i<@CpGpos;$i++) {
                $TpGsYes[$i] = 0;
            }
            for (my $i=0;$i<@CpGpos;$i++) {
                $CpGsNo[$i] = 0;
            }
            for (my $i=0;$i<@CpGpos;$i++) {
                $TpGsNo[$i] = 0;
            }
        }
        if  (index($line[0], $minDel) != -1 && $line[5] >= $cutOff ) {
            $factor = $line[8];
            for (my $i=0;$i<@CpGpos;$i++) {
                $sumsYes[$i] = $sumsYes[$i] + $factor;
                if (substr($line[0],$CpGpos[$i]-1,1) eq "C") {
                    $CpGsYes[$i] = $CpGsYes[$i] + $factor;
                }   elsif (substr($line[0],$CpGpos[$i]-1,1) eq "T") {
                    $TpGsYes[$i] = $TpGsYes[$i] + $factor;
                }

            }
            $counterYes = $counterYes + $factor;

        } elsif (index($line[0], $del) != -1) {
        } else {
            $factor = $line[8];
            for (my $i=0;$i<@CpGpos;$i++) {
                $sumsNo[$i] = $sumsNo[$i] + $factor;
                if (substr($line[0],$CpGpos[$i]-1,1) eq "C") {
                    $CpGsNo[$i] = $CpGsNo[$i] + $factor;
                }   elsif (substr($line[0],$CpGpos[$i]-1,1) eq "T") {
                    $TpGsNo[$i] = $TpGsNo[$i] + $factor;
                }
            }
            $counterNo = $counterNo + $factor;
        }
    }
}

# generating a summary output:
print OUT "summary";
for (my $i=0;$i<@CpGpos;$i++) {
    print OUT "\tpos.$CpGpos[$i]";
}
print OUT "\ntotal C in Crispr ($counterYes)";
for (my $i=0;$i<@CpGpos;$i++) {
    print OUT "\t$CpGsYes[$i]";
}
print OUT "\ntotal T in Crispr ($counterYes)";
for (my $i=0;$i<@CpGpos;$i++) {
    print OUT "\t$TpGsYes[$i]";
}
```

```
print OUT "\ntotal Crispr ";
for (my $i=0;$i<@CpGpos;$i++) {
    print OUT "\t$sumsYes[$i]";
}
print OUT "\n% deleted C";
for (my $i=0;$i<@CpGpos;$i++) {
    if ($sumsYes[$i] == 0) {
    $CrisprEdit= 0;
    } else {
    $CrisprEdit= 100 - int(($CpGsYes[$i]+$TpGsYes[$i])/$sumsYes[$i]*1000)/10;
    }
    print OUT "\t$CrisprEdit";
}
print OUT "\n% Crispr-edited alleles";
for (my $i=0;$i<@CpGpos;$i++) {
    $CrisprAllele= int($sumsYes[$i]/$sumsNo[$i]*1000)/10;
    print OUT "\t$CrisprAllele";
}
print OUT "\ntotal C in notCrispr ($counterNo)";
for (my $i=0;$i<@CpGpos;$i++) {
    print OUT "\t$CpGsNo[$i]";
}
print OUT "\ntotal T in notCrispr ($counterNo)";
for (my $i=0;$i<@CpGpos;$i++) {
    print OUT "\t$TpGsNo[$i]";
}
print OUT "\ntotal notCrispr ";
for (my $i=0;$i<@CpGpos;$i++) {
    print OUT "\t$sumsNo[$i]";
}
print OUT "\n% methylated in non-edited allels";
for (my $i=0;$i<@CpGpos;$i++) {
    if (($CpGsNo[$i]+$TpGsNo[$i]) == 0) {
    print OUT "\tNA";
    } else {
    $CrisprMet= int($CpGsNo[$i]/($CpGsNo[$i]+$TpGsNo[$i])*1000)/10;
    print OUT "\t$CrisprMet";
    }
}
print OUT "\n% methylated in Crispr";
for (my $i=0;$i<@CpGpos;$i++) {
    if (($CpGsYes[$i]+$TpGsYes[$i]) == 0) {
    print OUT "\tNA";
    } else {
    $CrisprMet= int($CpGsYes[$i]/($CpGsYes[$i]+$TpGsYes[$i])*1000)/10;
    print OUT "\t$CrisprMet";
    }
}
close IN;
close OUT;
exit;
```

```
# BS summary.txt was made up of each individual summary.txt in excel
# heat map was generated with R
x <- read.delim("/Users/michaelrehli/Dropbox/For JuliaM/phD thesis stuff/bisulfite
heatmap/BS summary.txt", header=TRUE, row.names="Sample", sep="\t")
data <- as.matrix(x)
mycol <- colorRampPalette(c("yellow","darkblue"))(199)
pdf(file="/Users/michaelrehli/Dropbox/For JuliaM/phD thesis stuff/bisulfite
heatmap/BS_heatmap.pdf")
heatmap <- heatmap.2(data, Rowv=NA, Colv=NA, col = mycol, scale="column", margins=c(5,10),
key=TRUE, density.info="none", trace="none", dendrogram="none")
dev.off()
```

### 10.1.4 Analysis of NGS Data of DAC-treated CTV-1 Cells

The following script was used to analyze the binding properties of PU.1 in CTV-1 cells treated with DAC. The script includes parts of edgeR (Robinson et al. 2010), the HOMER suite (Heinz et al. 2010) as well as parts of the R software (R Development Core Team 2008).

```
#!/bin/bash
#setting homer environment
DIR_PKG="/misc/software/ngs"
PATH_PERL=/misc/software/package/perl/perl-5.26.1/bin
PATH_SAMTOOLS=${DIR_PKG}/samtools/samtools-1.6/bin
PATH_HOMER=${DIR_PKG}/homer/v4.9/bin
PATH_R=/misc/software/package/RBioC/3.4.3/bin
export PATH=${PATH_R}:${PATH_PERL}:${PATH_SAMTOOLS}:${PATH_HOMER}:${PATH}
export PATH

BEDTOOLS="/misc/software/ngs/bedtools/bedtools2-2.27.1/bin/bedtools"
WORKDIR="/misc/data/analysis/project_PU1/CTV1"
TAGDIR="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/RNAtransfection/PU1"
CHIPDIR="${WORKDIR}/ChIP"
ATACDIR="${WORKDIR}/ATAC"
RNADIR="${WORKDIR}/RNAseq"
EXPRESSION="/misc/data/analysis/project_PU1/CTV1/RNAseq/basic"
ANALYSISDIR="${WORKDIR}/analysis/basic"
CHROMSIZES="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/hg19.chrom.sizes"
TMPDIR="/loctmp"
TMP="/loctmp/${RANDOM}.tmp."
FIGURESDIR="${WORKDIR}/figures/basic"
ALLMCIP="/misc/data/analysis/generalStuff/MCIpDetectable/All_MCIp_detected.bed"
CTV1MCIPDIR="/misc/data/processedData/tagDir/DNA/hg19/MCIp/CellLines/MCIp_CTV1_603"
CTV1PU1DIR="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/RNAtransfection/PU1/CTV1_Flag_
CNVnormRefChr_merged2"
CTV1DACPU1DIR="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/RNAtransfection/PU1/CTV1_DA
C_PU1_Flag_CNVnormRefChr_merged"
CNVFILECTV1="/misc/data/processedData/mapping/DNA/hg19/Input/CellLines/CNVdata/CTV1_genInput_6
20.sam_CNVs"
PU1MOTIF="/misc/data/analysis/project_PU1/PU1long.motif"
PU1MOTIFR="/misc/data/analysis/project_PU1/PU1long.rev.motif"
PU1MOTIFBED="/misc/data/analysis/project_PU1/homotypicClusters/PU.1_long_hg19_all.bed"
CGMOTIFBED="/misc/data/analysis/generalStuff/CpG/hg19/CGmotif.bed"
BLACKLIST_HG19="/misc/data/analysis/generalStuff/annotation/hg19/hg19.blacklist.bed"
HG19TRANS="/misc/data/analysis/generalStuff/annotation/hg19/gencode.v19.transcripts.txt"
HG19GTEX="/misc/data/analysis/generalStuff/annotation/hg19/GTEx_v7_eQTL_WBA.snpgenes.1000.bed"
PEAKDIR="${CHIPDIR}/peaksCNVcorr"
DIFFDIR="${CHIPDIR}/diffPeaksCNVcorr"
ATACPEAKDIR="${ATACDIR}/peaksCNVcorr"
ATACDIFFPEAKDIR="${ATACDIR}/diffPeaksCNVcorr"
MOTIFDIR="${WORKDIR}/motifs/basic"
KNOWNTFDIR="/misc/software/ngs/homer/v4.9/data/knownTFs/motifs"
ATACTAGDIR="/misc/data/processedData/tagDir/chromatin/hg19/ATAC/RNAtransfection/PU1/"
CHIPTAGDIR="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/RNAtransfection/PU1/"
BIGWIGDIR="/misc/data/processedData/bigWig/chromatin/hg19/ChIP/RNAtransfection/PU1/"
BIGWIGATAC="/misc/data/processedData/bigWig/chromatin/hg19/ATAC/RNAtransfection/PU1/"
PU1CTV1PEAKS="${PEAKDIR}/PU1_Flag_merged.factor.fdr05.ntag15.filtered.bed"
TAGDIRSETS="${CHIPTAGDIR}/CTV1_PU1mut_Flag_merged ${CHIPTAGDIR}/CTV1_PU1_Flag_merged
${ATACTAGDIR}/CTV1_PU1mut_merged ${ATACTAGDIR}/CTV1_PU1_merged
${CHIPTAGDIR}/CTV1_PU1mut_H3K27ac_merged ${CHIPTAGDIR}/CTV1_PU1_H3K27ac_merged"
HOMOCLUSTERDIR="/misc/data/analysis/project_PU1/homotypicClusters/"
GHISTDIR="${WORKDIR}/analysis/basic/ghist"
HISTDIR="${WORKDIR}/analysis/basic/hist"
HG19GTF="/misc/software/ngs/genome/annotation/hg19/gencode.v19.annotation.gtf"
CGIHG19="/misc/data/analysis/generalStuff/annotation/hg19/CGI_hg19_UCSC.bed"
GENOME_HG19="/misc/software/ngs/genome/sequence/hg19/hg19.fa"
PHYLOP="/misc/data/analysis/generalStuff/conservation/hg19/hg19.100way.phyloP100way.bedGraph"
PHASTCONS="/misc/data/analysis/generalStuff/conservation/hg19/hg19.100way.phastCons.bedGraph"
METASCAPE="/misc/data/analysis/project_PU1/CTV1/analysis/basic/Metascape"
ALLMCIP="/misc/data/analysis/generalStuff/MCIpDetectable/All_MCIp_detected.bed"
CTV1MCIPDIR="/misc/data/processedData/tagDir/DNA/hg19/MCIp/CellLines/MCIp_CTV1_603"
DACPU1CTV1PEAKS="${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.ntag15.filtered.bed"
noDACPU1CTV1PEAKS="${PEAKDIR}/PU1_Flag_merged_2sample.factor.fdr05.ntag15.filtered.bed"
```

```
mkdir -p ${RNADIR}/basic/DAC/
mkdir -p ${ATACDIR}/peaksCNVcorr/
mkdir -p ${ATACDIR}/diffPeaksCNVcorr/


# basic analysis of RNAseq data

cd "/misc/data/processedData/mapping/RNA/hg19/RNAseq/RNAtransfection/PU1"
paste /misc/software/ngs/genome/sequence/hg19/STAR_transcriptIDshort.txt \
CG12_CTV1_mock.trimmed.ReadsPerGene.txt \
UA122_CTV1_mock.ReadsPerGene.txt \
CG10_CTV1_PU1mRNA.trimmed.ReadsPerGene.txt \
UA120_CTV1_PU1mRNA.ReadsPerGene.txt \
CG11_CTV1_mutPU1mRNA.trimmed.ReadsPerGene.txt \
UA121_CTV1_mutPU1mRNA.ReadsPerGene.txt \
JM01_CTV1_PU1mRNA_DAC.ReadsPerGene.txt \
JM03_CTV1_PU1mRNA_DAC.ReadsPerGene.txt \
JM05_CTV1_PU1mRNA_DAC.ReadsPerGene.txt \
JM02_CTV1_mutPU1mRNA_DAC.ReadsPerGene.txt \
JM04_CTV1_mutPU1mRNA_DAC.ReadsPerGene.txt \
JM06_CTV1_mutPU1mRNA_DAC.ReadsPerGene.txt \
> /loctmp/tmpReadCountTable_RawDAC.txt

cut -f 1,4,8,12,16,20,24,28,32,36,40,44,48 /loctmp/tmpReadCountTable_RawDAC.txt >
/loctmp/tmpReadCountTable_RawRedDAC.txt

tail -n +5 /loctmp/tmpReadCountTable_RawRedDAC.txt >
/loctmp/tmpReadCountTable_RawRedDelDAC.txt

echo
$'Gene\tCTV1_noDAC_woPU1_rep1\tCTV1_noDAC_woPU1_rep2\tCTV1_noDAC_PU1_rep1\tCTV1_noDAC_PU1_rep2
\tCTV1_noDAC_PU1mut_rep1\tCTV1_noDAC_PU1mut_rep2\tCTV1_DAC_PU1_rep1\tCTV1_DAC_PU1_rep2\tCTV1_D
AC_PU1_rep3\tCTV1_DAC_PU1mut_rep1\tCTV1_DAC_PU1mut_rep2\tCTV1_DAC_PU1mut_rep3' | cat -
/loctmp/tmpReadCountTable_RawRedDelDAC.txt > ${RNADIR}/basic/DAC/ReadCountTable.txt

# basic analysis edgeR
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.edgeR.P.${_DATE}.R" <<EOF
library(edgeR)
library(ggplot2)
library(ggrepel)
data <- read.delim("${RNADIR}/basic/DAC/ReadCountTable.txt", row.names="Gene")
group <- factor(c(rep("CTV1_noDAC_woPU1", 2), rep("CTV1_noDAC_PU1", 2),
rep("CTV1_noDAC_PU1mut", 2), rep("CTV1_DAC_PU1", 3), rep("CTV1_DAC_PU1mut", 3)))
batch <- factor(c(1,2,1,2,1,2,2,2,2,2,2,2))
treatment <- factor(c(rep("noDAC_woPU1", 2), rep("noDAC_PU1", 2), rep("noDAC_PU1mut", 2),
rep("DAC_PU1", 3), rep("DAC_PU1mut", 3)))
d <- DGEList(counts=data,group=group)
keep <- rowSums(cpm(d) > 1) >= 2
summary(keep)
d <- d[keep, , keep.lib.sizes=FALSE]
d <- calcNormFactors(d)
d$samples
counts.per.m <- cpm(d, normalized.lib.sizes=TRUE)
logcpm <- cpm(d, prior.count=2, log=TRUE)

# removing batch effect for heatmaps etc.
design=model.matrix(~treatment)
corr.cpm <- removeBatchEffect(counts.per.m, batch=batch, design=design)
corr.logcpm <- removeBatchEffect(logcpm, batch=batch, design=design)
write.table (corr.cpm, file = "${RNADIR}/basic/DAC/ReadCountTable.norm.corr.cpm.txt", sep =
"\t", col.names=NA, quote=FALSE)
write.table (corr.logcpm, file = "${RNADIR}/basic/DAC/ReadCountTable.norm.corr.logcpm.txt",
sep = "\t", col.names=NA, quote=FALSE)
write.table (corr.logcpm, file =
"${RNADIR}/basic/DAC/ReadCountTable.norm.corr.logcpm.expression", sep = "\t", col.names=NA,
quote=FALSE)
corr.logcpm.scaled <- (scale(t(corr.logcpm)))
corr.logdata.scaled <- data.matrix(t(corr.logcpm.scaled))
write.table (corr.logdata.scaled, file =
"${RNADIR}/basic/DAC/ReadCountTable.norm.corr.zscore.txt", sep = "\t", col.names=NA,
quote=FALSE)

# MDS plot (batch corrected)
points <- c(15,16,17,1,2)
colors <- rep(c("goldenrod", "chocolate", "forestgreen","blue", "midnightblue"))
```

```
pdf(file="${FIGURESDIR}/edgeR/edgeR_DAC_RNAseq_MDSplot_batchCorr.pdf", height=5, width=5)
plotMDS(corr.logcpm, col=colors[group], pch=points[group])
legend("right", legend=levels(group), pch=points, col=colors, ncol=1)
dev.off()

# DGE analysis
design <- model.matrix(~0 + treatment + batch)
rownames(design) <- colnames(d)
d <- estimateDisp(d, design, robust=TRUE)
d\$common.dispersion
fit <- glmQLFit(d, design)
con <- makeContrasts((treatmentDAC_PU1 - treatmentDAC_PU1mut) - (treatmentnoDAC_PU1 -
treatmentnoDAC_PU1mut), levels=design)
qlf.2vs3 <- glmQLFTest(fit, contrast=con)
qstat.2vs3 <- topTags(qlf.2vs3, n=Inf)
write.table (qstat.2vs3, file = "${RNADIR}/basic/DAC/qstat_PU1_DACvsPU1_MOCK.glm.txt", sep =
"\t", col.names=NA, quote=FALSE)
summary(qdt.2vs3 <- decideTestsDGE(qlf.2vs3))
qisDE.2vs3 <- as.logical(qdt.2vs3)
qDEnames.2vs3 <- rownames(d)[qisDE.2vs3]
pdf(file="${FIGURESDIR}/edgeR/edgeR_DACvsMOCK_RNAseq_MvA_withBatchCorr.pdf", height=5,
width=5)
plotSmear(qlf.2vs3, de.tags=qDEnames.2vs3)
abline(h=c(-1,1), col="blue")
dev.off()
EOF
chmod 750 "/loctmp/R.edgeR.P.${_DATE}.R"
R < /loctmp/R.edgeR.P.${_DATE}.R  --no-save
rm /loctmp/R.edgeR.P.${_DATE}.R

# Down    47
# NotSig  14636
# Up      234

# extract expressed genes from table
awk -F '[\$]' 'BEGIN{OFS="\t"}{print $2}'
${RNADIR}/basic/DAC/ReadCountTable.norm.corr.logcpm.txt >
${RNADIR}/basic/DAC/expressedGenesList.txt

# assemble gene count table including statistics
cd ${RNADIR}/basic/DAC/
(head -n 1 qstat_PU1_DACvsPU1_MOCK.glm.txt && tail -n +2 qstat_PU1_DACvsPU1_MOCK.glm.txt |
sort -k 1,1 ) > /loctmp/tmpqstat_PU1_DACvsPU1_MOCK.glm.txt;
(head -n 1 ReadCountTable.norm.corr.cpm.txt && tail -n +2 ReadCountTable.norm.corr.cpm.txt |
sort -k 1,1 ) > /loctmp/tmpsortedReadCountTable.norm.cpm.txt;
(head -n 1 ReadCountTable.norm.corr.logcpm.txt && tail -n +2
ReadCountTable.norm.corr.logcpm.txt | sort -k 1,1 ) >
/loctmp/tmpsortedReadCountTable.norm.logcpm.txt;
paste /loctmp/tmpsortedReadCountTable.norm.logcpm.txt
/loctmp/tmpsortedReadCountTable.norm.cpm.txt /loctmp/tmpqstat_PU1_DACvsPU1_MOCK.glm.txt >
/loctmp/tmpsortedReadCountTable.norm.analysed.txt
sed -e '1s/^/Gene/g' /loctmp/tmpsortedReadCountTable.norm.analysed.txt >
${RNADIR}/basic/DAC/ReadCountTable.norm.corr.DACvsMOCK.analysed.txt

#subset DAC_PU1vsPU1 log transformed
cut -f1-13,28-32 ${RNADIR}/basic/DAC/ReadCountTable.norm.corr.DACvsMOCK.analysed.txt >
${RNADIR}/basic/DAC/ReadCountTable.norm.corr.DAC_PU1vsMOCK_PU1.analysed.txt

# generate heatmap with batch correction include logCPM cut off
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.heatmap.P.${_DATE}.R" <<EOF
library(gplots)
library(RColorBrewer)
x <- read.delim("${RNADIR}/basic/DAC/ReadCountTable.norm.corr.DAC_PU1vsMOCK_PU1.analysed.txt",
row.names="Gene", header = TRUE, dec = ".")
# Focus on genes that have an absolute fold change > 1.5, FDR <0.05, logCPM > 1
subdata <- subset(x,abs(logFC) > 1 & FDR < 0.05 & logCPM > 1,
select=c(CTV1_noDAC_PU1_rep1,CTV1_noDAC_PU1_rep2,CTV1_noDAC_PU1mut_rep1,CTV1_noDAC_PU1mut_rep2
,CTV1_DAC_PU1_rep1,CTV1_DAC_PU1_rep2,CTV1_DAC_PU1_rep3,CTV1_DAC_PU1mut_rep1,CTV1_DAC_PU1mut_re
p2,CTV1_DAC_PU1mut_rep3))
data.scaled <- (scale(t(subdata)))
data <- data.matrix(t(data.scaled))
colnames(data) <-
c("PU1_noDAC_A","PU1_noDAC_B","PU1mut_noDAC_A","PU1mut_noDAC_B","PU1_DAC_A","PU1_DAC_B","PU1_D
AC_C","PU1mut_DAC_A","PU1mut_DAC_B","PU1mut_DAC_C")
# Hierarchical clustering of Zscores
hc <- hclust(dist(t(data), method = "manhattan"), method="ward.D")
```

```
hr <- hclust(dist(data, method = "manhattan"), method="ward.D")
clustercol <- colorRampPalette(c("blue","white","red"))(299)
#col_breaks = c(seq(-1.1,-0.75,length=100), seq(-0.749,0.749,length=100),
seq(0.75,1.1,length=100))
col_breaks = c(seq(-1.5,-0.75,length=100), seq(-0.749,0.749,length=100),
seq(0.75,1.5,length=100))
# Mark clusters
mycl <- cutree(hr, h=max(hr\$height/3.5))
clusterCols <- rainbow(length(unique(mycl)))
myClusterSideBar <- clusterCols[mycl]
# Create data table for clustered data subset
splnames <- unlist(strsplit(as.character(rownames(t(data.scaled))),"[\$]"))
row.matrix <- matrix( splnames , ncol = 4 , byrow = TRUE )
colnames(row.matrix) <- c("EnsemblID","GeneSymbol","TranscriptLength","GeneType")
subdata.clustered <- cbind(t(data.scaled), row.matrix, clusterID=mycl)
clustered.data <- apply(subdata.clustered[hr\$order,], 2, rev)
write.table(clustered.data, file =
"${RNADIR}/basic/DAC/Table_ManhattanWardZscores_.qstat.corr.DAC_PU1vsMOCK_PU1allCPM.txt", sep
= "\t", col.names=NA, quote=FALSE)
pdf(file="${FIGURESDIR}/edgeR/edgeR_DAC_RNAseq_Heatmap_MW_Zscores_qstat.corr_DAC_PU1vsMOCK_PU1
.allCPM.pdf", height=8, width=8)
heatmap.2(data, Rowv=as.dendrogram(hr), Colv=as.dendrogram(hc), col = clustercol, breaks =
col_breaks, labRow = FALSE, na.rm=TRUE, scale="none", margins=c(8,5),  cexCol=1, key=TRUE,
density.info="none", trace="none", key.title = NA, key.xlab = "Z score",RowSideColors=
myClusterSideBar, main = "DGE in PU1-transfected cells\nabs(logFC) > 1 & FDR < 0.05 & logCPM >
1\n ")
dev.off()
EOF
chmod 750 "/loctmp/R.heatmap.P.${_DATE}.R"
R < /loctmp/R.heatmap.P.${_DATE}.R  --no-save
rm /loctmp/R.heatmap.P.${_DATE}.R

# Extracting lists of differentially expressed genes
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.extract.P.${_DATE}.R" <<EOF
list <- read.table("${RNADIR}/basic/DAC/qstat_PU1_DACvsPU1_MOCK.glm.txt", header=T, sep="\t")
splnames <- unlist(strsplit(as.character(list\$X),"[\$]"))
row.matrix <- matrix( splnames , ncol = 4 , byrow = TRUE )
colnames(row.matrix) <- c("EnsemblID","GeneSymbol","TranscriptLength","GeneType")
list <- cbind(list, row.matrix)
subdata <- subset(list,(logFC > 2 & FDR < 0.05), select=GeneSymbol)
write.table(as.character(subdata\$GeneSymbol), file =
"${RNADIR}/basic/DAC/Table_4foldup_PU1_DACvsPU1_MOCK.txt", sep = "\t", col.names=FALSE,
row.names=FALSE, quote=FALSE)
subdata <- subset(list,(logFC > 1.732 & FDR < 0.05), select=GeneSymbol)
write.table(as.character(subdata\$GeneSymbol), file =
"${RNADIR}/basic/DAC/Table_3foldup_PU1_DACvsPU1_MOCK.txt", sep = "\t", col.names=FALSE,
row.names=FALSE, quote=FALSE)
subdata <- subset(list,(logFC > 1 & FDR < 0.05), select=GeneSymbol)
write.table(as.character(subdata\$GeneSymbol), file =
"${RNADIR}/basic/DAC/Table_2foldup_PU1_DACvsPU1_MOCK.txt", sep = "\t", col.names=FALSE,
row.names=FALSE, quote=FALSE)
subdata <- subset(list,(logFC < -1 & logCPM > 1 & FDR < 0.05), select=GeneSymbol)
write.table(as.character(subdata\$GeneSymbol), file =
"${RNADIR}/basic/DAC/Table_2folddown_PU1_DACvsPU1_MOCK.txt", sep = "\t", col.names=FALSE,
row.names=FALSE, quote=FALSE)
EOF
chmod 750 "/loctmp/R.extract.P.${_DATE}.R"
R < /loctmp/R.extract.P.${_DATE}.R  --no-save
rm /loctmp/R.extract.P.${_DATE}.R

# generate volcano with batch correction
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.volcano.P.${_DATE}.R" <<EOF
library(ggplot2)
library(ggrepel)
res <- read.table("${RNADIR}/basic/DAC/qstat_PU1_DACvsPU1_MOCK.glm.txt", header=T, sep="\t")
genes <- read.table("${RNADIR}/basic/DAC/Table_2foldup_PU1_DACvsPU1_MOCK.txt", sep="\t")
genelist <- genes[["V1"]]
# Highlight genes that have an absolute fold change > 2 and FDR <0.05
res\$threshold = as.factor(abs(res\$logFC) > 1 & res\$FDR < 0.05)
res\$namethresh = as.factor(abs(res\$logFC) > 3  & res\$FDR < 0.05)
res\$fCategory <- factor(res\$threshold)
# split ID column
splnames <- unlist(strsplit(as.character(res\$X),"[\$]"))
row.matrix <- matrix( splnames , ncol = 4 , byrow = TRUE )
```

```
colnames(row.matrix) <- c("EnsemblID","GeneSymbol","TranscriptLength","GeneType")
res <- cbind(res, row.matrix)
# Construct the plot object
p <- ggplot(data=res, aes(x=logFC, y=-log10(FDR), color=threshold))
p <- p + theme_bw(base_size = 8, base_family = "Helvetica") + theme(legend.position = "none")
p <- p + geom_point(alpha=0.5, size=0.40) +
scale_color_manual(values=c("FALSE"="gray80","TRUE"="blue"))
p <- p + xlim(c(-5, 10.1)) + ylim(c(-0.01,5.5))
p <- p + xlab("log2 fold change") + ylab("-log10 q-value")
#p <- p + theme(panel.grid.major = element_line(size = .25, color = "grey"),panel.grid.minor =
element_line(size = .25, color = "grey"), panel.border = element_rect(size=.5, color =
"black"))
#p <- p + geom_text_repel(data=subset(res, res\$GeneSymbol %in% genelist) , aes(x=logFC, y=-
log10(FDR),label=GeneSymbol), segment.colour="black", segment.size=0.05,
min.segment.length=0.05, size=2, point.padding=.15, segment.alpha=0.5, alpha=1, color="black")
pdf(file="${FIGURESDIR}/edgeR/edgeR_DAC_RNAseq_Volcano_stat_PU1_DACvsPU1_MOCK.pdf",
height=2.8, width=2.8)
print(p)
dev.off()
EOF
chmod 750 "/loctmp/R.volcano.P.${_DATE}.R"
R < /loctmp/R.volcano.P.${_DATE}.R  --no-save
rm /loctmp/R.volcano.P.${_DATE}.R

# Comparison with blood cell CAGE data
# PU1 expression across FANTOM samples
makeBeanPlotFromCAGE.pl
/misc/data/analysis/generalStuff/annotation/FANTOM/HemaGeneExpressionGC19.txt
${RNADIR}/basic/DAC/Table_3foldup_PU1_DACvsPU1_MOCK.txt
bloodExpression.3foldup_PU1_DACvsPU1_MOCK ${FIGURESDIR}/edgeR -gene SPI1
makeBeanPlotFromCAGE.pl
/misc/data/analysis/generalStuff/annotation/FANTOM/HemaGeneExpressionGC19.txt
${RNADIR}/basic/DAC/Table_2foldup_PU1_DACvsPU1_MOCK.txt
bloodExpression.2foldup_PU1_DACvsPU1_MOCK ${FIGURESDIR}/edgeR -gene SPI1
```

## # basic analysis of ATAC & ChIPseq data

```
# CNV normalization of tagDirs
# normalize & reduce tagDir for better comparability with other data sets
declare -a oCHIPDIRS=("${TAGDIR}/CTV1_DAC_PU1_Flag_R1" "${TAGDIR}/CTV1_DAC_PU1_Flag_R2"
"${TAGDIR}/CTV1_DAC_PU1_Flag_R3")
declare -a oINPUTDIRS=("${TAGDIR}/CTV1_DAC_PU1mut_Flag_R1"
"${TAGDIR}/CTV1_DAC_PU1mut_Flag_R2" "${TAGDIR}/CTV1_DAC_PU1mut_Flag_R3")

COUNT=0
for SAMPLE in ${oCHIPDIRS[@]}; do
INPUT="${oINPUTDIRS[${COUNT}]}"
normalizeTagDirByCopyNumber.pl ${SAMPLE} -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl ${INPUT} -cnv "${CNVFILECTV1}" -remove
COUNT=$((COUNT+=1))
Done

cd ${TAGDIR}
makeTagDirectory CTV1_DAC_PU1_Flag_CNVnormRefChr_merged -d
CTV1_DAC_PU1_Flag_R1_CNVnormRefChr CTV1_DAC_PU1_Flag_R2_CNVnormRefChr
CTV1_DAC_PU1_Flag_R3_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_DAC_PU1mut_Flag_CNVnormRefChr_merged -d
CTV1_DAC_PU1mut_Flag_R1_CNVnormRefChr CTV1_DAC_PU1mut_Flag_R2_CNVnormRefChr
CTV1_DAC_PU1mut_Flag_R3_CNVnormRefChr -genome hg19 -checkGC

# normalization of PU.1_woDAC and ATAC tagDirs
cd ${TAGDIR}
normalizeTagDirByCopyNumber.pl CTV-1_PU.1_FLAG_S1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV-1_PU.1mut_FLAG_S1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV-1_PU.1_FLAG_S2 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV-1_Mock_FLAG_S2 -cnv "${CNVFILECTV1}" -remove

cd ${ATACTAGDIR}
normalizeTagDirByCopyNumber.pl CTV1_DAC_PU1_R1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_DAC_PU1_R2 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_DAC_PU1_R3 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_DAC_PU1mut_R1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_DAC_PU1mut_R2 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_DAC_PU1mut_R3 -cnv "${CNVFILECTV1}" -remove
```

```
# merge tag directories
cd ${TAGDIR}
makeTagDirectory CTV1_PU1_Flag_CNVnormRefChr_merged2 -d CTV-1_PU.1_FLAG_S2_CNVnormRefChr
CTV-1_PU.1_FLAG_S1_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_Flag_CNVnormRefChr_merged2 -d CTV-1_Mock_FLAG_S2_CNVnormRefChr
CTV-1_PU.1mut_FLAG_S1_CNVnormRefChr -genome hg19 -checkGC

cd ${ATACTAGDIR}
makeTagDirectory CTV1_DAC_PU1_CNVnormRefChr_merged -d CTV1_DAC_PU1_R1_CNVnormRefChr
CTV1_DAC_PU1_R2_CNVnormRefChr CTV1_DAC_PU1_R3_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_DAC_PU1mut_CNVnormRefChr_merged -d CTV1_DAC_PU1mut_R1_CNVnormRefChr
CTV1_DAC_PU1mut_R2_CNVnormRefChr CTV1_DAC_PU1mut_R3_CNVnormRefChr -genome hg19 -checkGC

# BigWigs for normalized TagDirs
TAGDIRSETS="CTV1_DAC_PU1_Flag_R1_CNVnormRefChr CTV1_DAC_PU1_Flag_R2_CNVnormRefChr
CTV1_DAC_PU1_Flag_R3_CNVnormRefChr \
CTV1_DAC_PU1mut_Flag_R1_CNVnormRefChr CTV1_DAC_PU1mut_Flag_R2_CNVnormRefChr
CTV1_DAC_PU1mut_Flag_R3_CNVnormRefChr \
CTV-1_PU.1_FLAG_S1_CNVnormRefChr CTV-1_PU.1_FLAG_S2_CNVnormRefChr
CTV-1_PU.1mut_FLAG_S1_CNVnormRefChr CTV-1_Mock_FLAG_S2_CNVnormRefChr "
for SAMPLE in ${TAGDIRSETS[@]}; do
makeUCSCfile ${CHIPTAGDIR}/$SAMPLE -bigWig $CHROMSIZES -o $BIGWIGDIR/$SAMPLE.bigwig
done

ATACDIRSETS="CTV1_DAC_PU1_R1_CNVnormRefChr CTV1_DAC_PU1_R2_CNVnormRefChr
CTV1_DAC_PU1_R3_CNVnormRefChr CTV1_DAC_PU1mut_R1_CNVnormRefChr
CTV1_DAC_PU1mut_R2_CNVnormRefChr CTV1_DAC_PU1mut_R3_CNVnormRefChr"
for SAMPLE in ${ATACDIRSETS[@]}; do
makeUCSCfile ${ATACTAGDIR}/$SAMPLE -bigWig $CHROMSIZES -fragLength 65 -o
$BIGWIGATAC/$SAMPLE.bigwig
done

# average bigwigs from triplicates
myAverageBigWig.pl -bw $BIGWIGATAC/CTV1_DAC_PU1_R1_CNVnormRefChr.bigwig \
$BIGWIGATAC/CTV1_DAC_PU1_R2_CNVnormRefChr.bigwig \
$BIGWIGATAC/CTV1_DAC_PU1_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGATAC/CTV1_DAC_aveATAC_PU1_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGATAC/CTV1_DAC_PU1mut_R1_CNVnormRefChr.bigwig \
$BIGWIGATAC/CTV1_DAC_PU1mut_R2_CNVnormRefChr.bigwig \
$BIGWIGATAC/CTV1_DAC_PU1mut_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGATAC/CTV1_DAC_aveATAC_PU1mut_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_DAC_PU1_Flag_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_DAC_PU1_Flag_R2_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_DAC_PU1_Flag_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_DAC_aveFlag_PU1_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_DAC_PU1mut_Flag_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_DAC_PU1mut_Flag_R2_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_DAC_PU1mut_Flag_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_DAC_aveFlag_PU1mut_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV-1_PU.1_FLAG_S1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV-1_PU.1_FLAG_S2_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_aveFlag2sample_PU1_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV-1_PU.1mut_FLAG_S1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV-1_Mock_FLAG_S2_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_aveFlag2sample_PU1mut_CNVnormRefChr.bigwig

# find ChIP peaks
cd ${TAGDIR}
findPeaks CTV1_DAC_PU1_Flag_CNVnormRefChr_merged -i
CTV1_DAC_PU1mut_Flag_CNVnormRefChr_merged -style factor -fdr 0.00001 -o
${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.peaks.txt
findPeaks CTV1_PU1_Flag_CNVnormRefChr_merged2 -i CTV1_Flag_CNVnormRefChr_merged2 -style
factor -fdr 0.00001 -o ${PEAKDIR}/PU1_Flag_merged_2sample.factor.fdr05.peaks.txt

# relevant peaks further filtered for peaks with at least 15 tags
myFilterFile.pl ${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.peaks.txt -column 6 -lowerlimit
15 > ${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.ntag15.peaks.txt
myFilterFile.pl ${PEAKDIR}/PU1_Flag_merged_2sample.factor.fdr05.peaks.txt -column 6
-lowerlimit 15 > ${PEAKDIR}/PU1_Flag_merged_2sample.factor.fdr05.ntag15.peaks.txt

# ChIP PU.1 peaks further filtered for mappability
pos2bed.pl ${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.ntag15.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Filtered out 3303 regions with mappability scores below 0.8, leaving 43561 regions.
pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt > ${DACPU1CTV1PEAKS}
pos2bed.pl ${PEAKDIR}/PU1_Flag_merged_2sample.factor.fdr05.ntag15.peaks.txt >
${TMPDIR}/tmp.3.bed
```

```
$BEDTOOLS intersect -a ${TMPDIR}/tmp.3.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.4.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.4.bed -g hg19 -f 0.8 -s 50
# Filtered out 2633 regions with mappability scores below 0.8, leaving 39955 regions.
pos2bed.pl ${TMPDIR}/tmp.4.mapScoreFiltered.txt > ${noDACPU1CTV1PEAKS}

# peaks induced by PU.1 in DAC treatment
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_DAC_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_DAC_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_DAC_PU1_Flag_R3_CNVnormRefChr \
-b ${TAGDIR}/CTV-1_PU.1mut_FLAG_S1_CNVnormRefChr ${TAGDIR}/CTV-1_PU.1_FLAG_S2_CNVnormRefChr
-genome hg19 \
-p "${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.ntag15.filtered.bed" >
"${DIFFDIR}/CTV1_DACspecificPU1.peaks.txt"
# Output Stats bg vs. target:
# Total Genes: 43561
# Total Up-regulated in target vs. bg: 1379 (3.166%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 0 (0.000%) [log2fold<-1, FDR<0.05]

# find ATAC peaks
for ENTRY in ${ATACTAGDIR}/*_CNVnormRefChr_merged
do
SAMPLEBASE=${ENTRY##*/}
echo "starting motif search in ${SAMPLEBASE}"
screen -dm -S ${SAMPLEBASE} bash -c "bash findATACpeaks.sh ${ENTRY} ${ATACPEAKDIR}"
done

# peaks induced by PU1 (against PU1mut) during DAC treatment
cd ${ATACPEAKDIR}
getDifferentialPeaksReplicates.pl -t ${ATACTAGDIR}/CTV1_DAC_PU1_R1_CNVnormRefChr
${ATACTAGDIR}/CTV1_DAC_PU1_R2_CNVnormRefChr ${ATACTAGDIR}/CTV1_DAC_PU1_R3_CNVnormRefChr \
-b ${ATACTAGDIR}/CTV1_DAC_PU1mut_R1_CNVnormRefChr
${ATACTAGDIR}/CTV1_DAC_PU1mut_R2_CNVnormRefChr
${ATACTAGDIR}/CTV1_DAC_PU1mut_R3_CNVnormRefChr -genome hg19 \
-p "${ATACPEAKDIR}/CTV1_DAC_PU1_CNVnormRefChr_merged.intersect.peaks.txt" >
"${ATACDIFFPEAKDIR}/CTV1_DAC_inducedByPU1.peaks.txt"
# Output Stats bg vs. target:
# Total Genes: 90113
# Total Up-regulated in target vs. bg: 33020 (36.643%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 3980 (4.417%) [log2fold<-1, FDR<0.05]

# peaks induced by PU.1 specific for DAC treatment
cd ${ATACPEAKDIR}
getDifferentialPeaksReplicates.pl -t ${ATACTAGDIR}/CTV1_DAC_PU1_R1_CNVnormRefChr
${ATACTAGDIR}/CTV1_DAC_PU1_R2_CNVnormRefChr ${ATACTAGDIR}/CTV1_DAC_PU1_R3_CNVnormRefChr \
-b ${ATACTAGDIR}/CTV1_PU1_R1_CNVnormRefChr ${ATACTAGDIR}/CTV1_PU1_R2_CNVnormRefChr
${ATACTAGDIR}/CTV1_PU1_R3_CNVnormRefChr -genome hg19 \
-p "${ATACPEAKDIR}/CTV1_DAC_PU1_CNVnormRefChr_merged.intersect.peaks.txt" >
"${ATACDIFFPEAKDIR}/CTV1_DACspecificPU1.peaks.txt"
# Output Stats bg vs. target:
# Total Genes: 90113
# Total Up-regulated in target vs. bg: 14186 (15.742%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 3750 (4.161%) [log2fold<-1, FDR<0.05]

# intersecting ATAC peak sets
pos2bed.pl ${ATACDIFFPEAKDIR}/CTV1_DACspecificPU1.peaks.txt >
${ATACDIFFPEAKDIR}/CTV1_DACspecificPU1.peaks.bed
pos2bed.pl ${ATACDIFFPEAKDIR}/CTV1_DAC_inducedByPU1.peaks.txt >
${ATACDIFFPEAKDIR}/CTV1_DAC_inducedByPU1.peaks.bed

cd ${ANALYSISDIR}
$BEDTOOLS intersect -a ${ATACDIFFPEAKDIR}/CTV1_DAC_inducedByPU1.peaks.bed -b
${ATACDIFFPEAKDIR}/CTV1_DACspecificPU1.peaks.bed > CTV1_DAC_PU1.ATACpeaks.overlap.bed
# 7614
$BEDTOOLS intersect -a ${ATACDIFFPEAKDIR}/CTV1_DAC_inducedByPU1.peaks.bed -b
${ATACDIFFPEAKDIR}/CTV1_DACspecificPU1.peaks.bed -v > CTV1_DAC_PU1.ATACpeaks.no.overlapA.bed
# 25406 PU.1_induced in DAC A
$BEDTOOLS intersect -a ${ATACDIFFPEAKDIR}/CTV1_DACspecificPU1.peaks.bed -b
${ATACDIFFPEAKDIR}/CTV1_DAC_inducedByPU1.peaks.bed -v >
CTV1_DAC_PU1.ATACpeaks.no.overlapB.bed
# 6572 DAC_specfic in DAC B

# venn diagram of differential ATAC peaks
mergePeaks -d 100 ${ATACDIFFPEAKDIR}/CTV1_DACspecificPU1.peaks.bed
${ATACDIFFPEAKDIR}/CTV1_DAC_inducedByPU1.peaks.bed -venn
${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.venn -matrix
${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.matrix -prefix ${ATACDIFFPEAKDIR}/peaks_sep >
${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.merged.txt
```

```
mv
${ATACDIFFPEAKDIR}/peaks_sep__misc_data_analysis_project_PU1_CTV1_ATAC_diffPeaksCNVcorr_CTV1_D
AC_inducedByPU1.peaks.bed ${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.PU1ind.txt
mv
${ATACDIFFPEAKDIR}/peaks_sep__misc_data_analysis_project_PU1_CTV1_ATAC_diffPeaksCNVcorr_CTV1_D
ACspecificPU1.peaks.bed ${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.DACspec.txt
mv
${ATACDIFFPEAKDIR}/peaks_sep__misc_data_analysis_project_PU1_CTV1_ATAC_diffPeaksCNVcorr_CTV1_D
ACspecificPU1.peaks.bed__misc_data_analysis_project_PU1_CTV1_ATAC_diffPeaksCNVcorr_CTV1_DAC_in
ducedByPU1.peaks.bed ${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.common.txt
pos2bed.pl ${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.PU1ind.txt >
${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.PU1ind.bed
pos2bed.pl ${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.DACspec.txt >
${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.DACspec.bed
pos2bed.pl ${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.common.txt >
${ATACDIFFPEAKDIR}/ATACpeaksDACspec_PU1ind.common.bed

_DATE=$(date +%s)
cat >"${TMPDIR}/R.venn.${_DATE}.R" <<EOF
library(VennDiagram)
venn <- venn.diagram(list(PU1 = 1:33020, DAC = 25407:39592),fill = c("blue", "slateblue4"),
alpha = c(0.5, 0.5), cex = 1.5, cat.cex=2, cat.fontface = 4, cat.fontfamily = "sans",lwd
=2, fontfamily = "sans", filename = NULL);
plot.new()
pdf(file="${FIGURESDIR}/PU1ind-DACspec.venn.pdf", width = 6, height = 6)
grid.draw(venn)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.venn.${_DATE}.R"
R < "${TMPDIR}/R.venn.${_DATE}.R" --no-save
rm "${TMPDIR}/R.venn.${_DATE}.R"

# overlap of differential ATAC peaks with PU.1 ChIP data of DAC-treated cells
cd ${ANALYSISDIR}
$BEDTOOLS intersect -a ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.overlap.bed -b
${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.ntag15.filtered.bed -wa -u >
CTV1_DAC_PU1.ATACpeaks.overlap.bound.PU1.bed
# 3722
$BEDTOOLS intersect -a ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.overlap.bed -b
${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.ntag15.filtered.bed -v >
CTV1_DAC_PU1.ATACpeaks.overlap.notbound.PU1.bed
# 3892
$BEDTOOLS intersect -a ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.bed -b
${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.ntag15.filtered.bed -wa -u >
CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1.bed
# 12575
$BEDTOOLS intersect -a ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.bed -b
${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.ntag15.filtered.bed -v >
CTV1_DAC_PU1.ATACpeaks.no.overlapA.notbound.PU1.bed
# 12831
$BEDTOOLS intersect -a ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.bed -b
${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.ntag15.filtered.bed -wa -u >
CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1.bed
# 973
$BEDTOOLS intersect -a ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.bed -b
${PEAKDIR}/DAC_PU1_Flag_merged.factor.fdr05.ntag15.filtered.bed -v >
CTV1_DAC_PU1.ATACpeaks.no.overlapB.notbound.PU1.bed
# 5599

# motif analyses of ATAC peaks and PU.1 bound vs non-bound peaks
# DAC_ATACpeaks induced by PU.1 and DAC
findMotifsGenome.pl ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.overlap.bed hg19r
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.overlap -size 200 -len 7,8,9,10,11,12,13,14 -p 4 -h
compareMotifs.pl ${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.overlap/homerMotifs.all.motifs
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.overlap/final -reduceThresh .75 -matchThresh .6 -pvalue
1e-12 -info 1.5 -cpu 2
# DAC_ATACpeaks induced by PU.1only
findMotifsGenome.pl ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.bed hg19r
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA -size 200 -len 7,8,9,10,11,12,13,14 -p 4 -h
compareMotifs.pl ${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA/homerMotifs.all.motifs
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA/final -reduceThresh .75 -matchThresh .6
-pvalue 1e-12 -info 1.5 -cpu 2
# DAC_ATACpeaks induced by DAConly
findMotifsGenome.pl ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.bed hg19r
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB -size 200 -len 7,8,9,10,11,12,13,14 -p 4 -h
compareMotifs.pl ${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB/homerMotifs.all.motifs
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB/final -reduceThresh .75 -matchThresh .6
-pvalue 1e-12 -info 1.5 -cpu 2
```

```
# DAC_ATACpeaks induced by PU.1 and DAC bound by PU.1
findMotifsGenome.pl ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.overlap.bound.PU1.bed hg19r
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.overlap.bound.PU1 -size 200 -len 7,8,9,10,11,12,13,14 -p
4 -h
compareMotifs.pl ${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.overlap.bound.PU1/homerMotifs.all.motifs
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.overlap.bound.PU1/final -reduceThresh .75 -matchThresh .6
-pvalue 1e-12 -info 1.5 -cpu 2
# DAC_ATACpeaks induced by PU.1 and DAC notbound by PU.1
findMotifsGenome.pl ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.overlap.notbound.PU1.bed hg19r
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.overlap.notbound.PU1 -size 200 -len 7,8,9,10,11,12,13,14
-p 4 -h
compareMotifs.pl
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.overlap.notbound.PU1/homerMotifs.all.motifs
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.overlap.notbound.PU1/final -reduceThresh .75 -matchThresh
.6 -pvalue 1e-12 -info 1.5 -cpu 2
# DAC_ATACpeaks induced by PU.1only bound by PU.1
findMotifsGenome.pl ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1.bed hg19r
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1 -size 200 -len 7,8,9,10,11,12,13,14
-p 4 -h
compareMotifs.pl
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1/homerMotifs.all.motifs
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1/final -reduceThresh .75
-matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 2
# DAC_ATACpeaks induced by PU.1only notbound by PU.1
findMotifsGenome.pl ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.notbound.PU1.bed hg19r
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.notbound.PU1 -size 200 -len
7,8,9,10,11,12,13,14 -p 4 -h
compareMotifs.pl
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.notbound.PU1/homerMotifs.all.motifs
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.notbound.PU1/final -reduceThresh .75
-matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 2
# DAC_ATACpeaks induced by DAConly bound by PU.1
findMotifsGenome.pl ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1.bed hg19r
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1 -size 200 -len 7,8,9,10,11,12,13,14
-p 4 -h
compareMotifs.pl
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1/homerMotifs.all.motifs
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1/final -reduceThresh .75
-matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 2
# DAC_ATACpeaks induced by DAConly notbound by PU.1
findMotifsGenome.pl ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.notbound.PU1.bed hg19r
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.notbound.PU1 -size 200 -len
7,8,9,10,11,12,13,14 -p 4 -h
compareMotifs.pl
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.notbound.PU1/homerMotifs.all.motifs
${MOTIFDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.notbound.PU1/final -reduceThresh .75
-matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 2

# overlap peaks with MAC PU.1 motif
cd ${ANALYSISDIR}
$BEDTOOLS intersect -a ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1.bed -b
${PU1MOTIFBED} -wa -u > CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1.PU1motif.bed
# 12042/12575
$BEDTOOLS intersect -a ${ANALYSISDIR}/CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1.bed -b
${PU1MOTIFBED} -wa -u > CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1.PU1motif.bed
# 874/973

# analysing the GC/CpG distribution across peaks sets overlapping with PU.1 motif
cd ${ANALYSISDIR}
annotatePeaks.pl CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1.PU1motif.bed hg19 -size given
-CpG > CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1.PU1motif.ann.txt -annStats
CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1.PU1motif.ann.stats.txt
annotatePeaks.pl CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1.PU1motif.bed hg19 -size given
-CpG > CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1.PU1motif.ann.txt -annStats
CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1.PU1motif.ann.stats.txt

cd ${ANALYSISDIR}
paste <(cut -f 20 CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1.PU1motif.ann.txt) \
<(cut -f 20 CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1.PU1motif.ann.txt) \
> "${TMP}tableCpG.txt"
tail -n +2 "${TMP}tableCpG.txt" > "${TMP}tableCpGnoHead.txt"
echo $'ATACpeaks.PU1induced\tATACpeaks.DACspecific.PU1' | cat - "${TMP}tableCpGnoHead.txt" >
ATAC-CpGCount-PU1motif.txt

_DATE=$(date +%s)
cat >"${TMPDIR}/R.bean.${_DATE}.R" <<EOF
setwd("${ANALYSISDIR}")
library(beanplot)
```

```
data <- read.table("ATAC-CpGCount-PU1motif.txt", header=T, sep="\t")
attach(data)
#data
pdf(file="${FIGURESDIR}/Beanplot_CpGinPU1-ATACpeaks.PU1motif.pdf", height=4, width=3)
plotcolors <- c("brown1","cyan3")
beanplot(as.data.frame(data),log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = "grey",
border = "grey" ,overallline = "median", method="jitter", boxwex = 1.25, beanlinewd = 1,
maxstripline = 0.8,ylim=c(0,.2),lwd=0.5)
par(new=TRUE)
boxplot(data,range=0,log="", style="quantile",axes = FALSE, col = plotcolors, border =
"black",overallline = "median", notch=TRUE, boxwex = 0.2, staplewex = 0.5,
ylim=c(0,.2),lwd=0.6)
axis(1,padj=-0.2,family="Helvetica",cex.axis=1,at=1:2,labels=c("PU.1induced","DACspecific"),la
s=2)
axis(2,padj=0.2,family="Helvetica",cex.axis=1)
mtext("% mCpG",family="Helvetica",side=2,line=2,cex=1.2,padj=0.4)
#mtext("peaks",family="Helvetica",ps=12,side=1,line=2,cex=1.2,padj=-0.6)
dev.off()
detach(data)
EOF
chmod 750 "${TMPDIR}/R.bean.${_DATE}.R"
R < ${TMPDIR}/R.bean.${_DATE}.R --no-save
rm ${TMPDIR}/R.bean.${_DATE}.R

cd ${ANALYSISDIR}
paste <(cut -f 21 CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1.PU1motif.ann.txt) \
<(cut -f 21 CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1.PU1motif.ann.txt) \
> "${TMP}tableCpG.txt"
tail -n +2 "${TMP}tableCpG.txt" > "${TMP}tableCpGnoHead.txt"
echo $'ATACpeaks.PU1induced\tATACpeaks.DACspecific.PU1' | cat - "${TMP}tableCpGnoHead.txt" >
ATAC-CGCount-PU1motif.txt

_DATE=$(date +%s)
cat >"${TMPDIR}/R.bean.${_DATE}.R" <<EOF
setwd("${ANALYSISDIR}")
library(beanplot)
data <- read.table("ATAC-CGCount-PU1motif.txt", header=T, sep="\t")
attach(data)
#data
pdf(file="${FIGURESDIR}/Beanplot_CGinPU1-ATACpeaks.PU1motif.pdf", height=4, width=3)
plotcolors <- c("brown1","cyan3")
beanplot(as.data.frame(data),log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = "grey",
border = "grey" ,overallline = "median", method="jitter", boxwex = 1.25, beanlinewd = 1,
maxstripline = 0.8,ylim=c(0,1),lwd=0.5)
par(new=TRUE)
boxplot(data,range=0,log="", style="quantile",axes = FALSE, col = plotcolors, border =
"black",overallline = "median", notch=TRUE, boxwex = 0.2, staplewex = 0.5,
ylim=c(0,1),lwd=0.6)
axis(1,padj=-0.2,family="Helvetica",cex.axis=1,at=1:2,labels=c("PU.1induced","DACspecific"),la
s=2)
axis(2,padj=0.2,family="Helvetica",cex.axis=1)
mtext("%GC",family="Helvetica",side=2,line=2,cex=1.2,padj=0.4)
#mtext("peaks",family="Helvetica",ps=12,side=1,line=2,cex=1.2,padj=-0.6)
dev.off()
detach(data)
EOF
chmod 750 "${TMPDIR}/R.bean.${_DATE}.R"
R < ${TMPDIR}/R.bean.${_DATE}.R --no-save
rm ${TMPDIR}/R.bean.${_DATE}.R

# overlap with CTV1 methylation data and PU1 ChIP data
# intersect between specific ATAC peaks and MCIp-detectable regions
cd ${ANALYSISDIR}
# DAC-specific ATAC peaks induced by PU.1 that were bound by PU.1
$BEDTOOLS intersect -a CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1.bed -b ${ALLMCIP} -wa -u
> CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1_MCIpOverlap.bed
# DAC-specific ATAC peaks induced by DAC that were bound by PU.1
$BEDTOOLS intersect -a CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1.bed -b ${ALLMCIP} -wa -u
> CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1_MCIpOverlap.bed

# hist plots
annotatePeaks.pl CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1_MCIpOverlap.bed hg19 -size
4200 -hist 25 -d ${CTV1MCIPDIR} >
CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1_MCIpOverlap.hist25.txt
annotatePeaks.pl CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1_MCIpOverlap.bed hg19 -size
4200 -hist 25 -d ${CTV1MCIPDIR} >
CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1_MCIpOverlap.hist25.txt
```

```
# comparison DAC-specific ATAC peaks that were bound by PU.1 induced with DAC or induced by
PU.1
_DATE=$(date +%s)
cat >"${TMPDIR}/R.hist.${_DATE}.R" <<EOF
setwd("${ANALYSISDIR}")
d1 <- read.table("CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1_MCIpOverlap.hist25.txt",
header=T, sep="\t")
data1 <- d1[,1:2]
d2 <- read.table("CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1_MCIpOverlap.hist25.txt",
header=T, sep="\t")
data2 <- d2[,2]
data <- cbind(data1,data2)
colnames(data) <- c("Distance","PU1","DAC.PU1")
plotcolors <- c("brown1","cyan3")
attach(data)
pdf(file="${FIGURESDIR}/Histplot_mCpGinPU1-ATACpeaks.inducedByDACvsPU1.bound.pdf", height=4,
width=4)
par(mar=c(3,3,.2,.2))
plot(Distance,PU1,type="l",col=plotcolors[1],xaxt="n",yaxt="n",xlab="",axes=FALSE,
ylab="",lwd=3, lty=1,ylim=c(1,8))
par(mar=c(3,3,.2,.2), new=TRUE)
plot(Distance,DAC.PU1,type="l",col=plotcolors[2],xaxt="n",yaxt="n",xlab="",axes=FALSE,
ylab="",lwd=3, lty=1,ylim=c(1,8))
axis(1,padj=-1.0,family="Helvetica",cex.axis=1,at=c(-2000,0,2000))
axis(2,padj=0.8,family="Helvetica",cex.axis=1, col="black", col.axis="black")
mtext("MCIp coverage",family="Helvetica", col="black", side=2,line=2,cex=1.3,padj=0.2)
mtext("Distance from ATAC peak
center",family="Helvetica",ps=12,side=1,line=2,cex=1.3,padj=-0.2)
legend("topleft", inset=.01, c("DACspecific","PU.1induced"), lwd=2, lty=1, cex=1.0, col =
c("cyan3","brown1"),bty="n")
dev.off()
detach(data)
EOF
chmod 750 "${TMPDIR}/R.hist.${_DATE}.R"
R < ${TMPDIR}/R.hist.${_DATE}.R --no-save
rm ${TMPDIR}/R.hist.${_DATE}.R

# intersect between specific ATAC peaks and PU.1 ChIP data bound
# hist plots
annotatePeaks.pl CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1.bed hg19 -size 4200 -hist 25
-d ${CTV1DACPU1DIR} > CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1_PU1overlap.hist25.txt
annotatePeaks.pl CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1.bed hg19 -size 4200 -hist 25
-d ${CTV1DACPU1DIR} > CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1_PU1overlap.hist25.txt

# comparison DAC-specific ATAC peaks that were bound by PU.1 induced with DAC or induced by
PU.1
_DATE=$(date +%s)
cat >"${TMPDIR}/R.hist.${_DATE}.R" <<EOF
setwd("${ANALYSISDIR}")
d1 <- read.table("CTV1_DAC_PU1.ATACpeaks.no.overlapA.bound.PU1_PU1overlap.hist25.txt",
header=T, sep="\t")
data1 <- d1[,1:2]
d2 <- read.table("CTV1_DAC_PU1.ATACpeaks.no.overlapB.bound.PU1_PU1overlap.hist25.txt",
header=T, sep="\t")
data2 <- d2[,2]
data <- cbind(data1,data2)
colnames(data) <- c("Distance","PU1","DAC.PU1")
plotcolors <- c("brown1","cyan3")
attach(data)
pdf(file="${FIGURESDIR}/Histplot_ChIPcovinPU1-ATACpeaks.inducedByDACvsPU1.bound.pdf",
height=4, width=4)
par(mar=c(3,3,.2,.2))
plot(Distance,PU1,type="l",col=plotcolors[1],xaxt="n",yaxt="n",xlab="",axes=FALSE,
ylab="",lwd=3, lty=1,ylim=c(1,12))
par(mar=c(3,3,.2,.2), new=TRUE)
plot(Distance,DAC.PU1,type="l",col=plotcolors[2],xaxt="n",yaxt="n",xlab="",axes=FALSE,
ylab="",lwd=3, lty=1,ylim=c(1,12))
axis(1,padj=-1.0,family="Helvetica",cex.axis=1,at=c(-2000,0,2000))
axis(2,padj=0.8,family="Helvetica",cex.axis=1, col="black", col.axis="black")
mtext("ChIPseq coverage",family="Helvetica", col="black", side=2,line=2,cex=1.3,padj=0.2)
mtext("Distance from ATAC peak
center",family="Helvetica",ps=12,side=1,line=2,cex=1.3,padj=-0.2)
legend("topleft", inset=.01, c("DACspecific","PU.1induced"), lwd=2, lty=1, cex=1.0, col =
c("cyan3","brown1"),bty="n")
dev.off()
detach(data)
EOF
chmod 750 "${TMPDIR}/R.hist.${_DATE}.R"
```

```
R < ${TMPDIR}/R.hist.${_DATE}.R --no-save
rm ${TMPDIR}/R.hist.${_DATE}.R
```

### 10.1.5 Analysis of NGS Data of Conventional CTV-1 Cells

The following script was used to analyze the binding properties of PU.1 in CTV-1 cells. The script includes parts of edgeR (Robinson et al. 2010), the HOMER suite (Heinz et al. 2010) as well as parts of the R software (R Development Core Team 2008).

```
#!/bin/bash

#setting homer environment
DIR_PKG="/misc/software/ngs"
PATH_PERL=/misc/software/package/perl/perl-5.26.1/bin
PATH_SAMTOOLS=${DIR_PKG}/samtools/samtools-1.6/bin
PATH_HOMER=${DIR_PKG}/homer/v4.9/bin
PATH_R=/misc/software/package/RBioC/3.4.3/bin
export PATH=${PATH_R}:${PATH_PERL}:${PATH_SAMTOOLS}:${PATH_HOMER}:${PATH}
export PATH

BEDTOOLS="/misc/software/ngs/bedtools/bedtools2-2.27.1/bin/bedtools"
WORKDIR="/misc/data/analysis/project_PU1/CTV1"
TAGDIR="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/RNAtransfection/PU1"
CHIPDIR="${WORKDIR}/ChIP"
ATACDIR="${WORKDIR}/ATAC"
RNADIR="${WORKDIR}/RNAseq"
EXPRESSION="/misc/data/analysis/project_PU1/CTV1/RNAseq/basic"
ANALYSISDIR="${WORKDIR}/analysis/basic"
CHROMSIZES="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/hg19.chrom.sizes"
TMPDIR="/loctmp"
TMP="/loctmp/${RANDOM}.tmp."
FIGURESDIR="${WORKDIR}/figures/basic"
ALLMCIP="/misc/data/analysis/generalStuff/MCIpDetectable/All_MCIp_detected.bed"
CTV1MCIPDIR="/misc/data/processedData/tagDir/DNA/hg19/MCIp/CellLines/MCIp_CTV1_603"
CNVFILECTV1="/misc/data/processedData/mapping/DNA/hg19/Input/CellLines/CNVdata/CTV1_genInput_6
20.sam_CNVs"
PU1MOTIF="/misc/data/analysis/project_PU1/PU1long.motif"
PU1MOTIFR="/misc/data/analysis/project_PU1/PU1long.rev.motif"
BLACKLIST_HG19="/misc/data/analysis/generalStuff/annotation/hg19/hg19.blacklist.bed"
HG19TRANS="/misc/data/analysis/generalStuff/annotation/hg19/gencode.v19.transcripts.txt"
HG19GTEX="/misc/data/analysis/generalStuff/annotation/hg19/GTEx_v7_eQTL_WBA.snpgenes.1000.bed"
PEAKDIR="${CHIPDIR}/peaksCNVcorr"
DIFFDIR="${CHIPDIR}/diffPeaksCNVcorr"
MOTIFDIR="${WORKDIR}/motifs/basic"
KNOWNTFDIR="/misc/software/ngs/homer/v4.9/data/knownTFs/motifs"
ATACTAGDIR="/misc/data/processedData/tagDir/chromatin/hg19/ATAC/RNAtransfection/PU1/"
CHIPTAGDIR="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/RNAtransfection/PU1/"
CHIPINDIR="/misc/data/processedData/tagDir/DNA/hg19/Input/RNAtransfection/PU1/"
BIGWIGDIR="/misc/data/processedData/bigWig/chromatin/hg19/ChIP/RNAtransfection/PU1/"
BIGWIGATAC="/misc/data/processedData/bigWig/chromatin/hg19/ATAC/RNAtransfection/PU1/"
PU1CTV1PEAKS="${PEAKDIR}/PU1_Flag_merged.factor.fdr05.ntag15.filtered.bed"
TAGDIRSETS="${CHIPTAGDIR}/CTV1_PU1mut_Flag_merged ${CHIPTAGDIR}/CTV1_PU1_Flag_merged
${ATACTAGDIR}/CTV1_PU1mut_merged ${ATACTAGDIR}/CTV1_PU1_merged
${CHIPTAGDIR}/CTV1_PU1mut_H3K27ac_merged ${CHIPTAGDIR}/CTV1_PU1_H3K27ac_merged"
HOMOCLUSTERDIR="/misc/data/analysis/project_PU1/homotypicClusters/"
GHISTDIR="${WORKDIR}/analysis/basic/ghist"
HISTDIR="${WORKDIR}/analysis/basic/hist"
HG19GTF="/misc/software/ngs/genome/annotation/hg19/gencode.v19.annotation.gtf"
CGIHG19="/misc/data/analysis/generalStuff/annotation/hg19/CGI_hg19_UCSC.bed"
GENOME_HG19="/misc/software/ngs/genome/sequence/hg19/hg19.fa"
PHYLOP="/misc/data/analysis/generalStuff/conservation/hg19/hg19.100way.phyloP100way.bedGraph"
PHASTCONS="/misc/data/analysis/generalStuff/conservation/hg19/hg19.100way.phastCons.bedGraph"
METASCAPE="/misc/data/analysis/project_PU1/CTV1/analysis/basic/Metascape"

# defining the colors
colPU1="blue"
colPU16u="darkblue"
colPU109u="dodgerblue1"
colDelP="blueviolet"
colDelQ="forestgreen"
colDelA="orange4"
```

```
colDelAQP="firebrick"
colMUT="gray50"
colClus="limegreen"
colSmot="mediumslateblue"
colNmot="black"
colCGI="darkgreen"
colnCGI="lightseagreen"
colCluM="springgreen4"
colETS1="slateblue4"

# directory for manually annotated motif lists
CUSTOMMOTIFS="/misc/data/analysis/project_PU1/allCellTypes/motifs/customAnnotation"

mkdir ${ANALYSISDIR}
mkdir ${MOTIFDIR}
mkdir -p ${FIGURESDIR}/R
mkdir ${FIGURESDIR}/edgeR
mkdir ${FIGURESDIR}/hist
mkdir ${GHISTDIR}
mkdir ${HISTDIR}
mkdir -p ${RNADIR}/basic/
mkdir ${ANALYSISDIR}/peakAssociatedGenes/


# basic analysis of RNAseq data

cd "/misc/data/processedData/mapping/RNA/hg19/RNAseq/RNAtransfection/PU1"
paste /misc/software/ngs/genome/sequence/hg19/STAR_transcriptIDshort.txt \
UA122_CTV1_mock.ReadsPerGene.txt \
CG12_CTV1_mock.trimmed.ReadsPerGene.txt \
UA120_CTV1_PU1mRNA.ReadsPerGene.txt \
CG10_CTV1_PU1mRNA.trimmed.ReadsPerGene.txt \
UA121_CTV1_mutPU1mRNA.ReadsPerGene.txt \
CG11_CTV1_mutPU1mRNA.trimmed.ReadsPerGene.txt \
JM112_CTV1_PU1mRNA_rep.ReadsPerGene.txt \
JM113_CTV1_mutPU1mRNA_rep.ReadsPerGene.txt \
> /loctmp/tmpReadCountTable_Raw.txt

cut -f 1,4,8,12,16,20,24,28,32,36,40 /loctmp/tmpReadCountTable_Raw.txt >
/loctmp/tmpReadCountTable_RawRed.txt

tail -n +5 /loctmp/tmpReadCountTable_RawRed.txt > /loctmp/tmpReadCountTable_RawRedDel.txt

echo
$'Gene\tCTV1_woPU1_rep2\tCTV1_woPU1_rep1_B\tCTV1_PU1_rep2\tCTV1_PU1_rep1_B\tCTV1_PU1mut_rep2\t
CTV1_PU1mut_rep1_B\tCTV1_PU1Rep\tCTV1_PU1mutRep' | cat -
/loctmp/tmpReadCountTable_RawRedDel.txt > ${RNADIR}/basic/ReadCountTable.txt

# basic analysis edgeR
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.edgeR.P.${_DATE}.R" <<EOF
library(edgeR)
library(ggplot2)
library(ggrepel)
data <- read.delim("${RNADIR}/basic/ReadCountTable.txt", row.names="Gene")
group <- factor(c(rep("CTV1_woPU1",2), rep("CTV1_PU1",2), rep("CTV1_PU1mut",2),
rep("CTV1_PU1Rep",1), rep("CTV1_PU1mutRep",1)))
batch <- factor(c(1,2,1,2,1,2,1,1))
treatment <- factor(c(rep("CTV1_woPU1",2), rep("CTV1_PU1",2), rep("CTV1_PU1mut",2),
rep("CTV1_PU1Rep",1), rep("CTV1_PU1mutRep",1)))
d <- DGEList(counts=data,group=group)
keep <- rowSums(cpm(d) > 1) >= 2
summary(keep)
d <- d[keep, , keep.lib.sizes=FALSE]
d <- calcNormFactors(d)
d$samples
counts.per.m <- cpm(d, normalized.lib.sizes=TRUE)
logcpm <- cpm(d, prior.count=2, log=TRUE)

# removing batch effect for heatmaps etc.
design=model.matrix(~treatment)
corr.cpm <- removeBatchEffect(counts.per.m, batch=batch, design=design)
corr.logcpm <- removeBatchEffect(logcpm, batch=batch, design=design)
write.table (corr.cpm, file = "${RNADIR}/basic/ReadCountTable.norm.corr.cpm.txt", sep =
"\t", col.names=NA, quote=FALSE)
write.table (corr.logcpm, file = "${RNADIR}/basic/ReadCountTable.norm.corr.logcpm.txt", sep
= "\t", col.names=NA, quote=FALSE)
write.table (corr.logcpm, file =
"${RNADIR}/basic/ReadCountTable.norm.corr.logcpm.expression", sep = "\t", col.names=NA,
quote=FALSE)
corr.logcpm.scaled <- (scale(t(corr.logcpm)))
corr.logdata.scaled <- data.matrix(t(corr.logcpm.scaled))
write.table (corr.logdata.scaled, file =
"${RNADIR}/basic/ReadCountTable.norm.corr.zscore.txt", sep = "\t", col.names=NA, quote=FALSE)

# Extracting expression data for ETS factors and IRF4/8
rm ${RNADIR}/basic/ETS_ReadCountTable.norm.corr.rpkm.txt
touch ${RNADIR}/basic/ETS_ReadCountTable.norm.corr.rpkm.txt
rm ${RNADIR}/basic/ETS_notDetected.txt
touch ${RNADIR}/basic/ETS_notDetected.txt
echo -e
$'Gene\tCTV1_woPU1_rep2\tCTV1_woPU1_rep1_B\tCTV1_PU1_rep2\tCTV1_PU1_rep1_B\tCTV1_PU1mut_rep2\t
CTV1_PU1mut_rep1_B\tCTV1_PU1Rep\tCTV1_PU1mutRep' >>
${RNADIR}/basic/ETS_ReadCountTable.norm.corr.rpkm.txt

declare -a ETSLIST=(ETS1 ETS2 ETV1 ETV2 ETV3 ETV4 ETV5 ELK1 ELK3 ELK4 ERF ERG FEV FLI1 GABPA
EHF ELF1 ELF2 ELF3 ELF4 ELF5 ETV6 ETV7 SPI1 SPIB SPIC SPDEF IRF4 IRF8)
for FACTOR in ${ETSLIST[@]}; do
# extracting rpkm values
awk -F '[\$\t]' -v "key=$FACTOR" 'BEGIN{OFS="\t"} $2 == key {print
$2,2^$5/$3*1000,2^$6/$3*1000,2^$7/$3*1000,2^$8/$3*1000,2^$9/$3*1000,2^$10/$3*1000,2^$11/$3*100
0,2^$12/$3*1000}' ${RNADIR}/basic/ReadCountTable.norm.corr.logcpm.txt >>
${RNADIR}/basic/ETS_ReadCountTable.norm.corr.rpkm.txt
done
for FACTOR in ${ETSLIST[@]}; do
if grep -q $FACTOR <(cut -f1 ${RNADIR}/basic/ETS_ReadCountTable.norm.corr.rpkm.txt)
then
echo -e "$FACTOR found\n"
else
echo -e "$FACTOR NOT found\n"
echo ${FACTOR} >> ${RNADIR}/basic/ETS_notDetected.txt
fi
done
cut -f1-5 ${RNADIR}/basic/ETS_ReadCountTable.norm.corr.rpkm.txt > ${TMPDIR}/etsExpression.txt

cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.strip.${_DATE}.R" <<EOF
library(ggplot2)
library(reshape2)
d <- read.table("${TMPDIR}/etsExpression.txt", header=T, sep="\t")
data <- melt(cbind(d[,-1], annotation = d\$Gene), id.vars = c('annotation'))
data\$replicates=grepl("woPU", data\$variable, fixed=TRUE)
ylabel = expression("expression (RPKM)")
xlabel = expression("ETS transcription factor")
data
##Construct the plot object
p <- ggplot(data, aes(x=factor(annotation,levels=unique(annotation)), y=value,
color=replicates, alpha=replicates)) + coord_trans(y="log2")
p <- p + theme_bw(base_size = 12, base_family = "Helvetica") + theme(legend.position =
"none", panel.grid.minor.y = element_blank()) + coord_cartesian(ylim=c(0.125,128))
p <- p + scale_y_continuous(trans = 'log2', breaks = c(.125,.25,.5,1,2,4,8,16,32,64,128),
labels = c(.125,.25,.5,1,2,4,8,16,32,64,128))
p <- p + scale_color_manual(values=c("TRUE"="gray50","FALSE"="blue"))
p <- p + scale_alpha_manual(values=c("TRUE"=1,"FALSE"=.65)) +
geom_jitter(position=position_jitter(0.2),size=.75)
p <- p + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
p <- p + labs(y = ylabel, x = xlabel)
pdf(file="${FIGURESDIR}/edgeR/etsExpression.pdf", height=2.2, width=3)
print(p)
dev.off()
EOF
chmod 750 "/loctmp/R.strip.${_DATE}.R"
R < /loctmp/R.strip.${_DATE}.R --no-save
rm /loctmp/R.strip.${_DATE}.R

# PCA with batch correction
prin_comp_wbatch <- prcomp(corr.logcpm, scale. = T)
PCAcolors <- c("red","red","blue","blue","goldenrod","goldenrod","blue","goldenrod")
embedding_w <- as.data.frame(prin_comp_wbatch\$rotation)
embedding_w\$Class <- as.factor(treatment)
embedding_w\$Color <- as.factor(PCAcolors)
p1 <- ggplot(embedding_w, aes(x=PC1, y=PC2, label=treatment)) +
geom_point(size=3, col=embedding_w\$Color) +
# geom_text(aes(label=colnames(corr.logcpm)), vjust = 1.75, col=embedding_w\$Color) +
geom_text_repel(aes(label=colnames(corr.logcpm)), col=embedding_w\$Color,
segment.colour="black", segment.size=0.2, min.segment.length=0.2, size=2.5,
point.padding=.5, segment.alpha=0.5, alpha=1, size=12) +
ggtitle("Principal Component Analysis\nlogCPM with batch correction") +
theme_light(base_size=16) # +
# xlim(0.348, 0.356) + ylim(-0.5, 0.75)
plot(p1, labels=TRUE)
pdf(file="${FIGURESDIR}/edgeR/edgeR_basicRNAseq_PCA_withBatchCorr.pdf", height=5, width=5)
plot(p1, labels=TRUE)
dev.off()

# DGE analysis
design <- model.matrix(~0 + treatment + batch)
rownames(design) <- colnames(d)
d <- estimateDisp(d, design, robust=TRUE)
d\$common.dispersion
fit <- glmQLFit(d, design)
con <- makeContrasts(treatmentCTV1_PU1 - treatmentCTV1_PU1mut, levels=design)
qlf.2vs3 <- glmQLFTest(fit, contrast=con)
qstat.2vs3 <- topTags(qlf.2vs3, n=Inf)
write.table (qstat.2vs3, file = "${RNADIR}/basic/qstat_PU1vsPU1mut.glm.txt", sep = "\t",
col.names=NA, quote=FALSE)
summary(qdt.2vs3 <- decideTestsDGE(qlf.2vs3))
qisDE.2vs3 <- as.logical(qdt.2vs3)
qDEnames.2vs3 <- rownames(d)[qisDE.2vs3]
pdf(file="${FIGURESDIR}/edgeR/edgeR_basicRNAseq_MvA_withBatchCorr.pdf", height=5, width=5)
plotSmear(qlf.2vs3, de.tags=qDEnames.2vs3)
abline(h=c(-1,1), col="blue")
dev.off()
EOF
chmod 750 "/loctmp/R.edgeR.P.${_DATE}.R"
R < /loctmp/R.edgeR.P.${_DATE}.R --no-save
rm /loctmp/R.edgeR.P.${_DATE}.R

# extract expressed genes from table
awk -F '[\$]' 'BEGIN{OFS="\t"}{print $2}'
${RNADIR}/basic/ReadCountTable.norm.corr.logcpm.txt > ${RNADIR}/basic/expressedGenesList.txt

# assemble gene count table including statistics
cd ${RNADIR}/basic/
(head -n 1 qstat_PU1vsPU1mut.glm.txt && tail -n +2 qstat_PU1vsPU1mut.glm.txt | sort -k 1,1 )
> /loctmp/tmpqstat_PU1vsPU1mut.glm.txt;
(head -n 1 ReadCountTable.norm.corr.cpm.txt && tail -n +2 ReadCountTable.norm.corr.cpm.txt |
sort -k 1,1 ) > /loctmp/tmpsortedReadCountTable.norm.cpm.txt;
(head -n 1 ReadCountTable.norm.corr.logcpm.txt && tail -n +2
ReadCountTable.norm.corr.logcpm.txt | sort -k 1,1 ) >
/loctmp/tmpsortedReadCountTable.norm.logcpm.txt;
paste /loctmp/tmpsortedReadCountTable.norm.logcpm.txt
/loctmp/tmpsortedReadCountTable.norm.cpm.txt /loctmp/tmpqstat_PU1vsPU1mut.glm.txt >
/loctmp/tmpsortedReadCountTable.norm.analysed.txt
sed -e '1s/^/Gene/g' /loctmp/tmpsortedReadCountTable.norm.analysed.txt >
ReadCountTable.norm.corr.analysed.txt

# subset PU1vsPU1mut log transformed
cut -f1-9,20-24 ReadCountTable.norm.corr.analysed.txt >
ReadCountTable.norm.corr.PU1vsPU1mutanalysed.txt

# generate heatmap with batch correction
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.heatmap.P.${_DATE}.R" <<EOF
library(gplots)
library(RColorBrewer)
x <- read.delim("${RNADIR}/basic/ReadCountTable.norm.corr.PU1vsPU1mutanalysed.txt",
row.names="Gene", header = TRUE, dec = ".")
# Focus on genes that have an absolute fold change > 1.5, FDR <0.05
subdata <- subset(x,abs(logFC) > 1 & FDR < 0.05,
select=c(CTV1_woPU1_rep2,CTV1_woPU1_rep1_B,CTV1_PU1_rep2,CTV1_PU1_rep1_B,CTV1_PU1mut_rep2,CTV1
_PU1mut_rep1_B,CTV1_PU1mutRep,CTV1_PU1Rep))
data.scaled <- (scale(t(subdata)))
data <- data.matrix(t(data.scaled))
colnames(data) <-
c("Control_A","Control_B","PU1_A","PU1_B","mutPU1_A","mutPU1_B","mutPU1rep","PU1rep")
# Hierarchical clustering of Zscores
hc <- hclust(dist(t(data), method = "manhattan"), method="ward.D")
hr <- hclust(dist(data, method = "manhattan"), method="ward.D")
clustercol <- colorRampPalette(c("blue","white","red"))(299)
```

```
col_breaks = c(seq(-1.5,-0.75,length=100), seq(-0.749,0.749,length=100),
seq(0.75,1.5,length=100))
# Mark clusters
mycl <- cutree(hr, h=max(hr\$height/3.5))
clusterCols <- rainbow(length(unique(mycl)))
myClusterSideBar <- clusterCols[mycl]
# Create data table for clustered data subset
splnames <- unlist(strsplit(as.character(rownames(t(data.scaled))),"[\$]"))
row.matrix <- matrix( splnames , ncol = 4 , byrow = TRUE )
colnames(row.matrix) <- c("EnsemblID","GeneSymbol","TranscriptLength","GeneType")
subdata.clustered <- cbind(t(data.scaled), row.matrix, clusterID=mycl)
clustered.data <- apply(subdata.clustered[hr\$order,], 2, rev)
write.table(clustered.data, file =
"${RNADIR}/basic/Table_ManhattanWardZscores_.qstat.corr_PU1mutvsPU1all.txt", sep = "\t",
col.names=NA, quote=FALSE)
pdf(file="${FIGURESDIR}/edgeR/edgeR_basicRNAseq_Heatmap_MW_Zscores_qstat.corr_PU1mutvsPU1all.p
df", height=8, width=8)
heatmap.2(data, Rowv=as.dendrogram(hr), Colv=as.dendrogram(hc), col = clustercol, breaks =
col_breaks, labRow = FALSE, na.rm=TRUE, scale="none", margins=c(8,5), cexCol=1, key=TRUE,
density.info="none", trace="none", key.title = NA, key.xlab = "Z score",RowSideColors=
myClusterSideBar, main = "DGE in PU1mut-/PU1-transfected cells\nabs(logFC) > 1 & FDR <
0.05\n ")
dev.off()
EOF
chmod 750 "/loctmp/R.heatmap.P.${_DATE}.R"
R < /loctmp/R.heatmap.P.${_DATE}.R --no-save
rm /loctmp/R.heatmap.P.${_DATE}.R

# generate volcano with batch correction
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.volcano.P.${_DATE}.R" <<EOF
library(ggplot2)
library(ggrepel)
res <- read.table("${RNADIR}/basic/qstat_PU1vsPU1mut.glm.txt", header=T, sep="\t")
genes <- read.table("${RNADIR}/basic/myeloidGenes.txt", sep="\t")
genelist <- genes[["V1"]]
# Highlight genes that have an absolute fold change > 2 and FDR <0.05
res\$threshold = as.factor(abs(res\$logFC) > 1 & res\$FDR < 0.05)
res\$namethresh = as.factor(abs(res\$logFC) > 3 & res\$FDR < 0.05)
res\$fCategory <- factor(res\$threshold)
# split ID column
splnames <- unlist(strsplit(as.character(res\$X),"[\$]"))
row.matrix <- matrix( splnames , ncol = 4 , byrow = TRUE )
colnames(row.matrix) <- c("EnsemblID","GeneSymbol","TranscriptLength","GeneType")
res <- cbind(res, row.matrix)
# Construct the plot object
p <- ggplot(data=res, aes(x=logFC, y=-log10(FDR), color=threshold))
#p <- p + ggtitle("Volcano Plot\nDGE in SIK3-siRNA treated cells/control")
p <- p + theme_bw(base_size = 8, base_family = "Helvetica") + theme(legend.position = "none")
p <- p + geom_point(alpha=0.5, size=0.40) +
scale_color_manual(values=c("FALSE"="gray80","TRUE"="blue"))
p <- p + xlim(c(-3, 10)) + ylim(c(0, 6.5))
p <- p + xlab("log2 fold change") + ylab("-log10 q-value")
p <- p + theme(panel.grid.major = element_line(size = .25, color = "grey"),panel.grid.minor
= element_line(size = .25, color = "grey"), panel.border = element_rect(size=.5, color =
"black"))
#p <- p + geom_text_repel(data=subset(res,res\$namethresh=="TRUE") , aes(x=logFC,
y=-log10(FDR),label=GeneSymbol), segment.colour="black", segment.size=0.2,
min.segment.length=0.2, size=2.5, point.padding=.05, segment.alpha=0.5, alpha=1, size=3,
color="black")
p <- p + geom_text_repel(data=subset(res, res\$GeneSymbol %in% genelist) , aes(x=logFC,
y=-log10(FDR),label=GeneSymbol), segment.colour="black", segment.size=0.05,
min.segment.length=0.05, size=2, point.padding=.15, segment.alpha=0.5, alpha=1, color="black")
#p <- p + annotate("text", x = -1 , y = 7, label ="FDR<0.05, absFC>2", size = 3,
colour="blue")
pdf(file="${FIGURESDIR}/edgeR/edgeR_basicRNAseq_Volcano_stat_PU1vsPU1mut.pdf", height=3,
width=4)
print(p)
dev.off()
EOF
chmod 750 "/loctmp/R.volcano.P.${_DATE}.R"
R < /loctmp/R.volcano.P.${_DATE}.R --no-save
rm /loctmp/R.volcano.P.${_DATE}.R

# Extracting lists of differentially expressed genes
cd ${ANALYSISDIR}
_DATE=$(date +%s)
```

```
cat >"/loctmp/R.extract.P.${_DATE}.R" <<EOF
list <- read.table("${RNADIR}/basic/qstat_PU1vsPU1mut.glm.txt", header=T, sep="\t")
splnames <- unlist(strsplit(as.character(list\$X),"[\$]"))
row.matrix <- matrix( splnames , ncol = 4 , byrow = TRUE )
colnames(row.matrix) <- c("EnsemblID","GeneSymbol","TranscriptLength","GeneType")
list <- cbind(list, row.matrix)
subdata <- subset(list,(logFC > 2 & FDR < 0.05), select=GeneSymbol)
write.table(as.character(subdata\$GeneSymbol), file =
"${RNADIR}/basic/Table_4foldup_PU1mutvsPU1.txt", sep = "\t", col.names=FALSE,
row.names=FALSE, quote=FALSE)
subdata <- subset(list,(logFC > 1.732 & FDR < 0.05), select=GeneSymbol)
write.table(as.character(subdata\$GeneSymbol), file =
"${RNADIR}/basic/Table_3foldup_PU1mutvsPU1.txt", sep = "\t", col.names=FALSE,
row.names=FALSE, quote=FALSE)
subdata <- subset(list,(logFC > 1 & FDR < 0.05), select=GeneSymbol)
write.table(as.character(subdata\$GeneSymbol), file =
"${RNADIR}/basic/Table_2foldup_PU1mutvsPU1.txt", sep = "\t", col.names=FALSE,
row.names=FALSE, quote=FALSE)
subdata <- subset(list,(logFC < -1 & logCPM > 1 & FDR < 0.05), select=GeneSymbol)
write.table(as.character(subdata\$GeneSymbol), file =
"${RNADIR}/basic/Table_2folddown_PU1mutvsPU1.txt", sep = "\t", col.names=FALSE,
row.names=FALSE, quote=FALSE)
EOF
chmod 750 "/loctmp/R.extract.P.${_DATE}.R"
R < /loctmp/R.extract.P.${_DATE}.R --no-save
rm /loctmp/R.extract.P.${_DATE}.R

# Comparison with blood cell CAGE data
# PU1 expression across FANTOM samples
makeBoxPlotFromCAGE.pl
/misc/data/analysis/generalStuff/annotation/FANTOM/HemaGeneExpressionGC19.txt SPI1
${FIGURESDIR}/edgeR
makeBeanPlotFromCAGE.pl
/misc/data/analysis/generalStuff/annotation/FANTOM/HemaGeneExpressionGC19.txt
${RNADIR}/basic/Table_3foldup_PU1mutvsPU1.txt bloodExpression.3foldup_PU1mutvsPU1
${FIGURESDIR}/edgeR -gene SPI1


# basic analysis of ATAC & ChIPseq data

# CNV normalization of tagDirs
# normalize & reduce tagDir for better comparability with other data sets
declare -a oCHIPDIRS=("${TAGDIR}/CTV1_PU1_Flag_R1" "${TAGDIR}/CTV1_PU1_Flag_R2"
"${TAGDIR}/CTV1_PU1_Flag_R3" "${TAGDIR}/CTV1_PU1_Flag_6ug" )
declare -a oINPUTDIRS=("${TAGDIR}/CTV1_PU1mut_Flag_R1" "${TAGDIR}/CTV1_PU1mut_Flag_R2"
"${TAGDIR}/CTV1_PU1mut_Flag_R3")

COUNT=0
for SAMPLE in ${oCHIPDIRS[@]}; do
INPUT="${oINPUTDIRS[${COUNT}]}"
normalizeTagDirByCopyNumber.pl ${SAMPLE} -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl ${INPUT} -cnv "${CNVFILECTV1}" -remove
COUNT=$((COUNT+=1))
done

cd ${TAGDIR}
makeTagDirectory CTV1_PU1_Flag_CNVnormRefChr_merged -d CTV1_PU1_Flag_R1_CNVnormRefChr
CTV1_PU1_Flag_R2_CNVnormRefChr CTV1_PU1_Flag_R3_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_PU1mut_Flag_CNVnormRefChr_merged -d CTV1_PU1mut_Flag_R1_CNVnormRefChr
CTV1_PU1mut_Flag_R2_CNVnormRefChr CTV1_PU1mut_Flag_R3_CNVnormRefChr -genome hg19 -checkGC

# normalization of H3K27ac and ATAC tagDirs
cd ${TAGDIR}
normalizeTagDirByCopyNumber.pl CTV1_PU1_H3K27ac_R1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1_H3K27ac_R2 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1_H3K27ac_R3 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1mut_H3K27ac_R1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1mut_H3K27ac_R2 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1mut_H3K27ac_R3 -cnv "${CNVFILECTV1}" -remove

# ETS1 ChIPseq
cd ${TAGDIR}
normalizeTagDirByCopyNumber.pl CTV1_DSG_ETS1_R1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_DSG_ETS1_R2 -cnv "${CNVFILECTV1}" -remove
cd ${CHIPINDIR}
normalizeTagDirByCopyNumber.pl CTV1_DSG_chrInput -cnv "${CNVFILECTV1}" -remove
```

```
cd ${ATACDIR}
normalizeTagDirByCopyNumber.pl CTV1_PU1_R1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1_R2 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1_R3 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1mut_R1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1mut_R2 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1mut_R3 -cnv "${CNVFILECTV1}" -remove

cd ${TAGDIR}
makeTagDirectory CTV1_PU1_H3K27ac_CNVnormRefChr_merged -d CTV1_PU1_H3K27ac_R1_CNVnormRefChr
CTV1_PU1_H3K27ac_R2_CNVnormRefChr CTV1_PU1_H3K27ac_R3_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged -d
CTV1_PU1mut_H3K27ac_R1_CNVnormRefChr CTV1_PU1mut_H3K27ac_R2_CNVnormRefChr
CTV1_PU1mut_H3K27ac_R3_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_DSG_ETS1_CNVnormRefChr_merged -d CTV1_DSG_ETS1_R1_CNVnormRefChr
CTV1_DSG_ETS1_R2_CNVnormRefChr -genome hg19 -checkGC

cd ${ATACDIR}
makeTagDirectory CTV1_PU1_CNVnormRefChr_merged -d CTV1_PU1_R1_CNVnormRefChr
CTV1_PU1_R2_CNVnormRefChr CTV1_PU1_R3_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_PU1mut_CNVnormRefChr_merged -d CTV1_PU1mut_R1_CNVnormRefChr
CTV1_PU1mut_R2_CNVnormRefChr CTV1_PU1mut_R3_CNVnormRefChr -genome hg19 -checkGC

cd ${TAGDIR}
findPeaks CTV1_PU1_Flag_CNVnormRefChr_merged -i CTV1_PU1mut_CNVnormRefChr_Flag_merged -style
factor -fdr 0.00001 -o ${PEAKDIR}/PU1_Flag_merged.factor.fdr05.peaks.txt
findPeaks CTV1_DSG_ETS1_CNVnormRefChr_merged -i ${CHIPINDIR}/CTV1_DSG_chrInput_CNVnormRefChr
-style factor -fdr 0.00001 -o ${PEAKDIR}/ETS1_Flag_merged.factor.fdr05.peaks.txt

# relevant peaks further filtered for peaks with at least 15 tags
myFilterFile.pl ${PEAKDIR}/PU1_Flag_merged.factor.fdr05.peaks.txt -column 6 -lowerlimit 15 >
${PEAKDIR}/PU1_Flag_merged.factor.fdr05.ntag15.peaks.txt
myFilterFile.pl ${PEAKDIR}/ETS1_Flag_merged.factor.fdr05.peaks.txt -column 6 -lowerlimit 15
> ${PEAKDIR}/ETS1_Flag_merged.factor.fdr05.ntag15.peaks.txt

# filtering the merged peak sets
pos2bed.pl ${PEAKDIR}/PU1_Flag_merged.factor.fdr05.ntag15.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Filtered out 3887 regions with mappability scores below 0.8, leaving 44629 regions.
pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt > ${PU1CTV1PEAKS}

pos2bed.pl ${PEAKDIR}/ETS1_Flag_merged.factor.fdr05.ntag15.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Filtered out 812 regions with mappability scores below 0.8, leaving 18221 regions.
pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/DSG_ETS1_merged.factor.fdr05.ntag15.filtered.bed

# BigWigs for normalized TagDirs
# generating individual bigwigs
TAGDIRSETS="CTV1_PU1_H3K27ac_R1_CNVnormRefChr CTV1_PU1_H3K27ac_R2_CNVnormRefChr
CTV1_PU1_H3K27ac_R3_CNVnormRefChr \
CTV1_PU1mut_H3K27ac_R1_CNVnormRefChr CTV1_PU1mut_H3K27ac_R2_CNVnormRefChr
CTV1_PU1mut_H3K27ac_R3_CNVnormRefChr \
CTV1_PU1_Flag_R1_CNVnormRefChr CTV1_PU1_Flag_R2_CNVnormRefChr CTV1_PU1_Flag_R3_CNVnormRefChr \
CTV1_PU1mut_Flag_R1_CNVnormRefChr CTV1_PU1mut_Flag_R2_CNVnormRefChr
CTV1_PU1mut_Flag_R3_CNVnormRefChr \
CTV1_DSG_ETS1_R1_CNVnormRefChr CTV1_DSG_ETS1_R2_CNVnormRefChr"

TAGDIRSETS="CTV1_DSG_ETS1_R1_CNVnormRefChr CTV1_DSG_ETS1_R2_CNVnormRefChr"
for SAMPLE in ${TAGDIRSETS[@]}; do
makeUCSCfile ${CHIPTAGDIR}/$SAMPLE -bigWig $CHROMSIZES -o $BIGWIGDIR/$SAMPLE.bigwig
done

ATACDIRSETS="CTV1_PU1_R1_CNVnormRefChr CTV1_PU1_R2_CNVnormRefChr CTV1_PU1_R3_CNVnormRefChr
CTV1_PU1mut_R1_CNVnormRefChr CTV1_PU1mut_R2_CNVnormRefChr CTV1_PU1mut_R3_CNVnormRefChr"

for SAMPLE in ${ATACDIRSETS[@]}; do
makeUCSCfile ${ATACTAGDIR}/$SAMPLE -bigWig $CHROMSIZES -fragLength 65 -o
$BIGWIGATAC/$SAMPLE.bigwig
Done

# average bigwigs from triplicates
myAverageBigWig.pl -bw $BIGWIGATAC/CTV1_PU1_R1_CNVnormRefChr.bigwig \
$BIGWIGATAC/CTV1_PU1_R2_CNVnormRefChr.bigwig \
$BIGWIGATAC/CTV1_PU1_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGATAC/CTV1_aveATAC_PU1_CNVnormRefChr.bigwig
```

```
myAverageBigWig.pl -bw $BIGWIGATAC/CTV1_PU1mut_R1_CNVnormRefChr.bigwig \
$BIGWIGATAC/CTV1_PU1mut_R2_CNVnormRefChr.bigwig \
$BIGWIGATAC/CTV1_PU1mut_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGATAC/CTV1_aveATAC_PU1mut_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_PU1_H3K27ac_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1_H3K27ac_R2_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1_H3K27ac_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_aveH3K27ac_PU1_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_PU1mut_H3K27ac_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1mut_H3K27ac_R2_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1mut_H3K27ac_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_aveH3K27ac_PU1mut_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_PU1_Flag_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1_Flag_R2_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1_Flag_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_aveFlag_PU1_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_PU1mut_Flag_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1mut_Flag_R2_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1mut_Flag_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_aveFlag_PU1mut_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_DSG_ETS1_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_DSG_ETS1_R2_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_aveDSG_ETS1_CNVnormRefChr.bigwig

#merging Peak sets for comparison

#PU.1 wt vs FLAG wt
cd ${PEAKDIR}
mergePeaks -d 100 CTV1_PU1_PU1_merged.factor.fdr05.peaks.txt
${PEAKDIR}/CTV1_PU1_Flag_merged.factor.fdr05.peaks.txt > /loctmp/tmp.1.txt
pos2bed.pl /loctmp/tmp.1.txt > /loctmp/tmp.1.bed
$BEDTOOLS intersect -a /loctmp/tmp.1.bed -b $BLACKLIST_HG19 -v > /loctmp/tmp.2.bed
filter4Mappability.sh -p /loctmp/tmp.2.bed -g hg19 -f 0.8 -s 50
pos2bed.pl /loctmp/tmp.2.mapScoreFiltered.txt > CTV1_PU1.merged.filtered.bed

annotatePeaks.pl CTV1_PU1.merged.filtered.bed hg19 -size 200 -d
${TAGDIR}/CTV1_PU1_PU1_merged ${TAGDIR}/CTV1_PU1_Flag_merged -noann -nogene >
CTV1_PU1.merged.filtered.ann.txt

cd ${PEAKDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.scatter.${_DATE}.R" <<EOF
library(ggplot2)
library(MASS)
library(scales)
data <- read.delim("CTV1_PU1.merged.filtered.ann.txt", header=T)
data.red <- data[,c(8:9)]
colnames(data.red) <- c("PU1","Flag")
attach(data.red)
d <- data.frame(log10 (data.red + 0.1))
lm_eqn = function(d){
m = lm(PU1 ~ Flag, d);
eq <- substitute(italic(r)^2~"="~r2,
list(r2 = format(summary(m)\$r.squared, digits = 3)))
as.character(as.expression(eq)); }
xlabel = expression("Tag count anti-PU.1")
ylabel = expression("Tag count anti-Flag")
p <- ggplot(data.red,aes(x=PU1, y=Flag)) + coord_trans(x="log10",y="log10")
p <- p + theme_bw(base_size = 12, base_family = "Helvetica") +
coord_cartesian(xlim=c(1,300),ylim=c(1,300))
p <- p + scale_y_continuous(trans = 'log10', breaks = c(1,10,100), labels = c(1,10,100))
p <- p + scale_x_continuous(trans = 'log10', breaks = c(1,10,100), labels = c(1,10,100))
p <- p + geom_jitter(size=.25,alpha=0.02,shape=20,fill="blue",color="blue",width=.1,height
=.1)
p <- p + annotate("text", x = 100, y = 1.4, label = lm_eqn(d), size = 4, colour="black",
parse = TRUE)
p <- p + annotation_logticks(base = 10, short = unit(0.05, "cm"), mid = unit(0.10, "cm"),
long = unit(0.15, "cm"))
p <- p + labs(x = xlabel, y = ylabel)
pdf(file="${FIGURESDIR}/Scatter.peaksPU1.PU1vsFlag.comb.pdf", height=3, width=3)
plot(p)
dev.off()
EOF
chmod 750 "/loctmp/R.scatter.${_DATE}.R"
R < /loctmp/R.scatter.${_DATE}.R --no-save
rm /loctmp/R.scatter.${_DATE}.R
```

```
# BASIC ANALYSIS OF PU1 BINDING IN CTV1 CELLS

# subdividing peaks based on ATAC
# annotate with ATAC data
annotatePeaks.pl ${PU1CTV1PEAKS} hg19 -size 300 -d ${ATACTAGDIR}/CTV1_PU1_R1_CNVnormRefChr
${ATACTAGDIR}/CTV1_PU1_R2_CNVnormRefChr ${ATACTAGDIR}/CTV1_PU1_R3_CNVnormRefChr \
${ATACTAGDIR}/CTV1_PU1mut_R1_CNVnormRefChr ${ATACTAGDIR}/CTV1_PU1mut_R2_CNVnormRefChr
${ATACTAGDIR}/CTV1_PU1mut_R3_CNVnormRefChr -nogene -noann -cpu 6 >
${ANALYSISDIR}/PU1_Flag_merged.ATACann.txt

tail -n +2 ${ANALYSISDIR}/PU1_Flag_merged.ATACann.txt |cut -f1-6,8-13 > ${TMPDIR}/tmp.1.txt
echo
$'ID\tChr\tStart\tEnd\tStrand\tno\tATAC_PU1_1\tATAC_PU1_2\tATAC_PU1_3\tATAC_PU1mut_1\tATAC_PU1
mut_2\tATAC_PU1mut_3' | cat - ${TMPDIR}/tmp.1.txt >
${ANALYSISDIR}/PU1_Flag_merged.ATACann.table.txt

# explorative: generate several Kmeans cluster solutions
rm ${TMPDIR}/PU1_Flag_merged*.*
_DATE=$(date +%s)
for ((i=9;i<=16;i++));do
cat >"${TMPDIR}/R.kmeans.${_DATE}.R" <<EOF
gc()
data <- read.delim("${ANALYSISDIR}/PU1_Flag_merged.ATACann.table.txt", row.names="ID")
d <- data.matrix(data[,6:11])
#cluster with Kmeans
fit <- kmeans(d, $i,iter.max=500)
clusd <- data.frame(data, fit\$cluster)
sorted <- clusd[order(-clusd\$fit.cluster), ]
pos <- sorted[,c(1:4,12)]
write.table(pos,file="${TMPDIR}/PU1_Flag_merged.ATACann.kmeans.$i.txt",sep="\t",
col.names=FALSE, quote=FALSE)
EOF
chmod 750 "${TMPDIR}/R.kmeans.${_DATE}.R"
R < ${TMPDIR}/R.kmeans.${_DATE}.R --no-save
rm ${TMPDIR}/R.kmeans.${_DATE}.R
done

# ghist plots including relevant data sets
cd ${ANALYSISDIR}
TAGDIRSETS="${CHIPTAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr_merged
${CHIPTAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged
${ATACTAGDIR}/CTV1_PU1mut_CNVnormRefChr_merged ${ATACTAGDIR}/CTV1_PU1_CNVnormRefChr_merged
${CHIPTAGDIR}/CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged
${CHIPTAGDIR}/CTV1_PU1_H3K27ac_CNVnormRefChr_merged"

_DATE=$(date +%s)
for KMEANS in ${TMPDIR}/PU1_Flag_merged.ATACann.kmeans.*.txt
do
NAMEBASE=${KMEANS##*/}
NAME=${NAMEBASE%.*}
cat >"${TMPDIR}/ghist.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
annotatePeaks.pl ${KMEANS} hg19 -size 1000 -hist 25 -ghist -d ${TAGDIRSETS} >
"${TMPDIR}/${NAME}.ghist.txt"
plotGHIST.sh -f "${TMPDIR}/${NAME}.ghist.txt" -h 2212 -w 1000 -d "${FIGURESDIR}" -n ${NAME}
-m 10 -c blue
EOF
chmod 750 "${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
echo "plotting ghist for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#---------------------------------------------
for KMEANS in ${TMPDIR}/PU1_Flag_merged.ATACann.kmeans.*.txt ; do
NAMEBASE=${KMEANS##*/}
NAME=${NAMEBASE%.*}
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
```

```
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------

mkdir -p ${FIGURESDIR}/explorativeKmeans/R
mv ${FIGURESDIR}/PU1*.* ${FIGURESDIR}/explorativeKmeans
mv ${FIGURESDIR}/R*.* ${FIGURESDIR}/explorativeKmeans/R

# solution with 14 clusters looks reasonable
mv ${TMPDIR}/PU1_Flag_merged.ATACann.kmeans.14.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.txt
# data analysis based on Kmeans 14 cluster solution
# generating histograms for available data sets across all 14 clusters
declare -a SETS=("${CHIPTAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr_merged"
"${CHIPTAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged"
"${ATACTAGDIR}/CTV1_PU1mut_CNVnormRefChr_merged"
"${ATACTAGDIR}/CTV1_PU1_CNVnormRefChr_merged"
"${CHIPTAGDIR}/CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged"
"${CHIPTAGDIR}/CTV1_PU1_H3K27ac_CNVnormRefChr_merged"
"${CHIPTAGDIR}/CTV1_PU1delA_Flag_CNVnormRefChr_merged"
"${CHIPTAGDIR}/CTV1_PU1delAQP_Flag_CNVnormRefChr_merged"
"${CHIPTAGDIR}/CTV1_PU1delQ_Flag_CNVnormRefChr_merged"
"${CHIPTAGDIR}/CTV1_PU1delP_Flag_CNVnormRefChr_merged")

for ((i=1;i<=14;i++));do
awk -v "key=$i" '$6 == key' ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.txt >
"${TMPDIR}/tmp.14.${i}.txt"
done

# copy clusters
for ((i=1;i<=14;i++));do
cp ${TMPDIR}/tmp.14.${i}.txt ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.${i}.txt
done

# determine cluster sizes:
for ((i=1;i<=14;i++));do
echo "cluster.$i size: $(wc -l <${TMPDIR}/tmp.14.${i}.txt)"
done
# cluster.1 size: 4515
# cluster.2 size: 3712
# cluster.3 size: 5091
# cluster.4 size: 3293
# cluster.5 size: 5023
# cluster.6 size: 3249
# cluster.7 size: 1797
# cluster.8 size: 3210
# cluster.9 size: 2276
# cluster.10 size: 4769
# cluster.11 size: 510
# cluster.12 size: 1512
# cluster.13 size: 1692
# cluster.14 size: 3980

# order clusters based on ATAC signal
KMEANSORDERORI="3 5 10 1 14 4 8 12 9 2 6 7 13 11"
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.3.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.1.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.5.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.2.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.10.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.3.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.1.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.4.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.14.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.5.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.4.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.6.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.8.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.7.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.12.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.8.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.9.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.9.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.2.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.10.txt
```

```
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.6.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.11.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.7.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.12.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.13.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.13.txt
cp ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.11.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.14.txt

cat ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.1.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.2.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.3.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.4.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.5.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.6.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.7.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.8.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.9.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.10.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.11.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.12.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.13.txt
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.14.txt>
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt
wc -l ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt #44629 peaks

pos2bed.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed


# average remodeling index for clusters
rm ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.remodelingIndices.txt
touch ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.remodelingIndices.txt
for ((i=1;i<=14;i++));do
annotatePeaks.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt hg19
-size 300 -d "${ATACTAGDIR}/CTV1_PU1mut_CNVnormRefChr_merged"
"${ATACTAGDIR}/CTV1_PU1_CNVnormRefChr_merged" -nogene -noann >
"${TMPDIR}/tmp.${i}.ATACsignal.txt"
awk -v "key=$i" '{ mutPU += $8 ; PU += $9} END { print "index for cluster"key": "
PU/mutPU }' "${TMPDIR}/tmp.$i.ATACsignal.txt" >>
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.remodelingIndices.txt
Done

# ghist plots ChIP & ATAC data
# combined plots
declare -a SETS=("${TAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged" \
"${ATACDIR}/CTV1_PU1mut_CNVnormRefChr_merged" "${ATACDIR}/CTV1_PU1_CNVnormRefChr_merged" \
"${TAGDIR}/CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_H3K27ac_CNVnormRefChr_merged" \
"${TAGDIR}/CTV1_PU1delA_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delAQP_Flag_CNVnormRefChr_merged" \
"${TAGDIR}/CTV1_PU1delQ_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delP_Flag_CNVnormRefChr_merged" \
"${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr" "${TAGDIR}/CTV1_DSG_ETS1_CNVnormRefChr_merged")

declare -a NAMES=(PU1mut PU1 ATACmut ATACPU1 H3K27acmut H3K27acPU1 delA delAQP delQ delP
PU1less ETS1)
declare -a SETS=("${TAGDIR}/CTV1_DSG_ETS1_CNVnormRefChr_merged")
declare -a NAMES=(ETS1)

COUNT=0
for SET in ${SETS[@]}; do
NAME="${NAMES[${COUNT}]}"
_DATE=$(date +%s)
cat >"/loctmp/ghist.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd /loctmp
annotatePeaks.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt hg19
-size 1000 -hist 25 -ghist -d ${SET} >
"${GHISTDIR}/kmeans.14.sorted.cleaned.${NAME}.ghist.txt"
EOF
COUNT=$((COUNT+=1))
```

```
chmod 750 "/loctmp/ghist.${NAME}.${_DATE}.sh"
echo "annotating ghist for ${NAME}"
screen -dm -S ${NAME} bash -c "bash /loctmp/ghist.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#--------------------------------------------
for NAME in ${NAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

# Histogram for PU.1, ATAC und H3K27ac
SCALES="25 25 15 15 15 15"
COLORS="blue blue navy navy dodgerblue3 dodgerblue3"
GHISTSAMPLES="${GHISTDIR}/kmeans.14.sorted.cleaned.${NAMES[0]}.ghist.txt"
for ((i=1;i<6;i++));do
GHISTSAMPLES="${GHISTSAMPLES} ${GHISTDIR}/kmeans.14.sorted.cleaned.${NAMES[$i]}.ghist.txt"
done
plotGHIST.sh -f "${GHISTSAMPLES}" -h 22315 -w 2000 -d "${FIGURESDIR}" -n PU.ATAC.H3K27ac -m
"${SCALES}" -c "${COLORS}"

# Histogram for ETS1, PU.1 & ATAC
declare -a NAMES=(PU1 ETS1 ATACmut)
SCALES="25 15 15"
COLORS="blue slateblue4 navy"
GHISTSAMPLES="${GHISTDIR}/kmeans.14.sorted.cleaned.${NAMES[0]}.ghist.txt"
for ((i=1;i<3;i++));do
GHISTSAMPLES="${GHISTSAMPLES} ${GHISTDIR}/kmeans.14.sorted.cleaned.${NAMES[$i]}.ghist.txt"
done
plotGHIST.sh -f "${GHISTSAMPLES}" -h 22315 -w 2000 -d "${FIGURESDIR}" -n PU.ETS.ATACmut -m
"${SCALES}" -c "${COLORS}"
plotGHIST.sh -f "${GHISTDIR}/kmeans.14.sorted.cleaned.ETS1.ghist.txt" -h 22315 -w 2000 -d
"${FIGURESDIR}" -n ETS1 -m 15 -c slateblue4

# "ghistplot" colorbar
_DATE=$(date +%s)
cat >"/loctmp/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
#library(RColorBrewer)
data <- read.delim("${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt",
header=F)
d <- data.matrix(data[,6])
mycol <-
c("darkorange3","deepskyblue2","gold","firebrick3","goldenrod2","dodgerblue1","dodgerblue3","f
irebrick","darkmagenta","darkorange1","blue4","deeppink","blue2","firebrick1")
#mycol <-
c("firebrick1","blue2","deeppink","blue4","darkorange1","darkmagenta","firebrick","dodgerblue3
","dodgerblue1","goldenrod2","firebrick3","gold","deepskyblue2","darkorange3")
png(filename="kmeans.14.ghistColorBar.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(d),col=mycol,zlim=range(c(0,14)))
dev.off()
EOF
chmod 750 "/loctmp/R.image.P.${_DATE}.R"
R < /loctmp/R.image.P.${_DATE}.R --no-save
rm /loctmp/R.image.P.${_DATE}.R


# annotation of peaks to genes
# find genes associated with 14 original clusters and get the corresponding gene expression
data
for ((i=1;i<=14;i++));do
getAllGeneEnhancerAssociations.pl
"${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt" \
"${HG19TRANS}" "PU1_Flag.kmeans.sorted.${i}" "${ANALYSISDIR}/peakAssociatedGenes" -gtex
"${HG19GTEX}" #-expr "${EXPRESSION}"/expressedGenesList.txt
tail -n +2 "${EXPRESSION}"/ReadCountTable.norm.corr.logcpm.txt > /loctmp/tmp.data.txt
tail -n +2
"${ANALYSISDIR}/peakAssociatedGenes/PU1_Flag.kmeans.sorted.${i}.total.GeneList.txt" |
sort -k1,1 > /loctmp/tmp.genes.txt
sed -e 's/[$]/\t/g' /loctmp/tmp.data.txt > /loctmp/tmp.data2.txt
cut -f2,5-12 /loctmp/tmp.data2.txt | sort -uk1,1 - > /loctmp/tmp.data3.txt
join -1 1 -2 1 -t $'\t' /loctmp/tmp.data3.txt /loctmp/tmp.genes.txt > /loctmp/tmp.pair.txt
echo
$'Gene\tCTV1_woPU1_rep2\tCTV1_woPU1_rep1_B\tCTV1_PU1_rep2\tCTV1_PU1_rep1_B\tCTV1_PU1mut_re
p2\tCTV1_PU1mut_rep1_B\tCTV1_PU1Rep\tCTV1_PU1mutRep' | cat - /loctmp/tmp.pair.txt >
"${EXPRESSION}"/PU1_Flag.kmeans.sorted.${i}.total.GeneList.expression.txt
wc -l "${EXPRESSION}"/PU1_Flag.kmeans.sorted.${i}.total.GeneList.expression.txt
done

# plotting mRNA expression data across clusters
for ((i=1;i<=14;i++));do
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.beanplot.P.${_DATE}.R" <<EOF
library(beanplot)
# importing data file
data <-
read.table("${EXPRESSION}/PU1_Flag.kmeans.sorted.${i}.total.GeneList.expression.txt",
header=T, row.names="Gene", sep="\t")
# combining data
a <- (data[-1,1] + data[-1,2]) / 2
b <- (data[-1,3] + data[-1,4]) / 2
c <- (data[-1,5] + data[-1,6]) / 2
d <- data[-1,7]
e <- data[-1,8]
# statistical testing
wilcb <- wilcox.test(c,b, paired=TRUE,alternative="less")
if (wilcb\$p.value < 0.001) {
wcb <- "***"
} else if (wilcb\$p.value < 0.01) {
wcb <- "**"
} else if (wilcb\$p.value < 0.05) {
wcb <- "*"
} else {
wcb <- "ns"
}
wilde <- wilcox.test(e,d, paired=TRUE,alternative="less")
if (wilde\$p.value < 0.001) {
wde <- "***"
} else if (wilde\$p.value < 0.01) {
wde <- "**"
} else if (wilde\$p.value < 0.05) {
wde <- "*"
} else {
wde <- "ns"
}
d <- data.frame(cbind(c,b,e,d))
# defining colors
beancol <- list("azure1","azure4","azure1","azure4")
bcol <- c("cyan","dodgerblue","cyan","dodgerblue")
boxcol <- adjustcolor(bcol, alpha.f = 0.5)
# plotting the beans
pdf(file="${FIGURESDIR}/Kmeans${i}_mRNAexpression.pdf", height=4, width=2)
par(mar=c(6,3,3,1))
beanplot(d,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = "gray"
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = 1, maxstripline =
0.8,ylim=c(-5,15),lwd=0.4)
# adding box plot on top
par(mar=c(6,3,3,1),new=TRUE)
boxplot(d,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.8, staplewex = 0.2, ylim=c(-5,15),lwd=0.6)
# axis and legends
title(main="PU.1 peaks in cluster ${i}", cex.main=0.8)
axis(1,padj=0.4,family="Helvetica",cex.axis=1,at=1:4,labels=c("mutPU1","PU1","mutPU1rep","PU1r
ep"),las=2)
axis(2,padj=0.4,family="Helvetica",cex.axis=1,las=1,mgp=c(2,.6,0))
mtext("norm. gene expression (log2)",family="Helvetica",side=2,line=2,cex=.8,padj=0.4)
# adding significance levels
text(1.5,15,wcb)
segments(1,14,2,14)
text(3.5,15,wde)
segments(3,14,4,14)
dev.off()
EOF
chmod 750 "/loctmp/R.beanplot.P.${_DATE}.R"
R < /loctmp/R.beanplot.P.${_DATE}.R --no-save
```

```
rm /loctmp/R.beanplot.P.${_DATE}.R
done


# motif analyses accros all clusters

# perform motif scan for all clusters
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >"${TMPDIR}/motifs.cl${i}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findMotifsGenome.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt hg19
${MOTIFDIR}/PU1_Kmeans14.${i} -size 200 -len 7,8,9,10,11,12,13,14 -p 2 -h
compareMotifs.pl ${MOTIFDIR}/PU1_Kmeans14.${i}/homerMotifs.all.motifs
${MOTIFDIR}/PU1_Kmeans14.${i}/final -reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info
1.5 -cpu 2
EOF
chmod 750 "${TMPDIR}/motifs.cl${i}.${_DATE}.sh"
echo "finding motifs for cluster ${i}"
screen -dm -S motif${i} bash -c "bash ${TMPDIR}/motifs.cl${i}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#---------------------------------------------
for ((i=1;i<=14;i++));do
while [ true ]; do # Endless loop.
pid=`screen -S motif${i} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------

# create a file containing all known motif enrichments
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >>"${TMPDIR}/knownMotifs.kmeans.${_DATE}.txt" <<EOF
${MOTIFDIR}/PU1_Kmeans14.${i}/knownResults.txt
EOF
done
mkdir "${MOTIFDIR}/PU1_Kmeans14.summary"
mySummarizeMotifResults.pl "${TMPDIR}/knownMotifs.kmeans.${_DATE}.txt" \
-outDir "${MOTIFDIR}/PU1_Kmeans14.summary" -hc -minp 0.0000000001 -minr 2.5 -limit 3

# need to reduce motifs
# creating a known motif file for annotation across clusters
cd ${KNOWNTFDIR}
cat /misc/data/analysis/project_PU1/PU1long.motif runx1.motif ap1.motif ctcf.motif
gata3.motif \
gfy.motif rfx2.motif nfy.motif ets-runx.motif pu1-irf.motif sp1.motif >
${MOTIFDIR}/knownMotifsInClusters.motifs

# redo knownMotif scan with known Motifs in Clusters
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >"${TMPDIR}/motifs.cl${i}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findMotifsGenome.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt hg19r
${MOTIFDIR}/PU1_Kmeans14.sorted.${i}.known -size 200 -mknown
/misc/data/analysis/project_PU1/CTV1/motifs/knownMotifsInClusters.motifs -nomotif -p 4 -h
EOF
chmod 750 "${TMPDIR}/motifs.cl${i}.${_DATE}.sh"
```

```
echo "finding motifs for cluster ${i}"
screen -dm -S motif${i} bash -c "bash ${TMPDIR}/motifs.cl${i}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#--------------------------------------------
for ((i=1;i<=14;i++));do
while [ true ]; do # Endless loop.
pid=`screen -S motif${i} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample motif${i}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

# create a file containing all known motif enrichments
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >>"${TMPDIR}/knownMotifs.kmeans.${_DATE}.txt" <<EOF
${MOTIFDIR}/PU1_Kmeans14.sorted.${i}.known/knownResults.txt
EOF
done
mkdir "${MOTIFDIR}/PU1_Kmeans14.known.summary"
mySummarizeMotifResults.pl "${TMPDIR}/knownMotifs.kmeans.${_DATE}.txt" \
-outDir "${MOTIFDIR}/PU1_Kmeans14.known.summary" -minp 0.001 -minr 1.25 -limit 3

# same analysis with de novo motifs
# create a combined motif file containing only top motif out of all Homer motifs in all
Celltypes
COMBINEDMOTIFS="${MOTIFDIR}/combinedHomerMotifs.all.motifs"
rm ${COMBINEDMOTIFS}
touch ${COMBINEDMOTIFS}
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat ${COMBINEDMOTIFS} ${MOTIFDIR}/PU1_Kmeans14.${i}/homerMotifs.all.motifs >
"${TMPDIR}/homerMotifs.names.${_DATE}.txt"
mv "${TMPDIR}/homerMotifs.names.${_DATE}.txt" ${COMBINEDMOTIFS}
done
compareMotifs.pl ${COMBINEDMOTIFS} ${MOTIFDIR}/combinedHomerMotifs/final -reduceThresh .75
-matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 12
rm ${FILTEREDMOTIFS}
FILTEREDMOTIFS="${MOTIFDIR}/combinedFilteredHomerMotifs.all.motifs"
touch ${FILTEREDMOTIFS}
for ((i=1;i<=381;i++)); do
MOTIFCOR=$(awk -F'[()]' 'NR==1 {print $2}'
"${MOTIFDIR}/combinedHomerMotifs/final/homerResults/motif${i}.motif")
if [[ ${MOTIFCOR} > 0.85 ]]; then
cat ${FILTEREDMOTIFS} ${MOTIFDIR}/combinedHomerMotifs/final/homerResults/motif${i}.motif
> "${TMPDIR}/homerMotifs.comb.${_DATE}.txt"
mv "${TMPDIR}/homerMotifs.comb.${_DATE}.txt" ${FILTEREDMOTIFS}
fi
done

_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >"${TMPDIR}/motifs.cl${i}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findMotifsGenome.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt hg19r
${MOTIFDIR}/PU1_Kmeans14.sorted.${i}.redhomer -size 200 -mknown
"${MOTIFDIR}/combinedFilteredHomerMotifs.red.motifs" -nomotif -p 4 -h
EOF
chmod 750 "${TMPDIR}/motifs.cl${i}.${_DATE}.sh"
echo "finding motifs for cluster ${i}"
screen -dm -S motif${i} bash -c "bash ${TMPDIR}/motifs.cl${i}.${_DATE}.sh"
done
```

```
# loop to check when screen sessions are done
#-------------------------------------------
for ((i=1;i<=14;i++));do
while [ true ]; do # Endless loop.
pid=`screen -S motif${i} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample motif${i}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#-------------------------------------------

# create a file containing all known motif enrichments
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >>"${TMPDIR}/knownMotifs.kmeans.${_DATE}.txt" <<EOF
${MOTIFDIR}/PU1_Kmeans14.sorted.${i}.redhomer/knownResults.txt
EOF
done
mkdir "${MOTIFDIR}/PU1_Kmeans14.redhomer.summary"
mySummarizeMotifResults.pl "${TMPDIR}/knownMotifs.kmeans.${_DATE}.txt" \
-outDir "${MOTIFDIR}/PU1_Kmeans14.redhomer.summary" -minp 0.001 -minr 1.25 -limit 2

# ballonplot for motif enrichment
_DATE=$(date +%s)
cat >"${TMPDIR}/R.ballon.${_DATE}.R" <<EOF
library(ggplot2)
library(reshape2)
library(ggpubr)
r <- read.table("${MOTIFDIR}/PU1_Kmeans14.redhomer.summary/cleanedRatioTable.txt", header=T,
sep="\t")
mr <- melt(r)
q <- read.table("${MOTIFDIR}/PU1_Kmeans14.redhomer.summary/cleanedqValueTable.txt",
header=T, sep="\t")
mq <- melt(q)
s <- read.table("${MOTIFDIR}/PU1_Kmeans14.redhomer.summary/short.motif.names.txt", header=T,
sep="\t")
rq <- merge(mr,mq,by=c("Motif.Name","variable"))
table <- merge(s,rq,by="Motif.Name")
table\$logq <- log10(table\$value.y+0.00005)
table\$enr <- (table\$value.x+0.01)
p <- ggballoonplot(table, x = "variable", y = "Short.Name", size = "enr", color="black",
fill = "logq", size.range = c(0, 10), ggtheme = theme_bw())
p <- p + scale_fill_gradient2(low = "mediumblue", mid = "gray90", high = "gray90", midpoint
= -1)
p <- p + scale_color_gradient2(low = "mediumblue", mid = "gray90", high = "gray90",
midpoint = -1)
pdf(file="${FIGURESDIR}/ballonplot.motifenrichment.Kmeans14.pdf", height=3.5, width=5)
plot(p)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.ballon.${_DATE}.R"
R < "${TMPDIR}/R.ballon.${_DATE}.R" --no-save
rm "${TMPDIR}/R.ballon.${_DATE}.R"

# motifscore distribution across clusters
for ((i=1;i<=14;i++));do
_DATE=$(date +%s)
cat >"${TMPDIR}/motifs.${i}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
annotatePeaks.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt hg19 -size
200 -m /misc/data/analysis/project_PU1/PU1long.motif -mscore -nogene -noann >
${TMPDIR}/tmp.14.${i}.txt
EOF
chmod 750 "${TMPDIR}/motifs.${i}.${_DATE}.sh"
echo "annotating PU1 motif for cluster ${i}"
screen -dm -S motif${i} bash -c "bash ${TMPDIR}/motifs.${i}.${_DATE}.sh"
done
```

```
# loop to check when screen sessions are done
#--------------------------------------------
for ((i=1;i<=14;i++));do
while [ true ]; do # Endless loop.
pid=`screen -S motif${i} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample motif${i}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

# combined bean- and box-plot
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.14.1.txt", header=T, sep="\t")
data2 <- read.table("${TMPDIR}/tmp.14.2.txt", header=T, sep="\t")
data3 <- read.table("${TMPDIR}/tmp.14.3.txt", header=T, sep="\t")
data4 <- read.table("${TMPDIR}/tmp.14.4.txt", header=T, sep="\t")
data5 <- read.table("${TMPDIR}/tmp.14.5.txt", header=T, sep="\t")
data6 <- read.table("${TMPDIR}/tmp.14.6.txt", header=T, sep="\t")
data7 <- read.table("${TMPDIR}/tmp.14.7.txt", header=T, sep="\t")
data8 <- read.table("${TMPDIR}/tmp.14.8.txt", header=T, sep="\t")
data9 <- read.table("${TMPDIR}/tmp.14.9.txt", header=T, sep="\t")
data10 <- read.table("${TMPDIR}/tmp.14.10.txt", header=T, sep="\t")
data11 <- read.table("${TMPDIR}/tmp.14.11.txt", header=T, sep="\t")
data12 <- read.table("${TMPDIR}/tmp.14.12.txt", header=T, sep="\t")
data13 <- read.table("${TMPDIR}/tmp.14.13.txt", header=T, sep="\t")
data14 <- read.table("${TMPDIR}/tmp.14.14.txt", header=T, sep="\t")
# combining the data columns (bit complicated, because each column has different length)
a <- data1[-1,10]
b <- data2[-1,10]
c <- data3[-1,10]
d <- data4[-1,10]
e <- data5[-1,10]
f <- data6[-1,10]
g <- data7[-1,10]
h <- data8[-1,10]
i <- data9[-1,10]
j <- data10[-1,10]
k <- data11[-1,10]
l <- data12[-1,10]
m <- data13[-1,10]
n <- data14[-1,10]
z <- c("a", "b", "c", "d", "e", "f", "g", "h", "i", "j", "k", "l", "m", "n")
x <- lapply(z, get, envir=environment())
names(x) <- z
# determining the length of each column
anum <- nrow(as.matrix(a))
bnum <- nrow(as.matrix(b))
cnum <- nrow(as.matrix(c))
dnum <- nrow(as.matrix(d))
enum <- nrow(as.matrix(e))
fnum <- nrow(as.matrix(f))
gnum <- nrow(as.matrix(g))
hnum <- nrow(as.matrix(h))
inum <- nrow(as.matrix(i))
jnum <- nrow(as.matrix(j))
knum <- nrow(as.matrix(k))
lnum <- nrow(as.matrix(l))
mnum <- nrow(as.matrix(m))
nnum <- nrow(as.matrix(n))
# define labels
laba <- paste("1 (",anum,")",sep="")
labb <- paste("2 (",bnum,")",sep="")
labc <- paste("3 (",cnum,")",sep="")
labd <- paste("4 (",dnum,")",sep="")
labe <- paste("5 (",enum,")",sep="")
labf <- paste("6 (",fnum,")",sep="")
labg <- paste("7 (",gnum,")",sep="")
labh <- paste("8 (",hnum,")",sep="")
labi <- paste("9 (",inum,")",sep="")
```

```
labj <- paste("10 (",jnum,")",sep="")
labk <- paste("11 (",knum,")",sep="")
labl <- paste("12 (",lnum,")",sep="")
labm <- paste("13 (",mnum,")",sep="")
labn <- paste("14 (",nnum,")",sep="")

# defining colors
beancol <-
list("gold","goldenrod2","darkorange1","darkorange3","firebrick1","firebrick3","firebrick","de
eppink","darkmagenta","deepskyblue2","dodgerblue1","dodgerblue3","blue2","blue4")
boxcol <-
c("gray25","gray30","gray35","gray40","gray45","gray50","gray55","gray60","gray65","gray70","g
ray75","gray80","gray85","gray90")
pdf(file="${FIGURESDIR}/motifScores.cl1-14.pdf", height=3, width=5)
par(mar=c(4.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(3,12),lwd=0.5)
# adding box plot on top
par(mar=c(4.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(3,12),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:14, labels=c(laba, labb, labc, labd,
labe, labf, labg, labh, labi, labj, labk, labl, labm, labn),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext(" Motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
mtext("Cluster",family="Helvetica",ps=12,side=1,line=2,cex=1.2,padj=2.0)
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.motifbean.${_DATE}.R"
R < ${TMPDIR}/R.motifbean.${_DATE}.R --no-save
rm ${TMPDIR}/R.motifbean.${_DATE}.R


# analysis of motif co-occurence
# redo known motif analysis with PU.1 motif masked
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >"${TMPDIR}/motifs.cl${i}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findMotifsGenome.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt hg19r
${MOTIFDIR}/PU1_Kmeans14.sorted.${i}.known.PU1masked -size 200 -mknown
/misc/software/ngs/homer/v4.9/data/knownTFs/vertebrates/known.motifs -maskMotif ${PU1MOTIF}
-nomotif -p 3 -h
EOF
chmod 750 "${TMPDIR}/motifs.cl${i}.${_DATE}.sh"
echo "finding motifs for cluster ${i}"
screen -dm -S motif${i} bash -c "bash ${TMPDIR}/motifs.cl${i}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#---------------------------------------------
for ((i=1;i<=14;i++));do
while [ true ]; do # Endless loop.
pid=`screen -S motif${i} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample motif${i}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------
```

```
# network of motif co-association for selected samples
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >"${TMPDIR}/motifs.cl${i}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
annotatePeaks.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt hg19 -size
200 -m /misc/software/ngs/homer/v4.9/data/knownTFs/vertebrates/known.motifs -fm ${PU1MOTIF}
-matrixMinDist 4 -nogene -noann -nmotifs -matrix ${TMPDIR}/tmp.cl${i}rm >
${TMPDIR}/tmp.m.14.cl${i}.rm.txt
annotatePeaks.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt hg19 -size
200 -m ${PU1MOTIF} -matrixMinDist 6 -nogene -noann -nmotifs -matrix ${TMPDIR}/.cl${i}pu >
${TMPDIR}/tmp.m.14.cl${i}.pu.txt
myFilterFile.pl
/misc/data/analysis/project_PU1/CTV1/motifs/basic/PU1_Kmeans14.sorted.${i}.known.PU1masked/kno
wnResults.txt -column 5 -upperlimit 0.05001 | cut -f1 > ${TMPDIR}/tmp.motifs.${i}.txt
join -1 1 -2 1 -t $'\t' <(sort -k1,1 ${CUSTOMMOTIFS}/knownMotifsCTV1.classes.txt ) <(sort
-k1,1 ${TMPDIR}/tmp.motifs.${i}.txt) > ${TMPDIR}/tmp.sign.motifs.${i}.txt
(head -n 1 ${TMPDIR}/tmp.m.14.cl${i}.rm.txt && tail -n +2 ${TMPDIR}/tmp.m.14.cl${i}.rm.txt |
sort -k 1,1 )> ${TMPDIR}/tmp.m.14.cl${i}.rm.sorted.txt
cut -f11- ${TMPDIR}/tmp.m.14.cl${i}.rm.sorted.txt | sed -e s/Distance\ From\
Peak\(sequence,strand,conservation\)//g > ${TMPDIR}/tmp.m3.14.cl${i}.txt
myTP.pl ${TMPDIR}/tmp.m3.14.cl${i}.txt > ${TMPDIR}/tmp.m3.14.cl${i}.tp.txt
join -1 1 -2 1 -t $'\t' <(sort -k1,1 ${TMPDIR}/tmp.sign.motifs.${i}.txt ) <(sort -k1,1
${TMPDIR}/tmp.m3.14.cl${i}.tp.txt) > ${TMPDIR}/tmp.m4.14.cl${i}.tp.txt
sort -k2,2 ${TMPDIR}/tmp.m4.14.cl${i}.tp.txt > ${TMPDIR}/tmp.m5.14.cl${i}.tp.txt
redMotifCounts.pl ${TMPDIR}/tmp.m5.14.cl${i}.tp.txt > ${TMPDIR}/tmp.m6.14.cl${i}.tp.txt
myTP.pl ${TMPDIR}/tmp.m6.14.cl${i}.tp.txt >${TMPDIR}/tmp.m6.14.cl${i}.txt
(head -n 1 ${TMPDIR}/tmp.m.14.cl${i}.pu.txt && tail -n +2 ${TMPDIR}/tmp.m.14.cl${i}.pu.txt |
sort -k 1,1 )> ${TMPDIR}/tmp.m.14.cl${i}.pu.sorted.txt
cut -f10- ${TMPDIR}/tmp.m.14.cl${i}.pu.sorted.txt | sed -e s/Distance\ From\
Peak\(sequence,strand,conservation\)//g > ${TMPDIR}/tmp.m3.14.cl${i}.pu.txt
redAutoMotifCounts.pl ${TMPDIR}/tmp.m3.14.cl${i}.pu.txt PU1 >
${TMPDIR}/tmp.m6.14.cl${i}.pu.txt
paste ${TMPDIR}/tmp.m6.14.cl${i}.pu.txt ${TMPDIR}/tmp.m6.14.cl${i}.txt >
${TMPDIR}/tmp.m7.14.cl${i}.txt
getNetworkMotifFiles.pl ${TMPDIR}/tmp.m7.14.cl${i}.txt ${ANALYSISDIR}/cl${i}
EOF
chmod 750 "${TMPDIR}/motifs.cl${i}.${_DATE}.sh"
echo "finding motifs for cluster ${i}"
screen -dm -S motif${i} bash -c "bash ${TMPDIR}/motifs.cl${i}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#----------------------------------------------
for ((i=1;i<=14;i++));do
while [ true ]; do # Endless loop.
pid=`screen -S motif${i} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample motif${i}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#----------------------------------------------

# plotting the networks
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >"${TMPDIR}/network.cl${i}.${_DATE}.R" <<EOF
library("igraph")
nodes <- read.delim("${ANALYSISDIR}/cl${i}.nodes.txt", header=T, as.is=T)
links <- read.delim("${ANALYSISDIR}/cl${i}.edges.txt", header=T, as.is=T)
net <- graph.data.frame(links, nodes, directed=F)
colrs <- c("blue","firebrick1")
V(net)\$color <- colrs[factor(V(net)\$type.label)]
V(net)\$size <- sqrt(V(net)\$fraction)*12
V(net)\$label <- V(net)\$tf.name
E(net)\$width <- E(net)\$weight/2
E(net)\$edge.color <- "black"
l <- layout.star(net)
```

```
pdf(file="${FIGURESDIR}/motif-coenrich.nw.cl${i}.pdf", height=6, width=6)
plot(net, layout=l, vertex.label.family="Helvetica", vertex.frame.color="black")
blob.size <- c(75,50,25,5)
legend(x=1.6,y=1, blob.size, pch=21, col="black", pt.bg="white",
pt.cex=sqrt(blob.size)/0.42, cex=.8, bty="n")
dev.off()
EOF
chmod 750 "${TMPDIR}/network.cl${i}.${_DATE}.R"
R < "${TMPDIR}/network.cl${i}.${_DATE}.R" --no-save
rm "${TMPDIR}/network.cl${i}.${_DATE}.R"
done


# % coassociated PU.1
rm ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.homotypeIndices.txt
touch ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.homotypeIndices.txt
for ((i=1;i<=14;i++));do
awk -v "key=$i" '$2 ~ "PU1_total" { print "PU1total_cluster"key": " $4 }'
"${ANALYSISDIR}/cl${i}.nodes.txt" >>
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.homotypeIndices.txt
done

for ((i=1;i<=14;i++));do
awk -v "key=$i" '$2 ~ "PU1_co" { print "PU1co_cluster"key": " $4 }'
"${ANALYSISDIR}/cl${i}.nodes.txt" >>
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.homotypeIndices.txt
done

# comparison with remodeling index
awk -v OFS='\t' 'BEGIN {x=1};{print "cluster"x,$3,($2/$1*100);x++ }'
paste <(head -n 14 ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.homotypeIndices.txt |
cut -d ' ' -f2) \
<(tail -n 14 ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.homotypeIndices.txt | cut -d '
' -f2) \
<(head -n 14 ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.remodelingIndices.txt | cut -d
' ' -f2) \
>"${TMPDIR}/tmp.table.txt"
awk -v OFS='\t' 'BEGIN {x=1};{print "cluster"x, $3 , ($2/$1*100) ;x++ }'
${TMPDIR}/tmp.table.txt > ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.indexComp.txt

# plotting relationship between remodeling index and % coassociation
_DATE=$(date +%s)
cat >"${TMPDIR}/scatter.${_DATE}.R" <<EOF
library(ggplot2)
library(ggpubr)
setwd("${ANALYSISDIR}")
data <- read.table("PU1_Flag_merged.ATACann.kmeans.14.indexComp.txt", header=F, sep="\t")
colnames(data) <- c("cluster","index","percent")
data\$logindex <- log2(data\$index)
###### starting the plot
p <- ggplot(data, aes(x=logindex,y=percent))
p <- p + geom_smooth(method=lm,color = "gray", fill = "lightgray") + stat_cor(method =
"pearson", size = 2)
p <- p + geom_point(shape=16, size=2,
color=c("gold","goldenrod2","darkorange1","darkorange3","firebrick1","firebrick3","firebrick",
"deeppink","darkmagenta","deepskyblue2","dodgerblue1","dodgerblue3","blue2","blue4"))
p <- p + xlab("remodeling index (log2)") + ylab("% homotypic clusters")
p <- p + theme_light(base_size=8)
pdf(file="${FIGURESDIR}/comparisonRemIndex_percentHomotypic.pdf", height=2, width=2)
plot(p)
dev.off()
EOF
chmod 750 "${TMPDIR}/scatter.${_DATE}.R"
R < "${TMPDIR}/scatter.${_DATE}.R" --no-save
rm "${TMPDIR}/scatter.${_DATE}.R"


# ATAC signal around PU.1 motifs in clusters
# center peaks in clusters on PU.1 motif and annotate ATAC cut sites
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >"${TMPDIR}/center.cl${i}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
```

```
cd ${TMPDIR}
annotatePeaks.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt hg19 -size
200 -center ${PU1MOTIF} >
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.${i}.txt
annotatePeaks.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.${i}.txt
hg19 -size 550 -hist 1 -len 1 -d "${ATACTAGDIR}/CTV1_PU1mut_CNVnormRefChr_merged"
"${ATACTAGDIR}/CTV1_PU1_CNVnormRefChr_merged" >
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.${i}.ATAChist1.txt
EOF
chmod 750 "${TMPDIR}/center.cl${i}.${_DATE}.sh"
echo "centering peaks for cluster ${i}"
screen -dm -S cluster${i} bash -c "bash ${TMPDIR}/center.cl${i}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#--------------------------------------------
for ((i=1;i<=14;i++));do
while [ true ]; do # Endless loop.
pid=`screen -S cluster${i} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample cluster ${i}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

for ((i=1;i<=14;i++));do
echo "cluster.$i size: $(wc -l
<${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.${i}.txt)"
done

# remaining cluster sizes after motif centering:
# cluster.1 size: 5026
# cluster.2 size: 4924
# cluster.3 size: 4611
# cluster.4 size: 4326
# cluster.5 size: 3791
# cluster.6 size: 3129
# cluster.7 size: 3005
# cluster.8 size: 1405
# cluster.9 size: 2054
# cluster.10 size: 3269
# cluster.11 size: 2646
# cluster.12 size: 1457
# cluster.13 size: 1216
# cluster.14 size: 338

# annotation of motifs with sequence and generation of random controls
cd ${PEAKDIR}
$BEDTOOLS intersect -a ${HOMOCLUSTERDIR}/PU.1_long_hg19_refChr.bed -b ${PU1CTV1PEAKS} -u
>${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.bed
homerTools extract ${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.bed ${GENOME_HG19}
>${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seq.txt
join -1 4 -2 1 -t $'\t' <(sort -k4,4 ${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.bed)
<(sort -k1,1 ${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seq.txt) >
/misc/data/tmp/PU.1_long_hg19_refChr.filtered.bed
awk -v OFS='\t' '{print $2,$3,$4,($1 "%" $7),$5,$6,$7 }'
/misc/data/tmp/PU.1_long_hg19_refChr.filtered.bed >
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seq.bed
wc -l ${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seq.bed # 66506 motifs
for w in `cut -f 7 ${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seq.bed`; do echo $w;
done|sort|uniq -c|sort -k1,1nr|sed 's/^ *//g'|tr [:blank:] \\t >
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seq.count.txt

# overlap with all motifs
join -1 2 -2 2 -t $'\t' <(sort -k2,2
${HOMOCLUSTERDIR}/PU.1_long_hg19_refChr.filtered.seq.count.txt) <(sort -k2,2
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seq.count.txt) >
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seqoverlap.count.txt
awk -v OFS='\t' '{print $1,$2,$3,($3/$2)}'
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seqoverlap.count.txt |sort -k4,4nr >
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seqoverlap.countratio.txt
```

```
# filter out all motifs that have >1 reads around them
annotatePeaks.pl ${HOMOCLUSTERDIR}/PU.1_long_hg19_refChr.bed hg19 -size 200 -d
${CHIPTAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged -nogene -noann >
${ANALYSISDIR}/PU.1_long_hg19_refChr.ann.txt
myFilterFile.pl ${ANALYSISDIR}/PU.1_long_hg19_refChr.ann.txt -column 8 -upperlimit 1 >
${TMPDIR}/tmp.PU.1_long_hg19_refChr.ann.txt
myFilterFile.pl ${TMPDIR}/tmp.PU.1_long_hg19_refChr.ann.txt -column 2 -substring chrY >
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.txt
pos2bed.pl ${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.txt >
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.peaks.bed
$BEDTOOLS intersect -a ${HOMOCLUSTERDIR}/PU.1_long_hg19_refChr.bed -b
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.peaks.bed -u
>${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.bed

homerTools extract ${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.bed ${GENOME_HG19}
>${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.seq.txt

# overlap with bound motifs
join -1 4 -2 1 -t $'\t' <(sort -k4,4
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.bed) <(sort -k1,1
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.seq.txt) >
/misc/data/tmp/PU.1_long_hg19_refChr.unbound.bed
awk -v OFS='\t' '{print $2,$3,$4,($1 "%" $7),$5,$6,$7 }'
/misc/data/tmp/PU.1_long_hg19_refChr.unbound.bed >
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.seq.bed
for w in `cut -f 7 ${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.seq.bed`; do echo
$w; done|sort|uniq -c|sort -k1,1nr|sed 's/^ *//g'|tr [:blank:] \\t >
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.seq.count.txt
join -1 2 -2 2 -t $'\t' <(sort -k2,2
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.seq.count.txt) <(sort -k2,2
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seq.count.txt) >
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seqoverlap.unbound.count.txt
# 3043 motifs

awk -v OFS='\t' '{print $1,$2,$3,($3/$2)}'
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seqoverlap.unbound.count.txt |sort -k4,4nr >
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seqoverlap.unbound.countratio.txt

# non-overlapping motifs
join -v 1 -1 2 -2 2 -t $'\t' <(sort -k2,2
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.seq.count.txt) <(sort -k2,2
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seq.count.txt) >
${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seqoverlap.neverbound.count.txt
# 294 motifs

# generate a random control set of unbound motifs
rm ${ANALYSISDIR}/PU.1_unbound.random.seq.bed
touch ${ANALYSISDIR}/PU.1_unbound.random.seq.bed

# for each sequence, an equivalent number of unbound motifs will be randomly extracted
while read SEQ NOA NOB NOC; do
echo $SEQ $NOA $NOB
awk -v "key=$SEQ" '$7 == key'
"${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.seq.bed" >"${TMPDIR}/tmp.seq.txt"
if [[ "${NOA}" -gt "${NOB}" ]];then
shuf -n ${NOB} "${TMPDIR}/tmp.seq.txt" >>"${ANALYSISDIR}/PU.1_unbound.random.seq.bed"
else
cat "${TMPDIR}/tmp.seq.txt" >>"${ANALYSISDIR}/PU.1_unbound.random.seq.bed"
fi
done <"${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seqoverlap.unbound.countratio.txt"
wc -l ${ANALYSISDIR}/PU.1_unbound.random.seq.bed # 57206 motifs

# generate additional set of motifs sequences that are never bound
rm ${ANALYSISDIR}/PU.1_never.seq.bed
touch ${ANALYSISDIR}/PU.1_never.seq.bed
while read SEQ NOA
do
echo $SEQ $NOA
awk -v "key=$SEQ" '$7 == key'
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.seq.bed > "${TMPDIR}/tmp.seq.txt"
cat "${TMPDIR}/tmp.seq.txt" >> ${ANALYSISDIR}/PU.1_never.seq.bed
done <${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seqoverlap.neverbound.count.txt
wc -l ${ANALYSISDIR}/PU.1_never.seq.bed # 16552 motifs

# ATAC signal before and after PU.1 transfection across bound, non-bound and never-bound
motifs
# recenter on motif
annotatePeaks.pl ${ANALYSISDIR}/PU1.ntag15.filtered.allMotifs.seq.bed hg19 -size 14 -center
```

```
${PU1MOTIF} > ${TMPDIR}/tmp.PU1.ntag15.filtered.allMotifs.seq.txt
annotatePeaks.pl ${ANALYSISDIR}/PU.1_never.seq.bed hg19 -size 14 -center ${PU1MOTIF} >
${TMPDIR}/tmp.PU.1_never.seq.seq.txt
annotatePeaks.pl ${ANALYSISDIR}/PU.1_unbound.random.seq.bed hg19 -size 14 -center
${PU1MOTIF} > ${TMPDIR}/tmp.PU.1_unbound.random.seq.txt
# annotate conservation sites around motifs (hist)
annotatePeaks.pl ${TMPDIR}/tmp.PU1.ntag15.filtered.allMotifs.seq.txt hg19 -size 50 -hist 1
-len 1 -bedGraph "${PHYLOP}" "${PHASTCONS}" >
${HISTDIR}/PU1.ntag15.filtered.allMotifs.conshist1.txt
annotatePeaks.pl ${TMPDIR}/tmp.PU.1_never.seq.seq.txt hg19 -size 50 -hist 1 -len 1 -bedGraph
"${PHYLOP}" "${PHASTCONS}" > ${HISTDIR}/PU.1_never.seq.conshist1.txt
annotatePeaks.pl ${TMPDIR}/tmp.PU.1_unbound.random.seq.txt hg19 -size 50 -hist 1 -len 1
-bedGraph "${PHYLOP}" "${PHASTCONS}" > ${HISTDIR}/PU.1_unbound.random.seq.conshist1.txt

# generation of unbound control regions for 14 Kmeans clusters with matching motif score
composition
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >"${TMPDIR}/control.cl${i}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
resizePosFile.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.${i}.txt 12
6 > ${TMPDIR}/cl${i}.txt
homerTools extract ${TMPDIR}/cl${i}.txt ${GENOME_HG19} > ${TMPDIR}/cl${i}.all.txt
awk -v OFS='\t' '{print \$1, substr(\$2, 1, length(\$2)-1)}' /loctmp/cl${i}.all.txt >
/loctmp/cl${i}.allMotifs.txt # output needs to be cut to 12 bp
join -1 1 -2 1 -t $'\t' <(sort -k1,1 ${TMPDIR}/cl${i}.txt) <(sort -k1,1
${TMPDIR}/cl${i}.allMotifs.txt) > ${TMPDIR}/cl${i}.allMotifs.seq.txt
cut -f 8 ${TMPDIR}/cl${i}.allMotifs.seq.txt > ${TMPDIR}/cl${i}.allMotifs.list.txt
for w in \$(cat ${TMPDIR}/cl${i}.allMotifs.list.txt) ; do echo \$w; done|sort|uniq -c|sort
-k1,1nr|sed 's/^ *//g'|tr [:blank:] \\\t > ${TMPDIR}/cl${i}.allMotifs.seq.count.txt

# overlap with all motifs
join -1 2 -2 2 -t $'\t' <(sort -k2,2
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.seq.count.txt) <(sort -k2,2
${TMPDIR}/cl${i}.allMotifs.seq.count.txt) >
${TMPDIR}/cl${i}.allMotifs.seqoverlap.unbound.count.txt
awk -v OFS='\t' '{print \$1,\$2,\$3,(\$3/\$2)}'
${TMPDIR}/cl${i}.allMotifs.seqoverlap.unbound.count.txt |sort -k4,4nr >
${TMPDIR}/cl${i}.allMotifs.seqoverlap.unbound.countratio.txt

# generate a random control set of unbound motifs
rm ${ANALYSISDIR}/PU.1_unbound.Kmeans14.cl${i}.random.seq.bed
touch ${ANALYSISDIR}/PU.1_unbound.Kmeans14.cl${i}.random.seq.bed
while read SEQ NOA NOB NOC
do
echo \$SEQ \$NOA \$NOB
awk -v "key=\$SEQ" '\$7 == key'
${ANALYSISDIR}/PU.1_long_hg19_refChr.unbound.allMotifs.seq.bed >
"${TMPDIR}/tmp.cl${i}.seq.txt"
if [[ "\$NOA" -gt "\$NOB" ]]; then
shuf -n \$NOB "${TMPDIR}/tmp.cl${i}.seq.txt" >>
${ANALYSISDIR}/PU.1_unbound.Kmeans14.cl${i}.random.seq.bed
else
cat "${TMPDIR}/tmp.cl${i}.seq.txt" >>
${ANALYSISDIR}/PU.1_unbound.Kmeans14.cl${i}.random.seq.bed
fi
done <${TMPDIR}/cl${i}.allMotifs.seqoverlap.unbound.countratio.txt
EOF
chmod 750 "${TMPDIR}/control.cl${i}.${_DATE}.sh"
echo "generating random peaks for cluster ${i}"
screen -dm -S cluster${i} bash -c "bash ${TMPDIR}/control.cl${i}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#--------------------------------------------
for ((i=1;i<=14;i++));do
while [ true ]; do # Endless loop.
pid=`screen -S cluster${i} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample cluster ${i}"
break # Test next one.
else
```

```
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

for ((i=1;i<=14;i++));do
echo "random unbound cluster.$i size: $(wc -l
<${ANALYSISDIR}/PU.1_unbound.Kmeans14.cl${i}.random.seq.bed)"
done

# remaining cluster sizes after motif centering:
# random unbound cluster.1 size: 4210
# random unbound cluster.2 size: 4535
# random unbound cluster.3 size: 4413
# random unbound cluster.4 size: 4249
# random unbound cluster.5 size: 3712
# random unbound cluster.6 size: 3095
# random unbound cluster.7 size: 2995
# random unbound cluster.8 size: 1405
# random unbound cluster.9 size: 2054
# random unbound cluster.10 size: 3269
# random unbound cluster.11 size: 2646
# random unbound cluster.12 size: 1457
# random unbound cluster.13 size: 1216
# random unbound cluster.14 size: 338

# recentering to match peak file
for ((i=1;i<=14;i++));do
bed2pos.pl ${ANALYSISDIR}/PU.1_unbound.Kmeans14.cl${i}.random.seq.bed >
"${TMPDIR}/tmp.cl${i}.pos.txt"
annotatePeaks.pl ${TMPDIR}/tmp.cl${i}.pos.txt hg19 -size 16 -center ${PU1MOTIF} >
${ANALYSISDIR}/PU.1_unbound.Kmeans14.cl${i}.random.pos.txt
done
head ${ANALYSISDIR}/PU.1_unbound.Kmeans14.cl1.random.pos.txt


# Evol. conservation of PU.1 motifs in Kmeans clusters
# use PU.1 centered peaks (both peaks and random controls) to annotate PhyloP and PhastCons
scores
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >"${TMPDIR}/center.cl${i}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
annotatePeaks.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.${i}.txt
hg19 -size 100 -hist 1 -len 1 -bedGraph "${PHYLOP}" "${PHASTCONS}" >
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.${i}.cons.hist1.txt
annotatePeaks.pl ${ANALYSISDIR}/PU.1_unbound.Kmeans14.cl${i}.random.pos.txt hg19 -size 100
-hist 1 -len 1 -bedGraph "${PHYLOP}" "${PHASTCONS}" >
${HISTDIR}/PU.1_unbound.Kmeans14.cl${i}.random.cons.hist1.txt
EOF
chmod 750 "${TMPDIR}/center.cl${i}.${_DATE}.sh"
echo "centering peaks for cluster ${i}"
screen -dm -S cluster${i} bash -c "bash ${TMPDIR}/center.cl${i}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#--------------------------------------------
for ((i=1;i<=14;i++));do
while [ true ]; do # Endless loop.
pid=`screen -S cluster${i} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample cluster ${i}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------
```

```
# phyloP:
paste ${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.1.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.2.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.3.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.4.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.5.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.6.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.7.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.8.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.9.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.10.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.11.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.12.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.13.cons.hist1.txt \
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.14.cons.hist1.txt \
> ${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.all.cons.hist1.txt

tail -n +2 ${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.all.cons.hist1.txt |
cut -f1,2,9,16,23,30,37,44,51,58,65,72,79,86,93 >
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.all.phyloP.hist1.txt
echo
$'Distance\tCluster_1\tCluster_2\tCluster_3\tCluster_4\tCluster_5\tCluster_6\tCluster_7\tClust
er_8\tCluster_9\tCluster_10\tCluster_11\tCluster_12\tCluster_13\tCluster_14' | cat -
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.all.phyloP.hist1.txt >
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.all.phyloP.table.txt

paste ${HISTDIR}/PU.1_unbound.Kmeans14.cl1.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl2.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl3.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl4.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl5.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl6.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl7.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl8.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl9.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl10.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl11.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl12.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl13.random.cons.hist1.txt \
${HISTDIR}/PU.1_unbound.Kmeans14.cl14.random.cons.hist1.txt \
> ${HISTDIR}/PU.1_unbound.Kmeans14.clAll.random.cons.hist1.txt

tail -n +2 ${HISTDIR}/PU.1_unbound.Kmeans14.clAll.random.cons.hist1.txt | cut
-f1,2,9,16,23,30,37,44,51,58,65,72,79,86,93 >
${HISTDIR}/PU.1_unbound.Kmeans14.clAll.random.phyloP.hist1.txt
echo
$'Distance\tCluster_1\tCluster_2\tCluster_3\tCluster_4\tCluster_5\tCluster_6\tCluster_7\tClust
er_8\tCluster_9\tCluster_10\tCluster_11\tCluster_12\tCluster_13\tCluster_14' | cat -
${HISTDIR}/PU.1_unbound.Kmeans14.clAll.random.phyloP.hist1.txt >
${HISTDIR}/PU.1_unbound.Kmeans14.clAll.random.phyloP.table.txt

# phastCons:
tail -n +2 ${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.all.cons.hist1.txt |
cut -f1,5,12,19,26,33,40,47,54,61,68,75,82,89,96 >
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.all.phastCons.hist1.txt
echo
$'Distance\tCluster_1\tCluster_2\tCluster_3\tCluster_4\tCluster_5\tCluster_6\tCluster_7\tClust
er_8\tCluster_9\tCluster_10\tCluster_11\tCluster_12\tCluster_13\tCluster_14' | cat -
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.all.phastCons.hist1.txt >
${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.all.phastCons.table.txt

tail -n +2 ${HISTDIR}/PU.1_unbound.Kmeans14.clAll.random.cons.hist1.txt | cut
-f1,5,12,19,26,33,40,47,54,61,68,75,82,89,96 >
${HISTDIR}/PU.1_unbound.Kmeans14.clAll.random.phastCons.hist1.txt
echo
$'Distance\tCluster_1\tCluster_2\tCluster_3\tCluster_4\tCluster_5\tCluster_6\tCluster_7\tClust
er_8\tCluster_9\tCluster_10\tCluster_11\tCluster_12\tCluster_13\tCluster_14' | cat -
${HISTDIR}/PU.1_unbound.Kmeans14.clAll.random.phastCons.hist1.txt >
${HISTDIR}/PU.1_unbound.Kmeans14.clAll.random.phastCons.table.txt

# Histogram plots for conservation
declare -a COLORS=(gold goldenrod2 darkorange1 darkorange3 firebrick1 firebrick3 firebrick
deeppink darkmagenta deepskyblue2 dodgerblue1 dodgerblue3 blue2 blue4)
declare -a GRAYCOLORS=(gray25 gray30 gray35 gray40 gray45 gray50 gray55 gray60 gray65 gray70
gray75 gray80 gray85 gray90)
```

```
# all in one for PhastCons
COUNT=0
_DATE=$(date +%s)
cat >"${TMPDIR}/R.hist.phastCons.${_DATE}.R" <<EOF
library(ggplot2)
library(grid)
setwd("${HISTDIR}")
d <-
read.table("${HISTDIR}/PU1_Flag_merged.ATACann.kmeans.14.motifCentered.all.phastCons.table.txt
", header=T, sep="\t")
dU <- read.table("${HISTDIR}/PU.1_unbound.Kmeans14.clAll.random.phastCons.table.txt",
header=T, sep="\t")
# shifting motif to middle (homer output is centered to 5'end)
d\$Distance=d\$Distance-5.5
dU\$Distance=dU\$Distance-5.5
# starting the plot
p <- ggplot(data=d, aes(x=Distance,y=Cluster_1))
EOF
for ((i=1;i<=14;i++));do
GRAYCOLOR="${GRAYCOLORS[${COUNT}]}"
cat >>"${TMPDIR}/R.hist.phastCons.${_DATE}.R" <<EOF
p <- p + geom_line(data=dU, aes(x=Distance,y=Cluster_${i}), color="${GRAYCOLOR}", size=0.5)
EOF
COUNT=$((COUNT+=1))
done
COUNT=0
for ((i=1;i<=14;i++));do
COLOR="${COLORS[${COUNT}]}"
cat >>"${TMPDIR}/R.hist.phastCons.${_DATE}.R" <<EOF
p <- p + geom_line(data=d, aes(x=Distance,y=Cluster_${i}), color="${COLOR}", size=0.5)
EOF
COUNT=$((COUNT+=1))
done
cat >>"${TMPDIR}/R.hist.phastCons.${_DATE}.R" <<EOF
p <- p + geom_vline(xintercept=c(-5.5,5.5), linetype="dashed", color="red", size=.25,
alpha=.5)
p <- p + xlab("Distance to motif center") + ylab("PhastCons score")
p <- p + theme_light(base_size=8) + theme(panel.grid.major =
element_blank(),panel.grid.minor = element_blank())
p <- p + scale_x_continuous(expand = c(0,0),limits=c(-15,15)) + scale_y_continuous(expand =
c(0,0),limit=c(0,.5))
pdf(file="${FIGURESDIR}/hist/PhastConsAcrossPU1motif_Kmeans.14.all.pdf", height=2, width=2)
plot(p)
dev.off()
EOF
COUNT=$((COUNT+=1))
chmod 750 "${TMPDIR}/R.hist.phastCons.${_DATE}.R"
R < "${TMPDIR}/R.hist.phastCons.${_DATE}.R" --no-save
rm "${TMPDIR}/R.hist.phastCons.${_DATE}.R"

# genome ontology
_DATE=$(date +%s)
for ((i=1;i<=14;i++));do
cat >"${TMPDIR}/center.cl${i}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
annotatePeaks.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt hg19 -size
given -gtf ${HG19GTF} -annStats
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.genome.stats.txt >
${TMPDIR}/tmp.GO.cl${i}.txt
EOF
chmod 750 "${TMPDIR}/center.cl${i}.${_DATE}.sh"
echo "centering peaks for cluster ${i}"
screen -dm -S cluster${i} bash -c "bash ${TMPDIR}/center.cl${i}.${_DATE}.sh"
done
```

```
# loop to check when screen sessions are done
#--------------------------------------------
for ((i=1;i<=14;i++));do
while [ true ]; do # Endless loop.
pid=`screen -S cluster${i} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample cluster ${i}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

paste ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.1.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.2.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.3.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.4.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.5.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.6.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.7.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.8.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.9.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.10.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.11.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.12.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.13.genome.stats.txt \
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.14.genome.stats.txt \
> ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.all.genome.stats.txt

head -n 6 ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.all.genome.stats.txt | cut
-f1,2,6,10,14,18,22,26,30,34,38,72,42,46,50,54 > ${TMPDIR}/tmp.all.txt
tail -n +2 ${TMPDIR}/tmp.all.txt > ${TMPDIR}/tmp2.all.txt
echo
$'Annotation\tCluster_1\tCluster_2\tCluster_3\tCluster_4\tCluster_5\tCluster_6\tCluster_7\tClu
ster_8\tCluster_9\tCluster_10\tCluster_11\tCluster_12\tCluster_13\tCluster_14' | cat -
${TMPDIR}/tmp2.all.txt >
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.all.barChart.txt

# stacked bar chart for all Kmeans clusters
cat >"${FIGURESDIR}/R/R.GOstackedBarchart.R" <<EOF
library(ggplot2)
library(reshape)
library(scales)
d <- read.table("${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.all.barChart.txt",
header=T, sep="\t")
colnames(d)=c("Ann","1","2","3","4","5","6","7","8","9","10","11","12","13","14")
colors = c("red","yellow","green","blue","black")
datm <- melt(cbind(d[,-1], Annotation = d\$Ann), id.vars = c('Annotation'))
p <- ggplot(datm,aes(x = variable, y = value,fill = Annotation))
p <- p + geom_bar(position = "fill",stat = "identity") +scale_fill_manual(values = colors)
p <- p + theme_light(base_size=8) + theme(panel.grid.major =
element_blank(),panel.grid.minor = element_blank())
p <- p + scale_x_discrete(expand = c(0,0)) + scale_y_continuous(expand = c(0,0),labels =
percent_format())
p <- p + xlab("Kmeans cluster") + ylab("")
pdf(file="${FIGURESDIR}/genomeOntologyBarChartKmeans.pdf", height=2, width=3)
plot(p)
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.GOstackedBarchart.R"
R < "${FIGURESDIR}/R/R.GOstackedBarchart.R" --no-save
```

# # comparison of ETS1, FLI-1 and PU1 binding sites

```
# CNV normalization of tagDirs & peak calling
cd ${TAGDIR}
normalizeTagDirByCopyNumber.pl CTV1_PU1mut_DSG_ETS1_R1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1mut_DSG_FLI1_R1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1_DSG_ETS1_R1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1_DSG_FLI1_R1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1_DSG_FLI1_R2 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1_DSG_ETS1_R2 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1mut_DSG_ETS1_R2 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1mut_DSG_FLI1_R2 -cnv "${CNVFILECTV1}" -remove
```

```
cd ${CHIPINDIR}
normalizeTagDirByCopyNumber.pl CTV1_PU1_DSG_2chrInput -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1mut_DSG_2chrInput -cnv "${CNVFILECTV1}" -remove

cd ${TAGDIR}
makeTagDirectory CTV1_PU1_DSG_FLI1_CNVnormRefChr_merged -d
CTV1_PU1_DSG_FLI1_R1_CNVnormRefChr CTV1_PU1_DSG_FLI1_R2_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_PU1_DSG_ETS1_CNVnormRefChr_merged -d
CTV1_PU1_DSG_ETS1_R1_CNVnormRefChr CTV1_PU1_DSG_ETS1_R2_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_PU1mut_DSG_ETS1_CNVnormRefChr_merged -d
CTV1_PU1mut_DSG_ETS1_R1_CNVnormRefChr CTV1_PU1mut_DSG_ETS1_R2_CNVnormRefChr -genome hg19
-checkGC
makeTagDirectory CTV1_PU1mut_DSG_FLI1_CNVnormRefChr_merged -d
CTV1_PU1mut_DSG_FLI1_R1_CNVnormRefChr CTV1_PU1mut_DSG_FLI1_R2_CNVnormRefChr -genome hg19
-checkGC

cd ${TAGDIR}
findPeaks CTV1_PU1_DSG_ETS1_CNVnormRefChr_merged -i
${CHIPINDIR}/CTV1_PU1_DSG_2chrInput_CNVnormRefChr -style factor -fdr 0.00001 -o
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.peaks.txt
findPeaks CTV1_PU1_DSG_FLI1_CNVnormRefChr_merged -i
${CHIPINDIR}/CTV1_PU1_DSG_2chrInput_CNVnormRefChr -style factor -fdr 0.00001 -o
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.peaks.txt
findPeaks CTV1_PU1mut_DSG_ETS1_CNVnormRefChr_merged -i
${CHIPINDIR}/CTV1_PU1mut_DSG_2chrInput_CNVnormRefChr -style factor -fdr 0.00001 -o
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.peaks.txt
findPeaks CTV1_PU1mut_DSG_FLI1_CNVnormRefChr_merged -i
${CHIPINDIR}/CTV1_PU1mut_DSG_2chrInput_CNVnormRefChr -style factor -fdr 0.00001 -o
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.peaks.txt

# less stringent peaks for overlapping with PU.1
cd ${TAGDIR}
findPeaks CTV1_PU1_DSG_ETS1_CNVnormRefChr_merged -i
${CHIPINDIR}/CTV1_PU1_DSG_2chrInput_CNVnormRefChr -style factor -o
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.lenient.peaks.txt
findPeaks CTV1_PU1_DSG_FLI1_CNVnormRefChr_merged -i
${CHIPINDIR}/CTV1_PU1_DSG_2chrInput_CNVnormRefChr -style factor -o
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.lenient.peaks.txt
findPeaks CTV1_PU1mut_DSG_ETS1_CNVnormRefChr_merged -i
${CHIPINDIR}/CTV1_PU1mut_DSG_2chrInput_CNVnormRefChr -style factor -o
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.lenient.peaks.txt
findPeaks CTV1_PU1mut_DSG_FLI1_CNVnormRefChr_merged -i
${CHIPINDIR}/CTV1_PU1mut_DSG_2chrInput_CNVnormRefChr -style factor -o
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.lenient.peaks.txt

# relevant peaks further filtered for peaks with at least 15 tags
myFilterFile.pl ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.peaks.txt -column 6 -lowerlimit
15 > ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.peaks.txt
myFilterFile.pl ${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.peaks.txt -column 6
-lowerlimit 15 > ${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.peaks.txt
myFilterFile.pl ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.peaks.txt -column 6 -lowerlimit
15 > ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.peaks.txt
myFilterFile.pl ${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.peaks.txt -column 6
-lowerlimit 15 > ${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.peaks.txt

# filtering the merged peak sets
pos2bed.pl ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Filtered out xxx regions with mappability scores below 0.8, leaving 15830 regions.

pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.bed
bed2pos.pl ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.bed >
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt
pos2bed.pl ${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.peaks.txt >
${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Filtered out xxx regions with mappability scores below 0.8, leaving 13485 regions.

pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.bed
bed2pos.pl ${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.bed >
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt
pos2bed.pl ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
```

```
# Filtered out xxx regions with mappability scores below 0.8, leaving 9949 regions.

pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.bed
bed2pos.pl ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.bed >
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt
pos2bed.pl ${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.peaks.txt >
${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Filtered out xxx regions with mappability scores below 0.8, leaving 10260 regions.

pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.bed
bed2pos.pl ${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.bed >
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt
pos2bed.pl ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.lenient.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Filtered out 4304 regions with mappability scores below 0.8, leaving 58793 regions.

pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.lenient.filtered.bed
bed2pos.pl ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.lenient.filtered.bed >
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.lenient.filtered.pos.txt
pos2bed.pl ${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.lenient.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Filtered out 2325 regions with mappability scores below 0.8, leaving 39651 regions.

pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.lenient.filtered.bed
bed2pos.pl ${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.lenient.filtered.bed >
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.lenient.filtered.pos.txt
pos2bed.pl ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.lenient.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Filtered out 3033 regions with mappability scores below 0.8, leaving 43358 regions.

pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.lenient.filtered.bed
bed2pos.pl ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.lenient.filtered.bed >
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.lenient.filtered.pos.txt
pos2bed.pl ${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.lenient.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Filtered out 1729 regions with mappability scores below 0.8, leaving 32454 regions.

pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.lenient.filtered.bed
bed2pos.pl ${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.lenient.filtered.bed >
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.lenient.filtered.pos.txt

# ETS1 and FLI1 consensus motifs (for stringent peaks)
findMotifsGenome.pl ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.bed hg19
${MOTIFDIR}/peaksETS1_DSG_PU1_merged -size 200 -len 7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl ${MOTIFDIR}/peaksETS1_DSG_PU1_merged/homerMotifs.all.motifs
${MOTIFDIR}/peaksETS1_DSG_PU1_merged/final -reduceThresh .75 -matchThresh .6 -pvalue 1e-12
-info 1.5 -cpu 8
findMotifsGenome.pl ${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.bed hg19
${MOTIFDIR}/peaksETS1_DSG_PU1mut_merged -size 200 -len 7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl ${MOTIFDIR}/peaksETS1_DSG_PU1mut_merged/homerMotifs.all.motifs
${MOTIFDIR}/peaksETS1_DSG_PU1mut_merged/final -reduceThresh .75 -matchThresh .6 -pvalue
1e-12 -info 1.5 -cpu 8
findMotifsGenome.pl ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.bed hg19
${MOTIFDIR}/peaksFLI1_DSG_PU1_merged -size 200 -len 7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl ${MOTIFDIR}/peaksFLI1_DSG_PU1_merged/homerMotifs.all.motifs
${MOTIFDIR}/peaksFLI1_DSG_PU1_merged/final -reduceThresh .75 -matchThresh .6 -pvalue 1e-12
-info 1.5 -cpu 8
findMotifsGenome.pl ${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.bed hg19
${MOTIFDIR}/peaksFLI1_DSG_PU1mut_merged -size 200 -len 7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl ${MOTIFDIR}/peaksFLI1_DSG_PU1mut_merged/homerMotifs.all.motifs
${MOTIFDIR}/peaksFLI1_DSG_PU1mut_merged/final -reduceThresh .75 -matchThresh .6 -pvalue
1e-12 -info 1.5 -cpu 8
```

```
# since there is a high degree of overlap, try to select the best motif (11-13mer to match
PU.1) detecting both FLI1 and ETS1.
cat ${MOTIFDIR}/peaksFLI1_DSG_PU1mut_merged/final/homerResults/motif1.motif
${MOTIFDIR}/peaksFLI1_DSG_PU1_merged/final/homerResults/motif1.motif
${MOTIFDIR}/peaksETS1_DSG_PU1mut_merged/final/homerResults/motif1.motif
${MOTIFDIR}/peaksETS1_DSG_PU1_merged/final/homerResults/motif1.motif
${MOTIFDIR}/ETSSPECIFIC_PU1mut/final/homerResults/motif1.motif ${PU1MOTIFR} >
${MOTIFDIR}/combinedETSmotifs.motifs
findMotifsGenome.pl ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.bed hg19
${MOTIFDIR}/peaksETS1_DSG_PU1_merged.combinedETSmotifs -size 200 -mknown
${MOTIFDIR}/combinedETSmotifs.motifs -nomotif -p 8 -h
findMotifsGenome.pl ${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.bed hg19
${MOTIFDIR}/peaksETS1_DSG_PU1mut_merged.combinedETSmotifs -size 200 -mknown
${MOTIFDIR}/combinedETSmotifs.motifs -nomotif -p 8 -h
findMotifsGenome.pl ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.bed hg19
${MOTIFDIR}/peaksFLI1_DSG_PU1_merged.combinedETSmotifs -size 200 -mknown
${MOTIFDIR}/combinedETSmotifs.motifs -nomotif -p 8 -h
findMotifsGenome.pl ${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.bed hg19
${MOTIFDIR}/peaksFLI1_DSG_PU1mut_merged.combinedETSmotifs -size 200 -mknown
${MOTIFDIR}/combinedETSmotifs.motifs -nomotif -p 8 -h

# chose the ETS-specific 12mer because it is most different from the PU1 motif
cp ${MOTIFDIR}/ETSSPECIFIC_PU1mut/final/homerResults/motif1.motif ${CLASS1ETSMOTIF}

# correlation between motif matrices
compareMotifs.pl ${MOTIFDIR}/combinedETSmotifs.motifs ${MOTIFDIR}/combinedETSmotifs -matrix
${ANALYSISDIR}/combinedETSmotifs.correlation.matrix.txt

# generate heatmap for correlation matrix
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.heatmap.P.${_DATE}.R" <<EOF
library(gplots)
library(reshape2)
my_data <- read.delim("${ANALYSISDIR}/combinedETSmotifs.correlation.matrix.txt", header =
TRUE)
data <- data.matrix(my_data[,2:7])
mycol <- colorRampPalette(c("blue","white","red"))(299)
col_breaks = c(seq(.7,.799,length=100), seq(0.8,0.899,length=100), seq(0.9,1,length=100))
pdf(file="${FIGURESDIR}/combinedETSmotifs.correlation.pdf")
heatmap <- heatmap.2(data, Rowv=NA, Colv=NA, col = mycol, breaks=col_breaks,
margins=c(5,10), key=TRUE, density.info="none", trace="none", dendrogram="none")
dev.off()
EOF
chmod 750 "/loctmp/R.heatmap.P.${_DATE}.R"
R < /loctmp/R.heatmap.P.${_DATE}.R --no-save
rm /loctmp/R.heatmap.P.${_DATE}.R

# ghist plots ChIP & ATAC data
# combined plots
declare -a SETS=("${TAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged" \
"${ATACDIR}/CTV1_PU1mut_CNVnormRefChr_merged" "${ATACDIR}/CTV1_PU1_CNVnormRefChr_merged" \
"${TAGDIR}/CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_H3K27ac_CNVnormRefChr_merged" \
"${TAGDIR}/CTV1_PU1delA_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delAQP_Flag_CNVnormRefChr_merged" \
"${TAGDIR}/CTV1_PU1delQ_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delP_Flag_CNVnormRefChr_merged" \
"${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr")
declare -a NAMES=(PU1mut PU1 ATACmut ATACPU1 H3K27acmut H3K27acPU1 delA delAQP delQ delP
PU1less)
declare -a SETS=("${TAGDIR}/CTV1_PU1mut_DSG_ETS1_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_DSG_ETS1_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_DSG_FLI1_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_DSG_FLI1_CNVnormRefChr_merged")
declare -a NAMES=(ETS1-PU1mut ETS1-PU1 FLI1-PU1mut FLI1-PU1)

COUNT=0
for SET in ${SETS[@]}; do
NAME="${NAMES[${COUNT}]}"
_DATE=$(date +%s)
cat >"/loctmp/ghist.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
```

```
export PATH
cd /loctmp
annotatePeaks.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt hg19
-size 1000 -hist 25 -ghist -d ${SET} >
"${GHISTDIR}/kmeans.14.sorted.cleaned.${NAME}.ghist.txt"
EOF
COUNT=$((COUNT+=1))
chmod 750 "/loctmp/ghist.${NAME}.${_DATE}.sh"
echo "annotating ghist for ${NAME}"
screen -dm -S ${NAME} bash -c "bash /loctmp/ghist.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#-------------------------------------------
for NAME in ${NAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

# Histogram for ETS1, FLI1 & ATAC
declare -a NAMES=(PU1mut PU1 ETS1-PU1mut ETS1-PU1 FLI1-PU1mut FLI1-PU1 ATACmut ATACPU1)
SCALES="25 25 15 15 15 15 15 15"
COLORS="blue blue slateblue4 slateblue4 magenta4 magenta4 navy navy"
GHISTSAMPLES="${GHISTDIR}/kmeans.14.sorted.cleaned.${NAMES[0]}.ghist.txt"
for ((i=1;i<8;i++));do
GHISTSAMPLES="${GHISTSAMPLES} ${GHISTDIR}/kmeans.14.sorted.cleaned.${NAMES[$i]}.ghist.txt"
done
plotGHIST.sh -f "${GHISTSAMPLES}" -h 22315 -w 2000 -d "${FIGURESDIR}" -n PU1.ETS.FLI.ATAC -m
"${SCALES}" -c "${COLORS}"

# ETS1 and FLI1 overlap (based on peak calling or DGE)
# overlap based on peak calling
mergePeaks -d 100 ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt -code -venn
${PEAKDIR}/FLI1_ETS1_DSG_PU1_merged_overlap.venn.txt -prefix ${PEAKDIR}/peaks_sep >
${TMPDIR}/tmp0.ann.txt
mv ${PEAKDIR}/peaks_sep_01 ${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_ETS1specifc.txt
mv ${PEAKDIR}/peaks_sep_10 ${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_FLI1specifc.txt
mv ${PEAKDIR}/peaks_sep_11 ${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_common.txt

mergePeaks -d 100 ${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt -code -venn
${PEAKDIR}/FLI1_ETS1_DSG_PU1mut_merged_overlap.venn.txt -prefix ${PEAKDIR}/peaks_sep >
${TMPDIR}/tmp1.ann.txt
mv ${PEAKDIR}/peaks_sep_01 ${PEAKDIR}/ETS1vsFLI1_DSG_PU1mut_merged_ETS1specifc.txt
mv ${PEAKDIR}/peaks_sep_10 ${PEAKDIR}/ETS1vsFLI1_DSG_PU1mut_merged_FLI1specifc.txt
mv ${PEAKDIR}/peaks_sep_11 ${PEAKDIR}/ETS1vsFLI1_DSG_PU1mut_merged_common.txt

mergePeaks -d 100 ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.lenient.peaks.txt -venn
${PEAKDIR}/FLI1_DSG_PU1_merged_overlap_ETS1lenient.venn.txt > ${TMPDIR}/tmp2.ann.txt

mergePeaks -d 100 ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.lenient.peaks.txt -venn
${PEAKDIR}/ETS1_DSG_PU1_merged_overlap_FLI1lenient.venn.txt > ${TMPDIR}/tmp3.ann.txt

mergePeaks -d 100 ${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.lenient.peaks.txt -venn
${PEAKDIR}/FLI1_DSG_PU1mut_merged_overlap_ETS1lenient.venn.txt > ${TMPDIR}/tmp4.ann.txt

mergePeaks -d 100 ${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.lenient.peaks.txt -venn
${PEAKDIR}/ETS1_DSG_PU1mut_merged_overlap_FLI1lenient.venn.txt > ${TMPDIR}/tmp5.ann.txt

# motif signatures in specific peaks - direct comparison
findMotifsGenome.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_FLI1specifc.txt hg19
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1_merged_FLI1specifc_direct -bg
${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_ETS1specifc.txt -size 200 -len 7,8,9,10,11,12,13,14 -p
8 -h
compareMotifs.pl
```

```
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1_merged_FLI1specifc_direct/homerMotifs.all.motifs
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1_merged_FLI1specifc_direct/final -reduceThresh .75
-matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 8

findMotifsGenome.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_ETS1specifc.txt hg19
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1_merged_ETS1specifc_direct -bg
${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_FLI1specifc.txt -size 200 -len 7,8,9,10,11,12,13,14 -p
8 -h
compareMotifs.pl
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1_merged_ETS1specifc_direct/homerMotifs.all.motifs
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1_merged_ETS1specifc_direct/final -reduceThresh .75
-matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 8

findMotifsGenome.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1mut_merged_FLI1specifc.txt hg19
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1mut_merged_FLI1specifc_direct -bg
${PEAKDIR}/ETS1vsFLI1_DSG_PU1mut_merged_ETS1specifc.txt -size 200 -len 7,8,9,10,11,12,13,14
-p 8 -h
compareMotifs.pl
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1mut_merged_FLI1specifc_direct/homerMotifs.all.motifs
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1mut_merged_FLI1specifc_direct/final -reduceThresh .75
-matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 8

findMotifsGenome.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1mut_merged_ETS1specifc.txt hg19
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1mut_merged_ETS1specifc_direct -bg
${PEAKDIR}/ETS1vsFLI1_DSG_PU1mut_merged_FLI1specifc.txt -size 200 -len 7,8,9,10,11,12,13,14
-p 8 -h
compareMotifs.pl
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1mut_merged_ETS1specifc_direct/homerMotifs.all.motifs
${MOTIFDIR}/peaksETS1vsFLI1_DSG_PU1mut_merged_ETS1specifc_direct/final -reduceThresh .75
-matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 8

# overlap stringent and lenient for venn diagram
mergePeaks -d 100 ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.lenient.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.lenient.filtered.pos.txt -code -venn
${ANALYSISDIR}/FLI1_ETS1_DSG_stringentAndLenient_overlap.venn.txt > ${TMPDIR}/tmp0.ann.txt

cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"${TMPDIR}/R.venneuler.${_DATE}.R" <<EOF
library(venneuler)
a0001=19547
a0010=4898
a0011=22612
a0101=1045
a0111=5899
a1010=259
a1011=804
a1111=8886
MyVenn <-
venneuler(c(C=a0010,D=a0001,"A&C"=a1010,"A&C&D"=a1011,"A&B&C&D"=a1111,"B&D"=a0101,"B&C&D"=a011
1,"C&D"=a0011))
MyVenn\$colors <- c(.2,.4,.6,.8)
pdf(file="${FIGURESDIR}/vennEulerETSFLI1_4way.pdf", height=5, width=3)
plot(MyVenn)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.motifbean.${_DATE}.R"
R < ${TMPDIR}/R.venneuler.${_DATE}.R --no-save
rm ${TMPDIR}/R.venneuler.${_DATE}.R

# overlap based on DGE
# ETS1 vs FLI1 (PU1mut)

cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_PU1mut_DSG_ETS1_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1mut_DSG_ETS1_R2_CNVnormRefChr \
-b ${TAGDIR}/CTV1_PU1mut_DSG_FLI1_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1mut_DSG_FLI1_R2_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt >
${DIFFDIR}/ETS1vsFLI1.DSG_PU1mut.peaks.DESEQ.txt
# Output Stats bg vs. target:
# Total Genes: 13485
# Total Up-regulated in target vs. bg: 286 (2.121%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 2 (0.015%) [log2fold<-1, FDR<0.05]
# FLI1 vs ETS1 (PU1mut)
```

```
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -b ${TAGDIR}/CTV1_PU1mut_DSG_ETS1_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1mut_DSG_ETS1_R2_CNVnormRefChr \
-t ${TAGDIR}/CTV1_PU1mut_DSG_FLI1_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1mut_DSG_FLI1_R2_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt >
${DIFFDIR}/FLI1vsETS1.DSG_PU1mut.peaks.DESEQ.txt
# Output Stats bg vs. target:
# Total Genes: 10260
# Total Up-regulated in target vs. bg: 1 (0.010%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 8 (0.078%) [log2fold<-1, FDR<0.05]
# ETS1 vs FLI1 (PU1)

cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_PU1_DSG_ETS1_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_DSG_ETS1_R2_CNVnormRefChr \
-b ${TAGDIR}/CTV1_PU1_DSG_FLI1_R1_CNVnormRefChr ${TAGDIR}/CTV1_PU1_DSG_FLI1_R2_CNVnormRefChr
-genome hg19 \
-p ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt >
${DIFFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.txt
# Output Stats bg vs. target:
# Total Genes: 15830
# Total Up-regulated in target vs. bg: 1135 (7.170%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 8 (0.051%) [log2fold<-1, FDR<0.05]
# FLI1 vs ETS1 (PU1)

cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -b ${TAGDIR}/CTV1_PU1_DSG_ETS1_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_DSG_ETS1_R2_CNVnormRefChr \
-t ${TAGDIR}/CTV1_PU1_DSG_FLI1_R1_CNVnormRefChr ${TAGDIR}/CTV1_PU1_DSG_FLI1_R2_CNVnormRefChr
-genome hg19 \
-p ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt >
${DIFFDIR}/FLI1vsETS1.DSG_PU1.peaks.DESEQ.txt
# Output Stats bg vs. target:
# Total Genes: 9949
# Total Up-regulated in target vs. bg: 4 (0.040%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 58 (0.583%) [log2fold<-1, FDR<0.05]
# ETS1 PU1 vs PU1mut

cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_PU1_DSG_ETS1_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_DSG_ETS1_R2_CNVnormRefChr \
-b ${TAGDIR}/CTV1_PU1mut_DSG_ETS1_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1mut_DSG_ETS1_R2_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt >
${DIFFDIR}/ETS1_PU1vsPU1mut.DSG_PU1.peaks.DESEQ.txt
# Output Stats bg vs. target:
# Total Genes: 15830
# Total Up-regulated in target vs. bg: 5080 (32.091%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 452 (2.855%) [log2fold<-1, FDR<0.05]
# FLI1 PU1 vs PU1mut

cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_PU1_DSG_FLI1_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_DSG_FLI1_R2_CNVnormRefChr \
-b ${TAGDIR}/CTV1_PU1mut_DSG_FLI1_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1mut_DSG_FLI1_R2_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt >
${DIFFDIR}/FLI1_PU1vsPU1mut.DSG_PU1.peaks.DESEQ.txt
# Output Stats bg vs. target:
# Total Genes: 9949
# Total Up-regulated in target vs. bg: 2518 (25.309%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 50 (0.503%) [log2fold<-1, FDR<0.05]

# only ETS1 vs FLI1 (PU1) has a significant enrichment of diff. peaks. Motif enrichment:
findMotifsGenome.pl ${DIFFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.txt hg19
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ -size 200 -len 7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl ${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ/homerMotifs.all.motifs
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ/final -reduceThresh .75 -matchThresh .6 -pvalue
1e-12 -info 1.5 -cpu 8

findMotifsGenome.pl ${DIFFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.txt hg19
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.PU1masked -maskMotif ${PU1MOTIF} -size 200 -len
7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl ${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.PU1masked/homerMotifs.all.motifs
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.PU1masked/final -reduceThresh .75 -matchThresh .6
-pvalue 1e-12 -info 1.5 -cpu 8
```

```
findMotifsGenome.pl ${DIFFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.txt hg19
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.bg -bg
${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_common.txt -size 200 -len 7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl ${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.bg/homerMotifs.all.motifs
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.bg/final -reduceThresh .75 -matchThresh .6
-pvalue 1e-12 -info 1.5 -cpu 8

findMotifsGenome.pl ${DIFFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.txt hg19
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.PU1masked.bg -bg
${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_common.txt -maskMotif ${PU1MOTIF} -size 200 -len
7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.PU1masked.bg/homerMotifs.all.motifs
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.PU1masked.bg/final -reduceThresh .75 -matchThresh
.6 -pvalue 1e-12 -info 1.5 -cpu 8

# compare with common sites
findMotifsGenome.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_common.txt hg19
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.common -size 200 -len 7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl ${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.common/homerMotifs.all.motifs
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.common/final -reduceThresh .75 -matchThresh .6 -pvalue
1e-12 -info 1.5 -cpu 8

findMotifsGenome.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_common.txt hg19
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.common.PU1masked -maskMotif ${PU1MOTIF} -size 200 -len
7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.common.PU1masked/homerMotifs.all.motifs
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1.peaks.common.PU1masked/final -reduceThresh .75 -matchThresh
.6 -pvalue 1e-12 -info 1.5 -cpu 8

findMotifsGenome.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1mut_merged_common.txt hg19
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1mut.peaks.common -size 200 -len 7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl ${MOTIFDIR}/ETS1vsFLI1.DSG_PU1mut.peaks.common/homerMotifs.all.motifs
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1mut.peaks.common/final -reduceThresh .75 -matchThresh .6
-pvalue 1e-12 -info 1.5 -cpu 8

findMotifsGenome.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1mut_merged_common.txt hg19
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1mut.peaks.common.PU1masked -maskMotif ${PU1MOTIF} -size 200
-len 7,8,9,10,11,12,13,14 -p 8 -h
compareMotifs.pl
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1mut.peaks.common.PU1masked/homerMotifs.all.motifs
${MOTIFDIR}/ETS1vsFLI1.DSG_PU1mut.peaks.common.PU1masked/final -reduceThresh .75
-matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 8

# ghist plot for differential ETS1 peaks
declare -a TAGDIRSETS=("${TAGDIR}/CTV1_PU1mut_DSG_ETS1_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_DSG_ETS1_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_DSG_FLI1_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_DSG_FLI1_CNVnormRefChr_merged")
annotatePeaks.pl ${DIFFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.txt hg19 -size 1000 -hist 25
-ghist -d ${TAGDIR}/CTV1_PU1_DSG_ETS1_CNVnormRefChr_merged >
"${GHISTDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.ETS1.ghist.txt"
annotatePeaks.pl ${DIFFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.txt hg19 -size 1000 -hist 25
-ghist -d ${TAGDIR}/CTV1_PU1mut_DSG_ETS1_CNVnormRefChr_merged >
"${GHISTDIR}/ETS1vsFLI1.DSG_PU1mut.peaks.DESEQ.ETS1.ghist.txt"
annotatePeaks.pl ${DIFFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.txt hg19 -size 1000 -hist 25
-ghist -d ${TAGDIR}/CTV1_PU1_DSG_FLI1_CNVnormRefChr_merged >
"${GHISTDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.FLI1.ghist.txt"
annotatePeaks.pl ${DIFFDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.txt hg19 -size 1000 -hist 25
-ghist -d ${TAGDIR}/CTV1_PU1mut_DSG_FLI1_CNVnormRefChr_merged >
"${GHISTDIR}/ETS1vsFLI1.DSG_PU1mut.peaks.DESEQ.FLI1.ghist.txt"
annotatePeaks.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_common.txt hg19 -size 1000 -hist 25
-ghist -d ${TAGDIR}/CTV1_PU1_DSG_ETS1_CNVnormRefChr_merged >
"${GHISTDIR}/ETS1vsFLI1.DSG_PU1.common.ETS1.ghist.txt"
annotatePeaks.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_common.txt hg19 -size 1000 -hist 25
-ghist -d ${TAGDIR}/CTV1_PU1mut_DSG_ETS1_CNVnormRefChr_merged >
"${GHISTDIR}/ETS1vsFLI1.DSG_PU1mut.common.ETS1.ghist.txt"
annotatePeaks.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_common.txt hg19 -size 1000 -hist 25
-ghist -d ${TAGDIR}/CTV1_PU1_DSG_FLI1_CNVnormRefChr_merged >
"${GHISTDIR}/ETS1vsFLI1.DSG_PU1.common.FLI1.ghist.txt"
annotatePeaks.pl ${PEAKDIR}/ETS1vsFLI1_DSG_PU1_merged_common.txt hg19 -size 1000 -hist 25
-ghist -d ${TAGDIR}/CTV1_PU1mut_DSG_FLI1_CNVnormRefChr_merged >
"${GHISTDIR}/ETS1vsFLI1.DSG_PU1mut.common.FLI1.ghist.txt"
```

```
# Histogram for ETS1, FLI1 & PU1

cd ${GHISTDIR}
# specific sites
plotHIST.sh \
-f "${GHISTDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.ETS1.ghist.txt
${GHISTDIR}/ETS1vsFLI1.DSG_PU1.peaks.DESEQ.FLI1.ghist.txt" \
-s "ETS1-PU1 FLI1-PU1" -c "slateblue4 magenta4" -x 1000 -y "0 18" -d ${FIGURESDIR}/hist -n
ETS-specific_ETS.FLI.PU1

# common sites
plotHIST.sh \
-f "${GHISTDIR}/ETS1vsFLI1.DSG_PU1.common.ETS1.ghist.txt
${GHISTDIR}/ETS1vsFLI1.DSG_PU1.common.FLI1.ghist.txt" \
-s "ETS1-PU1 FLI1-PU1" -c "slateblue4 magenta4" -x 1000 -y "0 18" -d ${FIGURESDIR}/hist -n
common_ETS.FLI.PU1

# ETS1 and FLI1 overlap with PU.1
# comparing stringent PU.1 peaks with both stringent and lenient ETS1/FLI1 peaks
mergePeaks -d 100 ${PU1CTV1PEAKS}
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt -code >
${TMPDIR}/tmp0.ann.txt
myFilterFile.pl ${TMPDIR}/tmp0.ann.txt -column 7 -lowerlimit 11 > ${TMPDIR}/tmp1.ann.txt
pos2bed.pl ${TMPDIR}/tmp1.ann.txt > ${PEAKDIR}/commonETS1_PU1.stringent.bed
mergePeaks -d 100 ${PU1CTV1PEAKS}
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.lenient.filtered.pos.txt -code > ${TMPDIR}/tmp0.ann.txt
myFilterFile.pl ${TMPDIR}/tmp0.ann.txt -column 7 -lowerlimit 11 > ${TMPDIR}/tmp1.ann.txt
pos2bed.pl ${TMPDIR}/tmp1.ann.txt > ${PEAKDIR}/commonETS1_PU1.lenient.bed
mergePeaks -d 100 ${PU1CTV1PEAKS}
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt -code >
${TMPDIR}/tmp0.ann.txt
myFilterFile.pl ${TMPDIR}/tmp0.ann.txt -column 7 -lowerlimit 11 > ${TMPDIR}/tmp1.ann.txt
pos2bed.pl ${TMPDIR}/tmp1.ann.txt > ${PEAKDIR}/commonETS1_PU1mut.stringent.bed
mergePeaks -d 100 ${PU1CTV1PEAKS}
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.lenient.filtered.pos.txt -code >
${TMPDIR}/tmp0.ann.txt
myFilterFile.pl ${TMPDIR}/tmp0.ann.txt -column 7 -lowerlimit 11 > ${TMPDIR}/tmp1.ann.txt
pos2bed.pl ${TMPDIR}/tmp1.ann.txt > ${PEAKDIR}/commonETS1_PU1mut.lenient.bed
mergePeaks -d 100 ${PU1CTV1PEAKS}
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt -code >
${TMPDIR}/tmp0.ann.txt
myFilterFile.pl ${TMPDIR}/tmp0.ann.txt -column 7 -lowerlimit 11 > ${TMPDIR}/tmp1.ann.txt
pos2bed.pl ${TMPDIR}/tmp1.ann.txt > ${PEAKDIR}/commonFLI1_PU1.stringent.bed
mergePeaks -d 100 ${PU1CTV1PEAKS}
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.lenient.filtered.pos.txt -code > ${TMPDIR}/tmp0.ann.txt
myFilterFile.pl ${TMPDIR}/tmp0.ann.txt -column 7 -lowerlimit 11 > ${TMPDIR}/tmp1.ann.txt
pos2bed.pl ${TMPDIR}/tmp1.ann.txt > ${PEAKDIR}/commonFLI1_PU1.lenient.bed
mergePeaks -d 100 ${PU1CTV1PEAKS}
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt -code >
${TMPDIR}/tmp0.ann.txt
myFilterFile.pl ${TMPDIR}/tmp0.ann.txt -column 7 -lowerlimit 11 > ${TMPDIR}/tmp1.ann.txt
pos2bed.pl ${TMPDIR}/tmp1.ann.txt > ${PEAKDIR}/commonFLI1_PU1mut.stringent.bed
mergePeaks -d 100 ${PU1CTV1PEAKS}
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.lenient.filtered.pos.txt -code >
${TMPDIR}/tmp0.ann.txt
myFilterFile.pl ${TMPDIR}/tmp0.ann.txt -column 7 -lowerlimit 11 > ${TMPDIR}/tmp1.ann.txt
pos2bed.pl ${TMPDIR}/tmp1.ann.txt > ${PEAKDIR}/commonFLI1_PU1mut.lenient.bed
mergePeaks -d 100 ${PU1CTV1PEAKS}
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt -code -venn
${PEAKDIR}/PU1_FLI1_ETS1_DSG_PU1_merged_overlap.venn.txt -prefix ${PEAKDIR}/peaks_sep >
${TMPDIR}/tmp0.ann.txt

# are ETS peaks enriched in the 14 Kmeans clusters compared to all peaks?
rm ${ANALYSISDIR}/ETS1Enrichment_PU1.Kmeans.14.lenient.txt
declare -a PEAKS=()
declare -a ETS1PEAKS=()
TOTALPEAKS=$(wc -l <"${PU1CTV1PEAKS}")
TOTALETS1=$(wc -l <"${PEAKDIR}/commonETS1_PU1.lenient.bed")
TAB=$(echo -e "\t")
for ((i=1;i<=14;i++));do
pos2bed.pl "${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt" >
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed"
PEAKS[${i}]=$(wc -l <"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed")
$BEDTOOLS intersect -a <(cut -f1-3 "${PEAKDIR}/commonETS1_PU1.lenient.bed") -b
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed" -u >
/misc/data/tmp/commonETS1_PU1.kmeans.14.sorted.${i}.bed
ETS1PEAKS[${i}]=$(wc -l <"/misc/data/tmp/commonETS1_PU1.kmeans.14.sorted.${i}.bed")
```

```
cat >>"${ANALYSISDIR}/ETS1Enrichment_PU1.Kmeans.14.lenient.txt" <<EOF
Cluster${i}$TAB$TOTALPEAKS$TAB$TOTALETS1$TAB${PEAKS[${i}]}$TAB${ETS1PEAKS[${i}]}
EOF
done

hyperTab.pl ${ANALYSISDIR}/ETS1Enrichment_PU1.Kmeans.14.lenient.txt >
${ANALYSISDIR}/ETS1Enrichment_PU1.Kmeans.14.lenient.fisher.txt
rm ${ANALYSISDIR}/ETS1Enrichment_PU1mut.Kmeans.14.stringent.txt
declare -a PEAKS=()
declare -a ETS1PEAKS=()
TOTALPEAKS=$(wc -l <"${PU1CTV1PEAKS}")
TOTALETS1=$(wc -l <"${PEAKDIR}/commonETS1_PU1mut.stringent.bed")
TAB=$(echo -e "\t")
for ((i=1;i<=14;i++));do
pos2bed.pl "${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt" >
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed"
PEAKS[${i}]=$(wc -l <"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed")
$BEDTOOLS intersect -a <(cut -f1-3 "${PEAKDIR}/commonETS1_PU1mut.stringent.bed") -b
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed" -u >
/misc/data/tmp/commonETS1_PU1mut.kmeans.14.sorted.${i}.bed
ETS1PEAKS[${i}]=$(wc -l <"/misc/data/tmp/commonETS1_PU1mut.kmeans.14.sorted.${i}.bed")
cat >>"${ANALYSISDIR}/ETS1Enrichment_PU1mut.Kmeans.14.stringent.txt" <<EOF
Cluster${i}$TAB$TOTALPEAKS$TAB$TOTALETS1$TAB${PEAKS[${i}]}$TAB${ETS1PEAKS[${i}]}
EOF
done

hyperTab.pl ${ANALYSISDIR}/ETS1Enrichment_PU1mut.Kmeans.14.stringent.txt >
${ANALYSISDIR}/ETS1Enrichment_PU1mut.Kmeans.14.stringent.fisher.txt
rm ${ANALYSISDIR}/ETS1Enrichment_PU1mut.Kmeans.14.lenient.txt
declare -a PEAKS=()
declare -a ETS1PEAKS=()
TOTALPEAKS=$(wc -l <"${PU1CTV1PEAKS}")
TOTALETS1=$(wc -l <"${PEAKDIR}/commonETS1_PU1mut.lenient.bed")
TAB=$(echo -e "\t")
for ((i=1;i<=14;i++));do
pos2bed.pl "${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt" >
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed"
PEAKS[${i}]=$(wc -l <"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed")
$BEDTOOLS intersect -a <(cut -f1-3 "${PEAKDIR}/commonETS1_PU1mut.lenient.bed") -b
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed" -u >
/misc/data/tmp/commonETS1_PU1mut.kmeans.14.sorted.${i}.bed
ETS1PEAKS[${i}]=$(wc -l <"/misc/data/tmp/commonETS1_PU1mut.kmeans.14.sorted.${i}.bed")
cat >>"${ANALYSISDIR}/ETS1Enrichment_PU1mut.Kmeans.14.lenient.txt" <<EOF
Cluster${i}$TAB$TOTALPEAKS$TAB$TOTALETS1$TAB${PEAKS[${i}]}$TAB${ETS1PEAKS[${i}]}
EOF
done

hyperTab.pl ${ANALYSISDIR}/ETS1Enrichment_PU1mut.Kmeans.14.lenient.txt >
${ANALYSISDIR}/ETS1Enrichment_PU1mut.Kmeans.14.lenient.fisher.txt
rm ${ANALYSISDIR}/FLI1Enrichment_PU1.Kmeans.14.stringent.txt
declare -a PEAKS=()
declare -a FLI1PEAKS=()
TOTALPEAKS=$(wc -l <"${PU1CTV1PEAKS}")
TOTALFLI1=$(wc -l <"${PEAKDIR}/commonFLI1_PU1.stringent.bed")
TAB=$(echo -e "\t")
for ((i=1;i<=14;i++));do
pos2bed.pl "${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt" >
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed"
PEAKS[${i}]=$(wc -l <"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed")
$BEDTOOLS intersect -a <(cut -f1-3 "${PEAKDIR}/commonFLI1_PU1.stringent.bed") -b
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed" -u >
/misc/data/tmp/commonFLI1_PU1.kmeans.14.sorted.${i}.bed
FLI1PEAKS[${i}]=$(wc -l <"/misc/data/tmp/commonFLI1_PU1.kmeans.14.sorted.${i}.bed")
cat >>"${ANALYSISDIR}/FLI1Enrichment_PU1.Kmeans.14.stringent.txt" <<EOF
Cluster${i}$TAB$TOTALPEAKS$TAB$TOTALFLI1$TAB${PEAKS[${i}]}$TAB${FLI1PEAKS[${i}]}
EOF
done

hyperTab.pl ${ANALYSISDIR}/FLI1Enrichment_PU1.Kmeans.14.stringent.txt >
${ANALYSISDIR}/FLI1Enrichment_PU1.Kmeans.14.stringent.fisher.txt
rm ${ANALYSISDIR}/FLI1Enrichment_PU1.Kmeans.14.lenient.txt
declare -a PEAKS=()
declare -a FLI1PEAKS=()
TOTALPEAKS=$(wc -l <"${PU1CTV1PEAKS}")
TOTALFLI1=$(wc -l <"${PEAKDIR}/commonFLI1_PU1.lenient.bed")
TAB=$(echo -e "\t")
for ((i=1;i<=14;i++));do
pos2bed.pl "${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt" >
```

```
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed"
PEAKS[${i}]=$(wc -l <"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed")
$BEDTOOLS intersect -a <(cut -f1-3 "${PEAKDIR}/commonFLI1_PU1.lenient.bed") -b
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed" -u >
/misc/data/tmp/commonFLI1_PU1.kmeans.14.sorted.${i}.bed
FLI1PEAKS[${i}]=$(wc -l <"/misc/data/tmp/commonFLI1_PU1.kmeans.14.sorted.${i}.bed")
cat >>"${ANALYSISDIR}/FLI1Enrichment_PU1.Kmeans.14.lenient.txt" <<EOF
Cluster${i}$TAB$TOTALPEAKS$TAB$TOTALFLI1$TAB${PEAKS[${i}]}$TAB${FLI1PEAKS[${i}]}
EOF
done

hyperTab.pl ${ANALYSISDIR}/FLI1Enrichment_PU1.Kmeans.14.lenient.txt >
${ANALYSISDIR}/FLI1Enrichment_PU1.Kmeans.14.lenient.fisher.txt
rm ${ANALYSISDIR}/FLI1Enrichment_PU1mut.Kmeans.14.stringent.txt
declare -a PEAKS=()
declare -a FLI1PEAKS=()
TOTALPEAKS=$(wc -l <"${PU1CTV1PEAKS}")
TOTALFLI1=$(wc -l <"${PEAKDIR}/commonFLI1_PU1mut.stringent.bed")
TAB=$(echo -e "\t")
for ((i=1;i<=14;i++));do
pos2bed.pl "${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt" >
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed"
PEAKS[${i}]=$(wc -l <"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed")
$BEDTOOLS intersect -a <(cut -f1-3 "${PEAKDIR}/commonFLI1_PU1mut.stringent.bed") -b
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed" -u >
/misc/data/tmp/commonFLI1_PU1mut.kmeans.14.sorted.${i}.bed
FLI1PEAKS[${i}]=$(wc -l <"/misc/data/tmp/commonFLI1_PU1mut.kmeans.14.sorted.${i}.bed")
cat >>"${ANALYSISDIR}/FLI1Enrichment_PU1mut.Kmeans.14.stringent.txt" <<EOF
Cluster${i}$TAB$TOTALPEAKS$TAB$TOTALFLI1$TAB${PEAKS[${i}]}$TAB${FLI1PEAKS[${i}]}
EOF
done

hyperTab.pl ${ANALYSISDIR}/FLI1Enrichment_PU1mut.Kmeans.14.stringent.txt >
${ANALYSISDIR}/FLI1Enrichment_PU1mut.Kmeans.14.stringent.fisher.txt
rm ${ANALYSISDIR}/FLI1Enrichment_PU1mut.Kmeans.14.lenient.txt
declare -a PEAKS=()
declare -a FLI1PEAKS=()
TOTALPEAKS=$(wc -l <"${PU1CTV1PEAKS}")
TOTALFLI1=$(wc -l <"${PEAKDIR}/commonFLI1_PU1mut.lenient.bed")
TAB=$(echo -e "\t")
for ((i=1;i<=14;i++));do
pos2bed.pl "${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt" >
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed"
PEAKS[${i}]=$(wc -l <"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed")
$BEDTOOLS intersect -a <(cut -f1-3 "${PEAKDIR}/commonFLI1_PU1mut.lenient.bed") -b
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed" -u >
/misc/data/tmp/commonFLI1_PU1mut.kmeans.14.sorted.${i}.bed
FLI1PEAKS[${i}]=$(wc -l <"/misc/data/tmp/commonFLI1_PU1mut.kmeans.14.sorted.${i}.bed")
cat >>"${ANALYSISDIR}/FLI1Enrichment_PU1mut.Kmeans.14.lenient.txt" <<EOF
Cluster${i}$TAB$TOTALPEAKS$TAB$TOTALFLI1$TAB${PEAKS[${i}]}$TAB${FLI1PEAKS[${i}]}
EOF
done

hyperTab.pl ${ANALYSISDIR}/FLI1Enrichment_PU1mut.Kmeans.14.lenient.txt >
${ANALYSISDIR}/FLI1Enrichment_PU1mut.Kmeans.14.lenient.fisher.txt

# barplots for percent overlap/cluster
cut -f 1,4,5,10 <(paste ${ANALYSISDIR}/ETS1Enrichment_PU1mut.Kmeans.14.stringent.txt
${ANALYSISDIR}/ETS1Enrichment_PU1.Kmeans.14.stringent.txt) >
${TMPDIR}/tmp.ETS1Enrichment.stringent.txt
cat >"${FIGURESDIR}/R/R.ETS1Enrichment.stringent.stackedBarchart.R" <<EOF
d <- read.table("${TMPDIR}/tmp.ETS1Enrichment.stringent.txt", sep="\t")
colnames(d)=c("Ann","total","PUmut","PU")
d\$percentPUmut <- (d\$PUmut / d\$total) * 100
d\$percentPU <- (d\$PU / d\$total) * 100
dr <- d[rev(rownames(d)),]
counts <- data.matrix(t(dr[,c(5,6)]))
pdf(file="${FIGURESDIR}/ETS1Enrichment.stringent.BarChartKmeans.pdf", height=5, width=3)
barplot(counts, xlab="% overlap ETS1/PU.1", xlim=c(0,100), ylab="cluster",
space=c(-0.2,0.4), col=c("gray50","slateblue4"), legend = c("mut","PU.1"),
names.arg=c(14,13,12,11,10,9,8,7,6,5,4,3,2,1) , beside=TRUE, args.legend =
list(x="topright",bty = "n",cex=1.5), cex.axis=1.5, cex.names=1.5, cex.lab=2, las=1,
horiz=TRUE)
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.ETS1Enrichment.stringent.stackedBarchart.R"
R < "${FIGURESDIR}/R/R.ETS1Enrichment.stringent.stackedBarchart.R" --no-save
```

```
cut -f 1,4,5,10 <(paste ${ANALYSISDIR}/ETS1Enrichment_PU1mut.Kmeans.14.lenient.txt
${ANALYSISDIR}/ETS1Enrichment_PU1.Kmeans.14.lenient.txt) >
${TMPDIR}/tmp.ETS1Enrichment.lenient.txt
cat >"${FIGURESDIR}/R/R.ETS1Enrichment.lenient.stackedBarchart.R" <<EOF
d <- read.table("${TMPDIR}/tmp.ETS1Enrichment.lenient.txt", sep="\t")
colnames(d)=c("Ann","total","PUmut","PU")
d\$percentPUmut <- (d\$PUmut / d\$total) * 100
d\$percentPU <- (d\$PU / d\$total) * 100
dr <- d[rev(rownames(d)),]
counts <- data.matrix(t(dr[,c(5,6)]))
pdf(file="${FIGURESDIR}/ETS1Enrichment.lenient.BarChartKmeans.pdf", height=5, width=3)
barplot(counts, xlab="% overlap ETS1/PU.1", xlim=c(0,100), ylab="cluster",
space=c(-0.2,0.4), col=c("gray50","slateblue4"), legend = c("mut","PU.1"),
names.arg=c(14,13,12,11,10,9,8,7,6,5,4,3,2,1) , beside=TRUE, args.legend =
list(x="topright",bty = "n",cex=1.5), cex.axis=1.5, cex.names=1.5, cex.lab=2, las=1,
horiz=TRUE)
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.ETS1Enrichment.lenient.stackedBarchart.R"
R < "${FIGURESDIR}/R/R.ETS1Enrichment.lenient.stackedBarchart.R" --no-save

cut -f 1,4,5,10 <(paste ${ANALYSISDIR}/FLI1Enrichment_PU1mut.Kmeans.14.stringent.txt
${ANALYSISDIR}/FLI1Enrichment_PU1.Kmeans.14.stringent.txt) >
${TMPDIR}/tmp.FLI1Enrichment.stringent.txt
cat >"${FIGURESDIR}/R/R.FLI1Enrichment.stringent.stackedBarchart.R" <<EOF
d <- read.table("${TMPDIR}/tmp.FLI1Enrichment.stringent.txt", sep="\t")
colnames(d)=c("Ann","total","PUmut","PU")
d\$percentPUmut <- (d\$PUmut / d\$total) * 100
d\$percentPU <- (d\$PU / d\$total) * 100
dr <- d[rev(rownames(d)),]
counts <- data.matrix(t(dr[,c(5,6)]))
pdf(file="${FIGURESDIR}/FLI1Enrichment.stringent.BarChartKmeans.pdf", height=5, width=3)
barplot(counts, xlab="% overlap FLI1/PU.1", xlim=c(0,100), ylab="cluster",
space=c(-0.2,0.4), col=c("gray50","magenta4"), legend = c("mut","PU.1"),
names.arg=c(14,13,12,11,10,9,8,7,6,5,4,3,2,1) , beside=TRUE, args.legend =
list(x="topright",bty = "n",cex=1.5), cex.axis=1.5, cex.names=1.5, cex.lab=2, las=1,
horiz=TRUE)
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.FLI1Enrichment.stringent.stackedBarchart.R"
R < "${FIGURESDIR}/R/R.FLI1Enrichment.stringent.stackedBarchart.R" --no-save

cut -f 1,4,5,10 <(paste ${ANALYSISDIR}/FLI1Enrichment_PU1mut.Kmeans.14.lenient.txt
${ANALYSISDIR}/FLI1Enrichment_PU1.Kmeans.14.lenient.txt) >
${TMPDIR}/tmp.FLI1Enrichment.lenient.txt
cat >"${FIGURESDIR}/R/R.FLI1Enrichment.lenient.stackedBarchart.R" <<EOF
d <- read.table("${TMPDIR}/tmp.FLI1Enrichment.lenient.txt", sep="\t")
colnames(d)=c("Ann","total","PUmut","PU")
d\$percentPUmut <- (d\$PUmut / d\$total) * 100
d\$percentPU <- (d\$PU / d\$total) * 100
dr <- d[rev(rownames(d)),]
counts <- data.matrix(t(dr[,c(5,6)]))
pdf(file="${FIGURESDIR}/FLI1Enrichment.lenient.BarChartKmeans.pdf", height=5, width=3)
barplot(counts, xlab="% overlap FLI1/PU.1", xlim=c(0,100), ylab="cluster",
space=c(-0.2,0.4), col=c("gray50","magenta4"), legend = c("mut","PU.1"),
names.arg=c(14,13,12,11,10,9,8,7,6,5,4,3,2,1) , beside=TRUE, args.legend =
list(x="topright",bty = "n",cex=1.5), cex.axis=1.5, cex.names=1.5, cex.lab=2, las=1,
horiz=TRUE)
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.FLI1Enrichment.lenient.stackedBarchart.R"
R < "${FIGURESDIR}/R/R.FLI1Enrichment.lenient.stackedBarchart.R" --no-save

# ETS peaks after transfection (including induced ones)
mergePeaks -d 100 ${PEAKDIR}/PU1.ntag15.filtered.12-50-bp.cluster.bound.pos.txt
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt -code -venn
${PEAKDIR}/cluster.bound_FLI1_ETS1_DSG_PU1_merged_overlap.venn.txt -prefix
${PEAKDIR}/cluster.bound.peaks_sep > ${TMPDIR}/tmp7.ann.txt
mv ${PEAKDIR}/cluster.bound.peaks_sep_100
${PEAKDIR}/cluster.bound.PU1-specific.stringent.pos.txt
cat ${PEAKDIR}/cluster.bound.peaks_sep_111 <(tail -n +2
${PEAKDIR}/cluster.bound.peaks_sep_110) <(tail -n +2 ${PEAKDIR}/cluster.bound.peaks_sep_101)
> ${PEAKDIR}/cluster.bound.PU1-ETS1-FLI1-shared.stringent.pos.txt
rm ${PEAKDIR}/cluster.bound.peaks_sep_*
mergePeaks -d 100 ${PEAKDIR}/PU1.ntag15.filtered.singleMotif.pos.txt
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt -code -venn
```

```
${PEAKDIR}/singleMotif_FLI1_ETS1_DSG_PU1_merged_overlap.venn.txt -prefix
${PEAKDIR}/singleMotif.peaks_sep > ${TMPDIR}/tmp8.ann.txt
mv ${PEAKDIR}/singleMotif.peaks_sep_100 ${PEAKDIR}/singleMotif.PU1-specific.stringent.pos.txt
cat ${PEAKDIR}/singleMotif.peaks_sep_111 <(tail -n +2 ${PEAKDIR}/singleMotif.peaks_sep_110)
<(tail -n +2 ${PEAKDIR}/singleMotif.peaks_sep_101) >
${PEAKDIR}/singleMotif.PU1-ETS1-FLI1-shared.stringent.pos.txt
rm ${PEAKDIR}/singleMotif.peaks_sep_*
mergePeaks -d 100 ${PEAKDIR}/PU1.ntag15.filtered.12-50-bp.cluster.bound.pos.txt
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.lenient.peaks.txt
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.lenient.peaks.txt -code -venn
${PEAKDIR}/cluster.bound_FLI1_ETS1_lenient_DSG_PU1_merged_overlap.venn.txt -prefix
${PEAKDIR}/cluster.bound.lenient.peaks_sep > ${TMPDIR}/tmp7.ann.txt
mv ${PEAKDIR}/cluster.bound.lenient.peaks_sep_100
${PEAKDIR}/cluster.bound.lenient.PU1-specific.stringent.pos.txt
cat ${PEAKDIR}/cluster.bound.lenient.peaks_sep_111 <(tail -n +2
${PEAKDIR}/cluster.bound.lenient.peaks_sep_110) <(tail -n +2
${PEAKDIR}/cluster.bound.lenient.peaks_sep_101) >
${PEAKDIR}/cluster.bound.lenient.PU1-ETS1-FLI1-shared.stringent.pos.txt
rm ${PEAKDIR}/cluster.bound.lenient.peaks_sep_*
mergePeaks -d 100 ${PEAKDIR}/PU1.ntag15.filtered.singleMotif.pos.txt
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.lenient.peaks.txt
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.lenient.peaks.txt -code -venn
${PEAKDIR}/singleMotif_FLI1_ETS1_lenient_DSG_PU1_merged_overlap.venn.txt -prefix
${PEAKDIR}/singleMotif.lenient.peaks_sep > ${TMPDIR}/tmp7.ann.txt
mv ${PEAKDIR}/singleMotif.lenient.peaks_sep_100
${PEAKDIR}/singleMotif.lenient.PU1-specific.stringent.pos.txt
cat ${PEAKDIR}/singleMotif.lenient.peaks_sep_111 <(tail -n +2
${PEAKDIR}/singleMotif.lenient.peaks_sep_110) <(tail -n +2
${PEAKDIR}/singleMotif.lenient.peaks_sep_101) >
${PEAKDIR}/singleMotif.lenient.PU1-ETS1-FLI1-shared.stringent.pos.txt
rm ${PEAKDIR}/singleMotif.lenient.peaks_sep_*

# ETS peaks without PU1 (includes competition or exchange)
mergePeaks -d 100 ${PEAKDIR}/PU1.ntag15.filtered.12-50-bp.cluster.bound.pos.txt
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt -code -venn
${PEAKDIR}/cluster.bound_FLI1_ETS1_DSG_PU1mut_merged_overlap.venn.txt -prefix
${PEAKDIR}/cluster.bound.peaksMut_sep > ${TMPDIR}/tmp7.ann.txt
mv ${PEAKDIR}/cluster.bound.peaksMut_sep_100
${PEAKDIR}/cluster.bound.PU1mut-specific.stringent.pos.txt
cat ${PEAKDIR}/cluster.bound.peaksMut_sep_111 <(tail -n +2
${PEAKDIR}/cluster.bound.peaksMut_sep_110) <(tail -n +2
${PEAKDIR}/cluster.bound.peaksMut_sep_101) >
${PEAKDIR}/cluster.bound.PU1mut-ETS1-FLI1-shared.stringent.pos.txt
rm ${PEAKDIR}/cluster.bound.peaksMut_sep_*
mergePeaks -d 100 ${PEAKDIR}/PU1.ntag15.filtered.singleMotif.pos.txt
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt -code -venn
${PEAKDIR}/singleMotif_FLI1_ETS1_DSG_PU1mut_merged_overlap.venn.txt -prefix
${PEAKDIR}/singleMotif.peaksMut_sep > ${TMPDIR}/tmp8.ann.txt
mv ${PEAKDIR}/singleMotif.peaksMut_sep_100
${PEAKDIR}/singleMotif.PU1mut-specific.stringent.pos.txt
cat ${PEAKDIR}/singleMotif.peaksMut_sep_111 <(tail -n +2
${PEAKDIR}/singleMotif.peaksMut_sep_110) <(tail -n +2
${PEAKDIR}/singleMotif.peaksMut_sep_101) >
${PEAKDIR}/singleMotif.PU1mut-ETS1-FLI1-shared.stringent.pos.txt
rm ${PEAKDIR}/singleMotif.peaksMut_sep_*
mergePeaks -d 100 ${PEAKDIR}/PU1.ntag15.filtered.12-50-bp.cluster.bound.pos.txt
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.lenient.peaks.txt
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.lenient.peaks.txt -code -venn
${PEAKDIR}/cluster.bound_FLI1_ETS1_lenient_DSG_PU1mut_merged_overlap.venn.txt -prefix
${PEAKDIR}/cluster.bound.lenient.peaksMut_sep > ${TMPDIR}/tmp7.ann.txt
mv ${PEAKDIR}/cluster.bound.lenient.peaksMut_sep_100
${PEAKDIR}/cluster.bound.lenient.PU1mut-specific.stringent.pos.txt
cat ${PEAKDIR}/cluster.bound.lenient.peaksMut_sep_111 <(tail -n +2
${PEAKDIR}/cluster.bound.lenient.peaksMut_sep_110) <(tail -n +2
${PEAKDIR}/cluster.bound.lenient.peaksMut_sep_101) >
${PEAKDIR}/cluster.bound.lenient.PU1mut-ETS1-FLI1-shared.stringent.pos.txt
rm ${PEAKDIR}/cluster.bound.lenient.peaksMut_sep_*
mergePeaks -d 100 ${PEAKDIR}/PU1.ntag15.filtered.singleMotif.pos.txt
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.lenient.peaks.txt
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.lenient.peaks.txt -code -venn
${PEAKDIR}/singleMotif_FLI1_ETS1_lenient_DSG_PU1mut_merged_overlap.venn.txt -prefix
${PEAKDIR}/singleMotif.lenient.peaksMut_sep > ${TMPDIR}/tmp7.ann.txt
mv ${PEAKDIR}/singleMotif.lenient.peaksMut_sep_100
${PEAKDIR}/singleMotif.lenient.PU1mut-specific.stringent.pos.txt
cat ${PEAKDIR}/singleMotif.lenient.peaksMut_sep_111 <(tail -n +2
${PEAKDIR}/singleMotif.lenient.peaksMut_sep_110) <(tail -n +2
${PEAKDIR}/singleMotif.lenient.peaksMut_sep_101) >
${PEAKDIR}/singleMotif.lenient.PU1mut-ETS1-FLI1-shared.stringent.pos.txt
rm ${PEAKDIR}/singleMotif.lenient.peaksMut_sep_*

# stringent ETS peaks overlapping with lenient PU1 peaks (class 1 ETS-specific)
mergePeaks -d 100 ${PEAKDIR}/PU1_Flag_merged.factor.lenient.peaks.txt
${PEAKDIR}/FLI1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1mut_merged.factor.fdr05.ntag15.filtered.pos.txt -code -venn
${PEAKDIR}/FLI1_ETS1_DSG_PU1mut_merged_overlap_PU1lenient.venn.txt -prefix
${PEAKDIR}/PU.bound.peaksLen_sep > ${TMPDIR}/tmp7.ann.txt
cat ${PEAKDIR}/PU.bound.peaksLen_sep_011 <(tail -n +2 ${PEAKDIR}/PU.bound.peaksLen_sep_010)
<(tail -n +2 ${PEAKDIR}/PU.bound.peaksLen_sep_001) >
${PEAKDIR}/ETS1-FLI1-shared.specific.stringent.mutPU1.pos.txt
rm ${PEAKDIR}/PU.bound.peaksLen_sep_*
mergePeaks -d 100 ${PEAKDIR}/PU1_Flag_merged.factor.lenient.peaks.txt
${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt
${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.pos.txt -code -venn
${PEAKDIR}/FLI1_ETS1_DSG_PU1_merged_overlap_PU1lenient.venn.txt -prefix
${PEAKDIR}/PU.bound.peaksLen_sep > ${TMPDIR}/tmp7.ann.txt
cat ${PEAKDIR}/PU.bound.peaksLen_sep_011 <(tail -n +2 ${PEAKDIR}/PU.bound.peaksLen_sep_010)
<(tail -n +2 ${PEAKDIR}/PU.bound.peaksLen_sep_001) >
${PEAKDIR}/ETS1-FLI1-shared.specific.stringent.PU1.pos.txt
rm ${PEAKDIR}/PU.bound.peaksLen_sep_*

# overlap between ETS1-FLI1-shared
mergePeaks -d 100 ${PEAKDIR}/ETS1-FLI1-shared.specific.stringent.mutPU1.pos.txt
${PEAKDIR}/ETS1-FLI1-shared.specific.stringent.PU1.pos.txt -code -venn
${PEAKDIR}/FLI1_ETS1_DSG_shared.specific_overlap.venn.txt >
${PEAKDIR}/FLI1_ETS1_DSG_shared.specific_overlap.pos.txt

# peak sets for further analysis
ETSSPECIFIC_PU1="${PEAKDIR}/ETS1-FLI1-shared.specific.stringent.PU1.pos.txt"
ETSSPECIFIC_PU1mut="${PEAKDIR}/ETS1-FLI1-shared.specific.stringent.mutPU1.pos.txt"
CLUSTERSPECIFIC_PU1="${PEAKDIR}/cluster.bound.lenient.PU1-specific.stringent.pos.txt"
CLUSTERSPECIFIC_PU1mut="${PEAKDIR}/cluster.bound.lenient.PU1mut-specific.stringent.pos.txt"
CLUSTERSHARED_PU1="${PEAKDIR}/cluster.bound.PU1-ETS1-FLI1-shared.stringent.pos.txt"
CLUSTERSHARED_PU1mut="${PEAKDIR}/cluster.bound.PU1mut-ETS1-FLI1-shared.stringent.pos.txt"
SINGLESPECIFIC_PU1="${PEAKDIR}/singleMotif.lenient.PU1-specific.stringent.pos.txt"
SINGLESPECIFIC_PU1mut="${PEAKDIR}/singleMotif.lenient.PU1mut-specific.stringent.pos.txt"
SINGLESHARED_PU1="${PEAKDIR}/singleMotif.PU1-ETS1-FLI1-shared.stringent.pos.txt"
SINGLESHARED_PU1mut="${PEAKDIR}/singleMotif.PU1mut-ETS1-FLI1-shared.stringent.pos.txt"
SINGLESPECIFIC_S_PU1="${PEAKDIR}/singleMotif.lenient.PU1-specific.stringent.single.txt"
SINGLESPECIFIC_S_PU1mut="${PEAKDIR}/singleMotif.lenient.PU1mut-specific.stringent.single.txt"
SINGLESHARED_S_PU1="${PEAKDIR}/singleMotif.PU1-ETS1-FLI1-shared.stringent.single.txt"
SINGLESHARED_S_PU1mut="${PEAKDIR}/singleMotif.PU1mut-ETS1-FLI1-shared.stringent.single.txt"
SINGLESPECIFIC_P_PU1="${PEAKDIR}/singleMotif.lenient.PU1-specific.stringent.paired.txt"
SINGLESPECIFIC_P_PU1mut="${PEAKDIR}/singleMotif.lenient.PU1mut-specific.stringent.paired.txt"
SINGLESHARED_P_PU1="${PEAKDIR}/singleMotif.PU1-ETS1-FLI1-shared.stringent.paired.txt"
SINGLESHARED_P_PU1mut="${PEAKDIR}/singleMotif.PU1mut-ETS1-FLI1-shared.stringent.paired.txt"

# single specific sites: divide into overlapping and separate ETS motifs
annotatePeaks.pl ${SINGLESHARED_PU1} hg19 -size 200 -m ${CLASS1ETSMOTIF} -fm ${PU1MOTIF}
-nogene -noann -nmotifs >
${PEAKDIR}/singleMotif.PU1-ETS1-FLI1-shared.stringent.ETSmotifann.txt
myFilterFile.pl ${PEAKDIR}/singleMotif.PU1-ETS1-FLI1-shared.stringent.ETSmotifann.txt
-column 11 -upperlimit 0 > ${PEAKDIR}/singleMotif.PU1-ETS1-FLI1-shared.stringent.single.txt
myFilterFile.pl ${PEAKDIR}/singleMotif.PU1-ETS1-FLI1-shared.stringent.ETSmotifann.txt
-column 11 -lowerlimit 1 > ${PEAKDIR}/singleMotif.PU1-ETS1-FLI1-shared.stringent.paired.txt
annotatePeaks.pl ${SINGLESHARED_PU1mut} hg19 -size 200 -m ${CLASS1ETSMOTIF} -fm ${PU1MOTIF}
-nogene -noann -nmotifs >
${PEAKDIR}/singleMotif.PU1mut-ETS1-FLI1-shared.stringent.ETSmotifann.txt
myFilterFile.pl ${PEAKDIR}/singleMotif.PU1mut-ETS1-FLI1-shared.stringent.ETSmotifann.txt
-column 11 -upperlimit 0 > ${PEAKDIR}/singleMotif.PU1mut-ETS1-FLI1-shared.stringent.single.txt
myFilterFile.pl ${PEAKDIR}/singleMotif.PU1mut-ETS1-FLI1-shared.stringent.ETSmotifann.txt
-column 11 -lowerlimit 1 > ${PEAKDIR}/singleMotif.PU1mut-ETS1-FLI1-shared.stringent.paired.txt
annotatePeaks.pl ${SINGLESPECIFIC_S_PU1} hg19 -size 200 -m ${CLASS1ETSMOTIF} -fm ${PU1MOTIF}
-nogene -noann -nmotifs >
${PEAKDIR}/singleMotif.lenient.PU1-specific.stringent.ETSmotifann.txt
myFilterFile.pl ${PEAKDIR}/singleMotif.lenient.PU1-specific.stringent.ETSmotifann.txt
-column 11 -upperlimit 0 > ${PEAKDIR}/singleMotif.lenient.PU1-specific.stringent.single.txt
myFilterFile.pl ${PEAKDIR}/singleMotif.lenient.PU1-specific.stringent.ETSmotifann.txt
-column 11 -lowerlimit 1 > ${PEAKDIR}/singleMotif.lenient.PU1-specific.stringent.paired.txt
annotatePeaks.pl ${SINGLESPECIFIC_S_PU1mut} hg19 -size 200 -m ${CLASS1ETSMOTIF} -fm
${PU1MOTIF} -nogene -noann -nmotifs >
${PEAKDIR}/singleMotif.lenient.PU1mut-specific.stringent.ETSmotifann.txt
myFilterFile.pl ${PEAKDIR}/singleMotif.lenient.PU1mut-specific.stringent.ETSmotifann.txt
-column 11 -upperlimit 0 > ${PEAKDIR}/singleMotif.lenient.PU1mut-specific.stringent.single.txt
myFilterFile.pl ${PEAKDIR}/singleMotif.lenient.PU1mut-specific.stringent.ETSmotifann.txt
```

```
-column 11 -lowerlimit 1 > ${PEAKDIR}/singleMotif.lenient.PU1mut-specific.stringent.paired.txt

# motifscore distribution for unseparated files
cd ${ANALYSISDIR}
declare -a PEAKSETS=("${SINGLESPECIFIC_PU1}" "${SINGLESHARED_PU1}" "${CLUSTERSPECIFIC_PU1}"
"${CLUSTERSHARED_PU1}" "${ETSSPECIFIC_PU1}" "${SINGLESPECIFIC_PU1mut}"
"${SINGLESHARED_PU1mut}" "${CLUSTERSPECIFIC_PU1mut}" "${CLUSTERSHARED_PU1mut}"
"${ETSSPECIFIC_PU1mut}")
declare -a NAMES=(SINGLESPECIFIC_PU1 SINGLESHARED_PU1 CLUSTERSPECIFIC_PU1 CLUSTERSHARED_PU1
ETSSPECIFIC_PU1 SINGLESPECIFIC_PU1mut SINGLESHARED_PU1mut CLUSTERSPECIFIC_PU1mut
CLUSTERSHARED_PU1mut ETSSPECIFIC_PU1mut)

COUNT=0
_DATE=$(date +%s)
for PEAKSET in ${PEAKSETS[@]}; do
NAME="${NAMES[${COUNT}]}"
cat >"${TMPDIR}/ghist.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
annotatePeaks.pl ${PEAKSET} hg19 -size 200 -m ${PU1MOTIF} ${CLASS1ETSMOTIF} -mscore -nogene
-noann > ${TMPDIR}/tmp.${NAME}.scoreAnn.txt
EOF
COUNT=$((COUNT+=1))
chmod 750 "${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
echo "plotting ghists for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#--------------------------------------------
for NAME in ${NAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

# combined bean- and box-plots
# PU1 motif in PU1 transfected
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.SINGLESPECIFIC_PU1.scoreAnn.txt", header=T, sep="\t")
data2 <- read.table("${TMPDIR}/tmp.SINGLESHARED_PU1.scoreAnn.txt", header=T, sep="\t")
data3 <- read.table("${TMPDIR}/tmp.CLUSTERSPECIFIC_PU1.scoreAnn.txt", header=T, sep="\t")
data4 <- read.table("${TMPDIR}/tmp.CLUSTERSHARED_PU1.scoreAnn.txt", header=T, sep="\t")
data5 <- read.table("${TMPDIR}/tmp.ETSSPECIFIC_PU1.scoreAnn.txt", header=T, sep="\t")
# combining the data columns (bit complicated, because each column has different length)
a <- data1[-1,10]
b <- data2[-1,10]
c <- data3[-1,10]
d <- data4[-1,10]
e <- data5[-1,10]
z <- c("a", "b", "c", "d", "e")
x <- lapply(z, get, envir=environment())
names(x) <- z
# define labels
laba <- "single PU.1 specific"
labb <- "single PU.1 shared"
labc <- "paired PU.1 specific"
labd <- "paired PU.1 shared"
labe <- "ETS specific"
# defining colors
beancol <- "${colPU1}"
boxcol <- "gray50"
```

```
pdf(file="${FIGURESDIR}/PU1motifScores.PU1.ETScomparison.pdf", height=5, width=3)
par(mar=c(7.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(0,12),lwd=0.5)
# adding box plot on top
par(mar=c(7.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(0,12),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:5, labels=c(laba, labb, labc, labd,
labe),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext("PU.1 motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.motifbean.${_DATE}.R"
R < ${TMPDIR}/R.motifbean.${_DATE}.R --no-save
rm ${TMPDIR}/R.motifbean.${_DATE}.R

# ETS motif in PU1 transfected
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.SINGLESPECIFIC_PU1.scoreAnn.txt", header=T, sep="\t")
data2 <- read.table("${TMPDIR}/tmp.SINGLESHARED_PU1.scoreAnn.txt", header=T, sep="\t")
data3 <- read.table("${TMPDIR}/tmp.CLUSTERSPECIFIC_PU1.scoreAnn.txt", header=T, sep="\t")
data4 <- read.table("${TMPDIR}/tmp.CLUSTERSHARED_PU1.scoreAnn.txt", header=T, sep="\t")
data5 <- read.table("${TMPDIR}/tmp.ETSSPECIFIC_PU1.scoreAnn.txt", header=T, sep="\t")
# combining the data columns (bit complicated, because each column has different length)
a <- data1[-1,11]
b <- data2[-1,11]
c <- data3[-1,11]
d <- data4[-1,11]
e <- data5[-1,11]
z <- c("a", "b", "c", "d", "e")
x <- lapply(z, get, envir=environment())
names(x) <- z
# define labels
laba <- "single PU.1 specific"
labb <- "single PU.1 shared"
labc <- "paired PU.1 specific"
labd <- "paired PU.1 shared"
labe <- "ETS specific"
# defining colors
beancol <- "${colETS1}"
boxcol <- "gray50"
pdf(file="${FIGURESDIR}/ETSmotifScores.PU1.ETScomparison.pdf", height=5, width=3)
par(mar=c(7.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(0,11),lwd=0.5)
# adding box plot on top
par(mar=c(7.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(0,11),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:5, labels=c(laba, labb, labc, labd,
labe),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext("ETS class1 motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
abline(h=6.738572,col="darkblue",lty=3,lwd=1)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.motifbean.${_DATE}.R"
R < ${TMPDIR}/R.motifbean.${_DATE}.R --no-save
rm ${TMPDIR}/R.motifbean.${_DATE}.R

# PU1 motif in PU1mut transfected
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
library(beanplot)
```

```
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.SINGLESPECIFIC_PU1mut.scoreAnn.txt", header=T, sep="\t")
data2 <- read.table("${TMPDIR}/tmp.SINGLESHARED_PU1mut.scoreAnn.txt", header=T, sep="\t")
data3 <- read.table("${TMPDIR}/tmp.CLUSTERSPECIFIC_PU1mut.scoreAnn.txt", header=T, sep="\t")
data4 <- read.table("${TMPDIR}/tmp.CLUSTERSHARED_PU1mut.scoreAnn.txt", header=T, sep="\t")
data5 <- read.table("${TMPDIR}/tmp.ETSSPECIFIC_PU1mut.scoreAnn.txt", header=T, sep="\t")
# combining the data columns (bit complicated, because each column has different length)
a <- data1[-1,10]
b <- data2[-1,10]
c <- data3[-1,10]
d <- data4[-1,10]
e <- data5[-1,10]
z <- c("a", "b", "c", "d", "e")
x <- lapply(z, get, envir=environment())
names(x) <- z
# define labels
laba <- "single PU.1 specific"
labb <- "single PU.1 shared"
labc <- "paired PU.1 specific"
labd <- "paired PU.1 shared"
labe <- "ETS specific"
# defining colors
beancol <- "${colPU1}"
boxcol <- "gray50"
pdf(file="${FIGURESDIR}/PU1motifScores.PU1mut.ETScomparison.pdf", height=5, width=3)
par(mar=c(7.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(0,12),lwd=0.5)
# adding box plot on top
par(mar=c(7.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(0,12),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:5, labels=c(laba, labb, labc, labd,
labe),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext("PU.1 motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.motifbean.${_DATE}.R"
R < ${TMPDIR}/R.motifbean.${_DATE}.R --no-save
rm ${TMPDIR}/R.motifbean.${_DATE}.R

# ETS motif in PU1mut transfected
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.SINGLESPECIFIC_PU1mut.scoreAnn.txt", header=T, sep="\t")
data2 <- read.table("${TMPDIR}/tmp.SINGLESHARED_PU1mut.scoreAnn.txt", header=T, sep="\t")
data3 <- read.table("${TMPDIR}/tmp.CLUSTERSPECIFIC_PU1mut.scoreAnn.txt", header=T, sep="\t")
data4 <- read.table("${TMPDIR}/tmp.CLUSTERSHARED_PU1mut.scoreAnn.txt", header=T, sep="\t")
data5 <- read.table("${TMPDIR}/tmp.ETSSPECIFIC_PU1mut.scoreAnn.txt", header=T, sep="\t")
# combining the data columns (bit complicated, because each column has different length)
a <- data1[-1,11]
b <- data2[-1,11]
c <- data3[-1,11]
d <- data4[-1,11]
e <- data5[-1,11]
z <- c("a", "b", "c", "d", "e")
x <- lapply(z, get, envir=environment())
names(x) <- z
# define labels
laba <- "single PU.1 specific"
labb <- "single PU.1 shared"
labc <- "paired PU.1 specific"
labd <- "paired PU.1 shared"
labe <- "ETS specific"
# defining colors
beancol <- "${colETS1}"
boxcol <- "gray50"
pdf(file="${FIGURESDIR}/ETSmotifScores.PU1mut.ETScomparison.pdf", height=5, width=3)
par(mar=c(7.5,2.5,1,1))
# plotting the beans
```

```
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(0,11),lwd=0.5)
# adding box plot on top
par(mar=c(7.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(0,11),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:5, labels=c(laba, labb, labc, labd,
labe),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext("ETS class1 motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
abline(h=6.738572,col="darkblue",lty=3,lwd=1)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.motifbean.${_DATE}.R"
R < ${TMPDIR}/R.motifbean.${_DATE}.R --no-save
rm ${TMPDIR}/R.motifbean.${_DATE}.R

# motifscore distribution for unseparated files
cd ${ANALYSISDIR}
declare -a PEAKSETS=("${SINGLESPECIFIC_S_PU1}" "${SINGLESHARED_S_PU1}"
"${SINGLESPECIFIC_S_PU1mut}" "${SINGLESHARED_S_PU1mut}" "${SINGLESPECIFIC_P_PU1}"
"${SINGLESHARED_P_PU1}" "${SINGLESPECIFIC_P_PU1mut}" "${SINGLESHARED_P_PU1mut}")
declare -a NAMES=(SINGLESPECIFIC_S_PU1 SINGLESHARED_S_PU1 SINGLESPECIFIC_S_PU1mut
SINGLESHARED_S_PU1mut SINGLESPECIFIC_P_PU1 SINGLESHARED_P_PU1 SINGLESPECIFIC_P_PU1mut
SINGLESHARED_P_PU1mut)
COUNT=0
_DATE=$(date +%s)
for PEAKSET in ${PEAKSETS[@]}; do
NAME="${NAMES[${COUNT}]}"
cat >"${TMPDIR}/ghist.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
annotatePeaks.pl ${PEAKSET} hg19 -size 200 -m ${PU1MOTIF} ${CLASS1ETSMOTIF} -mscore -nogene
-noann > ${TMPDIR}/tmp.${NAME}.scoreAnn.txt
EOF
COUNT=$((COUNT+=1))
chmod 750 "${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
echo "plotting ghists for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#---------------------------------------------
for NAME in ${NAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------

# combined bean- and box-plots

# PU1 motif in PU1 transfected
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.SINGLESPECIFIC_S_PU1.scoreAnn.txt", header=T, sep="\t")
data2 <- read.table("${TMPDIR}/tmp.SINGLESPECIFIC_P_PU1.scoreAnn.txt", header=T, sep="\t")
data3 <- read.table("${TMPDIR}/tmp.SINGLESHARED_S_PU1.scoreAnn.txt", header=T, sep="\t")
data4 <- read.table("${TMPDIR}/tmp.SINGLESHARED_P_PU1.scoreAnn.txt", header=T, sep="\t")
# combining the data columns (bit complicated, because each column has different length)
a <- data1[-1,10]
b <- data2[-1,10]
```

```
c <- data3[-1,10]
d <- data4[-1,10]
z <- c("a", "b", "c", "d")
x <- lapply(z, get, envir=environment())
names(x) <- z
# define labels
laba <- "single specific"
labb <- "single shared"
labc <- "paired specific"
labd <- "paired shared"
# defining colors
beancol <- "${colPU1}"
boxcol <- "gray50"
pdf(file="${FIGURESDIR}/PU1motifScores.PU1.ETScomparison.singleSplit.pdf", height=5,
width=2.5)
par(mar=c(7.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(0,12),lwd=0.5)
# adding box plot on top
par(mar=c(7.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(0,12),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:4, labels=c(laba, labb, labc,
labd),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext(" PU.1 motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.motifbean.${_DATE}.R"
R < ${TMPDIR}/R.motifbean.${_DATE}.R --no-save
rm ${TMPDIR}/R.motifbean.${_DATE}.R

# PU1 motif in PU1mut transfected
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"${TMPDIR}/R.motifbean.${_DATE}.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.SINGLESPECIFIC_S_PU1mut.scoreAnn.txt", header=T, sep="\t")
data2 <- read.table("${TMPDIR}/tmp.SINGLESPECIFIC_P_PU1mut.scoreAnn.txt", header=T, sep="\t")
data3 <- read.table("${TMPDIR}/tmp.SINGLESHARED_S_PU1mut.scoreAnn.txt", header=T, sep="\t")
data4 <- read.table("${TMPDIR}/tmp.SINGLESHARED_P_PU1mut.scoreAnn.txt", header=T, sep="\t")
# combining the data columns (bit complicated, because each column has different length)
a <- data1[-1,10]
b <- data2[-1,10]
c <- data3[-1,10]
d <- data4[-1,10]
z <- c("a", "b", "c", "d")
x <- lapply(z, get, envir=environment())
names(x) <- z
# define labels
laba <- "single specific"
labb <- "single shared"
labc <- "paired specific"
labd <- "paired shared"
# defining colors
beancol <- "${colPU1}"
boxcol <- "gray50"
pdf(file="${FIGURESDIR}/PU1motifScores.PU1mut.ETScomparison.singleSplit.pdf", height=5,
width=2.5)
par(mar=c(7.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(0,12),lwd=0.5)
# adding box plot on top
par(mar=c(7.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(0,12),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:4, labels=c(laba, labb, labc,
labd),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext(" PU.1 motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
```

```
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.motifbean.${_DATE}.R"
R < ${TMPDIR}/R.motifbean.${_DATE}.R --no-save
rm ${TMPDIR}/R.motifbean.${_DATE}.R

# motif analyses for all peak sets
cd ${ANALYSISDIR}
declare -a PEAKSETS=("${SINGLESPECIFIC_PU1}" "${SINGLESHARED_PU1}" "${CLUSTERSPECIFIC_PU1}"
"${CLUSTERSHARED_PU1}" "${ETSSPECIFIC_PU1}" "${SINGLESPECIFIC_PU1mut}"
"${SINGLESHARED_PU1mut}" "${CLUSTERSPECIFIC_PU1mut}" "${CLUSTERSHARED_PU1mut}"
"${ETSSPECIFIC_PU1mut}" "${SINGLESPECIFIC_S_PU1}" "${SINGLESHARED_S_PU1}"
"${SINGLESPECIFIC_S_PU1mut}" "${SINGLESHARED_S_PU1mut}" "${SINGLESPECIFIC_P_PU1}"
"${SINGLESHARED_P_PU1}" "${SINGLESPECIFIC_P_PU1mut}" "${SINGLESHARED_P_PU1mut}")
declare -a NAMES=(SINGLESPECIFIC_PU1 SINGLESHARED_PU1 CLUSTERSPECIFIC_PU1 CLUSTERSHARED_PU1
ETSSPECIFIC_PU1 SINGLESPECIFIC_PU1mut SINGLESHARED_PU1mut CLUSTERSPECIFIC_PU1mut
CLUSTERSHARED_PU1mut ETSSPECIFIC_PU1mut SINGLESPECIFIC_S_PU1 SINGLESHARED_S_PU1
SINGLESPECIFIC_S_PU1mut SINGLESHARED_S_PU1mut SINGLESPECIFIC_P_PU1 SINGLESHARED_P_PU1
SINGLESPECIFIC_P_PU1mut SINGLESHARED_P_PU1mut)

COUNT=0
_DATE=$(date +%s)
for PEAKSET in ${PEAKSETS[@]}; do
NAME="${NAMES[${COUNT}]}"
cat >"${TMPDIR}/motifs.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findMotifsGenome.pl ${PEAKSET} hg19 "${MOTIFDIR}/${NAME}" -size 200 -len
7,8,9,10,11,12,13,14 -p 2 -h
compareMotifs.pl "${MOTIFDIR}/${NAME}/homerMotifs.all.motifs" "${MOTIFDIR}/${NAME}/final"
-reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 2
EOF
COUNT=$((COUNT+=1))
chmod 750 "${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
echo "searching motifs for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#-----------------------------------------------
for NAME in ${NAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#-----------------------------------------------

# motif analyses with PU.1 motif masked
cd ${ANALYSISDIR}
declare -a PEAKSETS=("${SINGLESPECIFIC_PU1}" "${SINGLESHARED_PU1}" "${CLUSTERSPECIFIC_PU1}"
"${CLUSTERSHARED_PU1}")
declare -a NAMES=(SINGLESPECIFIC_PU1 SINGLESHARED_PU1 CLUSTERSPECIFIC_PU1 CLUSTERSHARED_PU1)

COUNT=0
_DATE=$(date +%s)
for PEAKSET in ${PEAKSETS[@]}; do
NAME="${NAMES[${COUNT}]}"
cat >"${TMPDIR}/motifs.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findMotifsGenome.pl ${PEAKSET} hg19 "${MOTIFDIR}/${NAME}_PU1masked" -maskMotif ${PU1MOTIF}
```

```
-size 200 -len 7,8,9,10,11,12,13,14 -p 2 -h
compareMotifs.pl "${MOTIFDIR}/${NAME}_PU1masked/homerMotifs.all.motifs"
"${MOTIFDIR}/${NAME}_PU1masked/final" -reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info
1.5 -cpu 2
EOF
COUNT=$((COUNT+=1))
chmod 750 "${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
echo "searching motifs for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#---------------------------------------------
for NAME in ${NAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------

# motif analyses for peak sets that do not overlap with ETS factors but contain the motif
cd ${ANALYSISDIR}
declare -a PEAKSETS=("${SINGLESPECIFIC_P_PU1}" "${SINGLESHARED_P_PU1}"
"${SINGLESPECIFIC_P_PU1mut}" "${SINGLESHARED_P_PU1mut}")
declare -a NAMES=(SINGLESPECIFIC_P_PU1 SINGLESHARED_P_PU1 SINGLESPECIFIC_P_PU1mut
SINGLESHARED_P_PU1mut)

COUNT=0
_DATE=$(date +%s)
for PEAKSET in ${PEAKSETS[@]}; do
NAME="${NAMES[${COUNT}]}"
cat >"${TMPDIR}/motifs.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findMotifsGenome.pl ${PEAKSET} hg19 "${MOTIFDIR}/${NAME}_PU1masked" -maskMotif ${PU1MOTIF}
-size 200 -len 7,8,9,10,11,12,13,14 -p 2 -h
compareMotifs.pl "${MOTIFDIR}/${NAME}_PU1masked/homerMotifs.all.motifs"
"${MOTIFDIR}/${NAME}_PU1masked/final" -reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info
1.5 -cpu 2
EOF
COUNT=$((COUNT+=1))
chmod 750 "${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
echo "searching motifs for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#---------------------------------------------
for NAME in ${NAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------
```

```
# motif analyses for peak sets that do not overlap with ETS factors and do not contain the
motif
cd ${ANALYSISDIR}
declare -a PEAKSETS=("${SINGLESPECIFIC_S_PU1}" "${SINGLESHARED_S_PU1}"
"${SINGLESPECIFIC_S_PU1mut}" "${SINGLESHARED_S_PU1mut}")
declare -a NAMES=(SINGLESPECIFIC_S_PU1 SINGLESHARED_S_PU1 SINGLESPECIFIC_S_PU1mut
SINGLESHARED_S_PU1mut)

COUNT=0
_DATE=$(date +%s)
for PEAKSET in ${PEAKSETS[@]}; do
NAME="${NAMES[${COUNT}]}"
cat >"${TMPDIR}/motifs.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
findMotifsGenome.pl ${PEAKSET} hg19 "${MOTIFDIR}/${NAME}_PU1masked" -maskMotif ${PU1MOTIF}
-size 200 -len 7,8,9,10,11,12,13,14 -p 2 -h
compareMotifs.pl "${MOTIFDIR}/${NAME}_PU1masked/homerMotifs.all.motifs"
"${MOTIFDIR}/${NAME}_PU1masked/final" -reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info
1.5 -cpu 2
EOF
COUNT=$((COUNT+=1))
chmod 750 "${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
echo "searching motifs for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/motifs.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#---------------------------------------------
for NAME in ${NAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------

# histogram plots for all peaksets
# ghist plots including relevant data sets
cd ${ANALYSISDIR}
declare -a PEAKSETS=("${SINGLESPECIFIC_PU1}" "${SINGLESHARED_PU1}" "${CLUSTERSPECIFIC_PU1}"
"${CLUSTERSHARED_PU1}" "${ETSSPECIFIC_PU1}" "${SINGLESPECIFIC_PU1mut}"
"${SINGLESHARED_PU1mut}" "${CLUSTERSPECIFIC_PU1mut}" "${CLUSTERSHARED_PU1mut}"
"${ETSSPECIFIC_PU1mut}" "${SINGLESPECIFIC_S_PU1}" "${SINGLESHARED_S_PU1}"
"${SINGLESPECIFIC_S_PU1mut}" "${SINGLESHARED_S_PU1mut}" "${SINGLESPECIFIC_P_PU1}"
"${SINGLESHARED_P_PU1}" "${SINGLESPECIFIC_P_PU1mut}" "${SINGLESHARED_P_PU1mut}")
declare -a PEAKNAMES=(SINGLESPECIFIC_PU1 SINGLESHARED_PU1 CLUSTERSPECIFIC_PU1
CLUSTERSHARED_PU1 ETSSPECIFIC_PU1 SINGLESPECIFIC_PU1mut SINGLESHARED_PU1mut
CLUSTERSPECIFIC_PU1mut CLUSTERSHARED_PU1mut ETSSPECIFIC_PU1mut SINGLESPECIFIC_S_PU1
SINGLESHARED_S_PU1 SINGLESPECIFIC_S_PU1mut SINGLESHARED_S_PU1mut SINGLESPECIFIC_P_PU1
SINGLESHARED_P_PU1 SINGLESPECIFIC_P_PU1mut SINGLESHARED_P_PU1mut)
declare -a SETS=("${TAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged" \
"${ATACTAGDIR}/CTV1_PU1mut_CNVnormRefChr_merged"
"${ATACTAGDIR}/CTV1_PU1_CNVnormRefChr_merged" \
"${TAGDIR}/CTV1_PU1mut_DSG_ETS1_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_DSG_ETS1_CNVnormRefChr_merged" \
"${TAGDIR}/CTV1_PU1mut_DSG_FLI1_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_DSG_FLI1_CNVnormRefChr_merged" \
"${TAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged.4samples"
"${TAGDIR}/CTV1_PU1_Flag_3ug_CNVnormRefChr" \
"${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr")
declare -a NAMES=(PU1mut PU1 ATACmut ATACPU1 ETS1-PU1mut ETS1-PU1 FLI1-PU1mut FLI1-PU1
100perc 50perc 15perc)
```

```
COUNT=0
_DATE=$(date +%s)
for PEAKSET in ${PEAKSETS[@]}; do
PEAKNAME="${PEAKNAMES[${COUNT}]}"
echo $PEAKNAME
cat >"${TMPDIR}/ghist.${PEAKNAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
EOF
COUNT2=0
for SET in ${SETS[@]}; do
NAME="${NAMES[${COUNT2}]}"
cat >>"${TMPDIR}/ghist.${PEAKNAME}.${_DATE}.sh" <<EOF
annotatePeaks.pl ${PEAKSET} hg19 -size 2000 -hist 25 -ghist -d ${SET} >
${TMPDIR}/tmp.ghist.${PEAKNAME}.${NAME}.txt
EOF
COUNT2=$((COUNT2+=1))
done
chmod 750 "${TMPDIR}/ghist.${PEAKNAME}.${_DATE}.sh"
echo "plotting ghists for ${PEAKNAME}"
screen -dm -S ${PEAKNAME} bash -c "bash ${TMPDIR}/ghist.${PEAKNAME}.${_DATE}.sh"
COUNT=$((COUNT+=1))
done

# loop to check when screen sessions are done
#---------------------------------------------
for PEAKNAME in ${PEAKNAMES[@]}; do
while [ true ]; do # Endless loop.
pid=`screen -S ${PEAKNAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${PEAKNAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------

cd ${ANALYSISDIR}
declare -a PEAKSETS=("${SINGLESPECIFIC_PU1}" "${SINGLESHARED_PU1}" "${CLUSTERSPECIFIC_PU1}"
"${CLUSTERSHARED_PU1}" "${SINGLESPECIFIC_PU1mut}" "${SINGLESHARED_PU1mut}"
"${CLUSTERSPECIFIC_PU1mut}" "${CLUSTERSHARED_PU1mut}" "${SINGLESPECIFIC_S_PU1}"
"${SINGLESHARED_S_PU1}" "${SINGLESPECIFIC_S_PU1mut}" "${SINGLESHARED_S_PU1mut}"
"${SINGLESPECIFIC_P_PU1}" "${SINGLESHARED_P_PU1}" "${SINGLESPECIFIC_P_PU1mut}"
"${SINGLESHARED_P_PU1mut}")
declare -a PEAKNAMES=(SINGLESPECIFIC_PU1 SINGLESHARED_PU1 CLUSTERSPECIFIC_PU1
CLUSTERSHARED_PU1 SINGLESPECIFIC_PU1mut SINGLESHARED_PU1mut CLUSTERSPECIFIC_PU1mut
CLUSTERSHARED_PU1mut SINGLESPECIFIC_S_PU1 SINGLESHARED_S_PU1 SINGLESPECIFIC_S_PU1mut
SINGLESHARED_S_PU1mut SINGLESPECIFIC_P_PU1 SINGLESHARED_P_PU1 SINGLESPECIFIC_P_PU1mut
SINGLESHARED_P_PU1mut)

COUNT=0
_DATE=$(date +%s)
for PEAKSET in ${PEAKSETS[@]}; do
PEAKNAME="${PEAKNAMES[${COUNT}]}"
plotHIST.sh \
-f "${TMPDIR}/tmp.ghist.${PEAKNAME}.PU1mut.txt ${TMPDIR}/tmp.ghist.${PEAKNAME}.PU1.txt
${TMPDIR}/tmp.ghist.${PEAKNAME}.ETS1-PU1mut.txt
${TMPDIR}/tmp.ghist.${PEAKNAME}.ETS1-PU1.txt" \
-s "mutPU1 PU1 ETS1-mutPU1 ETS1-PU1" -c "${colMUT} ${colPU1} ${colMUTETS1} ${colETS1}" -x
1000 -y "0 40" -d ${FIGURESDIR}/hist -n ETS1_${PEAKNAME}
plotHIST.sh \
-f "${TMPDIR}/tmp.ghist.${PEAKNAME}.PU1mut.txt ${TMPDIR}/tmp.ghist.${PEAKNAME}.PU1.txt
${TMPDIR}/tmp.ghist.${PEAKNAME}.FLI1-PU1mut.txt
${TMPDIR}/tmp.ghist.${PEAKNAME}.FLI1-PU1.txt" \
-s "mutPU1 PU1 FLI1-mutPU1 FLI1-PU1" -c "${colMUT} ${colPU1} ${colMUTFLI1} ${colFLI1}" -x
1000 -y "0 40" -d ${FIGURESDIR}/hist -n FLI1_${PEAKNAME}
```

```
# redo plots for CLUSTERSHARED PU1/mut for scaling purposes
plotHIST.sh \
-f "${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.PU1mut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.PU1.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.ETS1-PU1mut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.ETS1-PU1.txt" \
-s "mutPU1 PU1 ETS1-mutPU1 ETS1-PU1" -c "${colMUT} ${colPU1} ${colMUTETS1} ${colETS1}" -x
1000 -y "0 40" -d ${FIGURESDIR}/hist -n ETS1_CLUSTERSHARED_PU1

plotHIST.sh \
-f "${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.PU1mut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.PU1.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.FLI1-PU1mut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.FLI1-PU1.txt" \
-s "mutPU1 PU1 FLI1-mutPU1 FLI1-PU1" -c "${colMUT} ${colPU1} ${colMUTFLI1} ${colFLI1}" -x
1000 -y "0 40" -d ${FIGURESDIR}/hist -n FLI1_CLUSTERSHARED_PU1

plotHIST.sh \
-f "${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.ATACmut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.ATACPU1.txt" \
-s "mutPU1 PU1" -c "${colMUT} ${colPU1}" -x 1000 -y "0 20" -d ${FIGURESDIR}/hist -n
ATAC_CLUSTERSHARED_PU1

plotHIST.sh \
-f "${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.PU1mut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.15perc.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.50perc.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1.100perc.txt" \
-s "mutPU1 15% 50% 100%" -c "${colMUT} ${colPU109u} ${colPU13u} ${colPU1}" -x 1000 -y "0 40"
-d ${FIGURESDIR}/hist -n Titration_CLUSTERSHARED_PU1

plotHIST.sh \
-f "${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.PU1mut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.PU1.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.ETS1-PU1mut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.ETS1-PU1.txt" \
-s "mutPU1 PU1 ETS1-mutPU1 ETS1-PU1" -c "${colMUT} ${colPU1} ${colMUTETS1} ${colETS1}" -x
1000 -y "0 40" -d ${FIGURESDIR}/hist -n ETS1_CLUSTERSHARED_PU1mut

plotHIST.sh \
-f "${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.PU1mut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.PU1.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.FLI1-PU1mut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.FLI1-PU1.txt" \
-s "mutPU1 PU1 FLI1-mutPU1 FLI1-PU1" -c "${colMUT} ${colPU1} ${colMUTFLI1} ${colFLI1}" -x
1000 -y "0 40" -d ${FIGURESDIR}/hist -n FLI1_CLUSTERSHARED_PU1mut

plotHIST.sh \
-f "${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.ATACmut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.ATACPU1.txt" \
-s "mutPU1 PU1" -c "${colMUT} ${colPU1}" -x 1000 -y "0 20" -d ${FIGURESDIR}/hist -n
ATAC_CLUSTERSHARED_PU1mut

plotHIST.sh \
-f "${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.PU1mut.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.15perc.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.50perc.txt
${TMPDIR}/tmp.ghist.CLUSTERSHARED_PU1mut.100perc.txt" \
-s "mutPU1 15% 50% 100%" -c "${colMUT} ${colPU109u} ${colPU13u} ${colPU1}" -x 1000 -y "0 40"
-d ${FIGURESDIR}/hist -n Titration_CLUSTERSHARED_PU1mut

# "ghistplot" for presence of ${SINGLESPECIFIC_PU1}" "${SINGLESHARED_PU1}"
"${CLUSTERSPECIFIC_PU1}" "${CLUSTERSHARED_PU1}"

COUNT=0
_DATE=$(date +%s)
for PEAKSET in ${PEAKSETS[@]}; do
PEAKNAME="${PEAKNAMES[${COUNT}]}"
pos2bed.pl ${PEAKSET} > ${TMPDIR}/tmp.bed
$BEDTOOLS intersect -a
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed -b ${TMPDIR}/tmp.bed
-c > ${TMPDIR}/tmp.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}'
${TMPDIR}/tmp.peaks.bound.bed
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.sorted.txt
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
```

```
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.sorted.txt", header=T)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
mycol <- colorRampPalette(c("white","${colSmot}"))(3)
png(filename="${PEAKNAME}inK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R
COUNT=$((COUNT+=1))
done

# is there a preferred distance between PU.1 and ETS1 motifs in pairs?
# define all unbound pairs
$BEDTOOLS intersect -a <(cut -f1-4
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.filtered.be
d) -b "${PU1CTV1PEAKS}" -v > ${TMPDIR}/tmp1.unbound_pairs.txt
join -1 4 -2 4 -t $'\t' <(sort -k4,4 ${TMPDIR}/tmp1.unbound_pairs.txt) <(sort -k4,4
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.filtered.be
d) > ${TMPDIR}/tmp.unbound_pairs.txt
join -1 4 -2 4 -t $'\t' <(sort -k4,4 ${TMPDIR}/tmp1.unbound_pairs.txt) <(sort -k4,4
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.filtered.be
d) | cut -f 16,17 > ${TMPDIR}/tmp.unbound_pairs.txt
# 276629/495877 are not bound in CTV1

# analysis for all peaks
$BEDTOOLS intersect -a <(cut -f1-4
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.filtered.be
d) -b "${PU1CTV1PEAKS}" -u > ${TMPDIR}/pairs.all.bed
join -1 4 -2 4 -t $'\t' <(sort -k4,4 ${TMPDIR}/pairs.all.bed) <(sort -k4,4
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.filtered.be
d) | cut -f 16,17 > ${TMPDIR}/tmp.bound_pairs.txt
cat >"${TMPDIR}/R.distplot.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}/hist")
library(ggplot2)
library(grid)
# read in unbound data
udata <- read.table("${TMPDIR}/tmp.unbound_pairs.txt", header=F, sep="\t")
ud <- data.frame(udata)
colnames(ud) <- c("distance","orientation")
# seperating sense and antisense pairs
uantisense <- subset(ud, orientation==0)
usense <- subset(ud, orientation==1)
# generating count tables
uas_counttable <- as.data.frame(table(uantisense\$distance))
us_counttable <- as.data.frame(table(usense\$distance))
# read in bound data
bdata <- read.table("${TMPDIR}/tmp.bound_pairs.txt", header=F, sep="\t")
bd <- data.frame(bdata)
colnames(bd) <- c("distance","orientation")
# seperating sense and antisense pairs
bantisense <- subset(bd, orientation==0)
bsense <- subset(bd, orientation==1)
# generating count tables
bas_counttable <- as.data.frame(table(bantisense\$distance))
bs_counttable <- as.data.frame(table(bsense\$distance))
# merge sense & antisense tables and perform calculations (% frequency, hypergeom. test,
...)
sense <- merge(bs_counttable, us_counttable, by.x = "Var1" , by.y = "Var1")
sense\$perc.x <- (sense\$Freq.x * 100 / sum(sense\$Freq.x))
sense\$perc.y <- (sense\$Freq.y * 100 / sum(sense\$Freq.y))
antisense <- merge(bas_counttable, uas_counttable, by.x = "Var1" , by.y = "Var1")
antisense\$perc.x <- (antisense\$Freq.x * 100 / sum(antisense\$Freq.x))
antisense\$perc.y <- (antisense\$Freq.y * 100 / sum(antisense\$Freq.y))
maxperc <- round(max(c(sense\$perc.x,sense\$perc.y,antisense\$perc.x,antisense\$perc.y)),0)
+ 1
# perform calculations for sense (% frequency, hypergeom. test, ...)
sense\$ratio <- (sense\$Freq.x * 100 / sum(sense\$Freq.x)) / (sense\$Freq.y * 100 /
sum(sense\$Freq.y))
sense\$totpair <- (sum(sense\$Freq.x) + sum(sense\$Freq.y))
sense\$pair <- sum(sense\$Freq.x)
sense\$totbound <- sense\$Freq.x + sense\$Freq.y
sense\$bound <- sense\$Freq.x
```

```
sense\$dist <- as.numeric(levels(sense\$Var1))[sense\$Var1]
sense\$hyper <- phyper(sense\$bound,sense\$totbound,sense\$totpair,sense\$pair,lower.tail =
FALSE, log.p = FALSE)
sense\$sig <- factor(as.numeric(sense\$hyper < 0.05))
grobsense <- grobTree(textGrob("sense - sense", x=.05, y=.95, hjust=0,
gp=gpar(col="black", fontsize=6, fontface="italic")))
# generating the sense plot
p <- ggplot(data=sense, aes(x=dist,y=perc.x,fill=sig))
p <- p + geom_bar(stat="identity", position=position_dodge(width = 0),alpha=.5)
p <- p + geom_bar(data=sense, aes(x=dist,y=perc.y), stat="identity",
position=position_dodge(width = 0), fill="gray20", alpha=.3)
p <- p + xlab("Distance between motifs") + ylab("% motifs") + annotation_custom(grobsense)
p <- p + theme_light(base_size=8) + theme(panel.grid.major =
element_blank(),panel.grid.minor = element_blank())
p <- p + scale_x_continuous(expand = c(0,0),limits=c(0,150)) + scale_y_continuous(expand =
c(0,0),limits=c(0,maxperc)) + scale_x_reverse(expand = c(0,0),limits=c(150,0))
p <- p + scale_fill_manual(values=c("cornflowerblue", "red")) + theme(legend.position="none")
pdf(file="PU1ETSpairDist.allPeaks.s.hist.pdf", height=2, width=1.4)
plot(p)
dev.off()
# perform calculations for antisense (% frequency, hypergeom. test, ...)
antisense\$ratio <- (antisense\$Freq.x * 100 / sum(antisense\$Freq.x)) / (antisense\$Freq.y
* 100 / sum(antisense\$Freq.y))
antisense\$totpair <- (sum(antisense\$Freq.x) + sum(antisense\$Freq.y))
antisense\$pair <- sum(antisense\$Freq.x)
antisense\$totbound <- antisense\$Freq.x + antisense\$Freq.y
antisense\$bound <- antisense\$Freq.x
antisense\$dist <- as.numeric(levels(antisense\$Var1))[antisense\$Var1]
antisense\$hyper <-
phyper(antisense\$bound,antisense\$totbound,antisense\$totpair,antisense\$pair,lower.tail =
FALSE, log.p = FALSE)
antisense\$sig <- factor(as.numeric(antisense\$hyper < 0.05))
grobantisense <- grobTree(textGrob("sense - antisense", x=0.95, y=0.95, hjust=1,
gp=gpar(col="black", fontsize=6, fontface="italic")))
# generating the antisense plot
p <- ggplot(data=antisense, aes(x=dist,y=perc.x,fill=sig))
p <- p + geom_bar(stat="identity", position=position_dodge(width = 0),alpha=.5)
p <- p + geom_bar(data=antisense, aes(x=dist,y=perc.y), stat="identity",
position=position_dodge(width = 0), fill="gray20", alpha=.3)
p <- p + xlab("Distance between motifs") + ylab("% motifs") + annotation_custom(grobantisense)
p <- p + theme_light(base_size=8) + theme(panel.grid.major =
element_blank(),panel.grid.minor = element_blank())
p <- p + scale_x_continuous(expand = c(0,0),limits=c(0,150)) + scale_y_continuous(expand =
c(0,0),limits=c(0,maxperc),position="right")
p <- p + scale_fill_manual(values=c("cornflowerblue", "red")) + theme(legend.position="none")
pdf(file="PU1ETSpairDist.allPeaks.as.hist.pdf", height=2, width=1.4)
plot(p)
dev.off()
EOF
chmod 750 "\${TMPDIR}/R.distplot.P.${_DATE}.R"
R < ${TMPDIR}/R.distplot.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.distplot.P.${_DATE}.R

# stringent PU.1 peaks that overlap with close ETS-paired PU.1 motifs
myFilterFile.pl
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.filtered.be
d -column 13 -range -rangeU 50 -rangeL 12 >
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.sizefiltere
d.bed
wc -l
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.sizefiltere
d.bed # 137217
$BEDTOOLS intersect -b <(cut -f1-4
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.sizefiltere
d.bed) -a "${PU1CTV1PEAKS}" -u > ${ANALYSISDIR}/PU.1peaks_overlapETSPU1.pairs.12-50bp.bed
wc -l ${ANALYSISDIR}/PU.1peaks_overlapETSPU1.pairs.12-50bp.bed #10246 PU1 peaks overlap with
PU1-ETS motif pairs

# stringent FLI1 peaks that overlap with close ETS-paired PU.1 motifs
$BEDTOOLS intersect -b <(cut -f1-4
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.sizefiltere
d.bed) -a ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.bed -u >
${ANALYSISDIR}/FLI1peaks_PU1_overlapETSPU1.pairs.12-50bp.bed
wc -l ${ANALYSISDIR}/FLI1peaks_PU1_overlapETSPU1.pairs.12-50bp.bed #3111 PU1 peaks overlap
with PU1-ETS motif pairs
```

```
# stringent ETS1 peaks that overlap with close ETS-paired PU.1 motifs
$BEDTOOLS intersect -b <(cut -f1-4
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.sizefiltere
d.bed) -a ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.fdr05.ntag15.filtered.bed -u >
${ANALYSISDIR}/ETS1peaks_PU1_overlapETSPU1.pairs.12-50bp.bed
wc -l ${ANALYSISDIR}/ETS1peaks_PU1_overlapETSPU1.pairs.12-50bp.bed #3965 PU1 peaks overlap
with PU1-ETS motif pairs

# lenient FLI1 peaks that overlap with close ETS-paired PU.1 motifs
$BEDTOOLS intersect -b <(cut -f1-4
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.sizefiltere
d.bed) -a ${PEAKDIR}/FLI1_DSG_PU1_merged.factor.lenient.filtered.bed -u >
${ANALYSISDIR}/FLI1peaks.lenient_PU1_overlapETSPU1.pairs.12-50bp.bed
wc -l ${ANALYSISDIR}/FLI1peaks.lenient_PU1_overlapETSPU1.pairs.12-50bp.bed #10128 PU1 peaks
overlap with PU1-ETS motif pairs

# lenient ETS1 peaks that overlap with close ETS-paired PU.1 motifs
$BEDTOOLS intersect -b <(cut -f1-4
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.sizefiltere
d.bed) -a ${PEAKDIR}/ETS1_DSG_PU1_merged.factor.lenient.filtered.bed -u >
${ANALYSISDIR}/ETS1peaks.lenient_PU1_overlapETSPU1.pairs.12-50bp.bed
wc -l ${ANALYSISDIR}/ETS1peaks.lenient_PU1_overlapETSPU1.pairs.12-50bp.bed #11831 PU1 peaks
overlap with PU1-ETS motif pairs
cd ${ANALYSISDIR}
mergePeaks -d 100 PU.1peaks_overlapETSPU1.pairs.12-50bp.bed
FLI1peaks_PU1_overlapETSPU1.pairs.12-50bp.bed ETS1peaks_PU1_overlapETSPU1.pairs.12-50bp.bed
-venn peaks_overlapETSPU1.pairs.venn.txt -code > ${TMPDIR}/peaksPU1_ETS1.merged.txt
mergePeaks -d 100 PU.1peaks_overlapETSPU1.pairs.12-50bp.bed
FLI1peaks.lenient_PU1_overlapETSPU1.pairs.12-50bp.bed
ETS1peaks.lenient_PU1_overlapETSPU1.pairs.12-50bp.bed -venn
peaks.lenient_overlapETSPU1.pairs.venn.txt -code > ${TMPDIR}/peaksPU1_ETS1.merged.txt

# "ghistplot" for PU.1peaks_overlapETSPU1.pairs.12-50bp.bed
$BEDTOOLS intersect -a ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed
-b ${ANALYSISDIR}/PU.1peaks_overlapETSPU1.pairs.12-50bp.bed -c > ${TMPDIR}/tmp.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}' ${TMPDIR}/tmp.peaks.bound.bed
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.sorted.txt
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.sorted.txt", header=T)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
mycol <- colorRampPalette(c("white","darkviolet"))(3)
png(filename="PU.1peaks_overlapETSPU1.pairs.12-50bpinK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# "ghistplot" for PU.1peaks with a locally overlapping ETS PU1 motif (at the same site)
$BEDTOOLS intersect -a ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed
-b
/misc/data/analysis/project_PU1/heterotypicClusters/PU1class1ETSspecific_intersect.ident.bed
-c > ${TMPDIR}/tmp.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}' ${TMPDIR}/tmp.peaks.bound.bed
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.sorted.txt
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.sorted.txt", header=T)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
mycol <- colorRampPalette(c("white","lightpink4"))(3)
png(filename="PU.1peaks_overlapETSPU1.singleinK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
```

```
rm ${TMPDIR}/R.image.P.${_DATE}.R
```

### 10.1.6 Analysis of Homotypic PU.1 Clusters

The following script was used to analyze the occurrence of homotypic PU.1 clusters in CTV-1 cells.

```
#!/bin/bash
#setting homer environment
DIR_PKG="/misc/software/ngs"
PATH_PERL=/misc/software/package/perl/perl-5.26.1/bin
PATH_SAMTOOLS=${DIR_PKG}/samtools/samtools-1.6/bin
PATH_HOMER=${DIR_PKG}/homer/v4.9/bin
PATH_R=/misc/software/package/RBioC/3.4.3/bin
export PATH=${PATH_R}:${PATH_PERL}:${PATH_SAMTOOLS}:${PATH_HOMER}:${PATH}
export PATH

BEDTOOLS="/misc/software/ngs/bedtools/bedtools2-2.27.1/bin/bedtools"
BLACKLIST_HG19="/misc/data/analysis/generalStuff/annotation/hg19/hg19.blacklist.bed"
WORKDIR="/misc/data/analysis/project_PU1/homotypicClusters/"
CHROMSIZES_HG19="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/hg19.chrom.sizes"
GENOME_HG19="/misc/software/ngs/genome/sequence/hg19/hg19.fa"
MOTIFDIR="${WORKDIR}/motifs"
TMPDIR="/loctmp"

mkdir ${MOTIFDIR}
mkdir ${WORKDIR}

# ANALYSIS OF HOMOTYPIC PU1 CLUSTERS

# HG19 PU.1 motifs repeat masked and wo
# all motifs across the genome
#----------------------------
scanMotifGenomeWide.pl /misc/data/analysis/project_PU1/PU1long.motif \
hg19 -keepAll -p 23 > ${WORKDIR}/PU.1_long_hg19_all.pos.txt
pos2bed.pl ${WORKDIR}/PU.1_long_hg19_all.pos.txt > ${TMPDIR}/tmp.PU.1_long_hg19_all.bed
# add score again
join -1 1 -2 4 -t $'\t' ${WORKDIR}/PU.1_long_hg19_all.pos.txt
${TMPDIR}/tmp.PU.1_long_hg19_all.bed > ${TMPDIR}/tmp2.PU.1_long_hg19_all.bed
awk -v OFS='\t' '{print $8,$9,$10,$1,$6,$5}' ${TMPDIR}/tmp2.PU.1_long_hg19_all.bed >
${WORKDIR}/PU.1_long_hg19_all.bed
wc -l PU.1_long_hg19_all.bed #2242815
filterMotifClusters.py PU.1_long_hg19_all.bed PU.1_long_hg19_all ${WORKDIR}
# Finding homotypic clusters with a maximum distance of 150 bp
# Done. Filtered 552025 motifs/249129 clusters in 8 seconds.

# with repeats masked
#--------------------
cd ${WORKDIR}
scanMotifGenomeWide.pl /misc/data/analysis/project_PU1/PU1long.motif \
hg19 -mask -keepAll -p 23 > ${WORKDIR}/PU.1_long_hg19_rm.pos.txt
pos2bed.pl ${WORKDIR}/PU.1_long_hg19_rm.pos.txt > ${TMPDIR}/tmp.PU.1_long_hg19_rm.bed
# add score again
join -1 1 -2 4 -t $'\t' ${WORKDIR}/PU.1_long_hg19_rm.pos.txt
${TMPDIR}/tmp.PU.1_long_hg19_rm.bed > ${TMPDIR}/tmp2.PU.1_long_hg19_rm.bed
awk -v OFS='\t' '{print $8,$9,$10,$1,$6,$5}' ${TMPDIR}/tmp2.PU.1_long_hg19_rm.bed >
${WORKDIR}/PU.1_long_hg19_rm.bed
wc -l PU.1_long_hg19_rm.bed #1180218
filterMotifClusters.py PU.1_long_hg19_rm.bed PU.1_long_hg19_rm ${WORKDIR}
# Finding homotypic clusters with a maximum distance of 150 bp
# Done. Filtered 253632 motifs/117326 clusters in 4 seconds.
# equivalent set for peaks and reduced to reference chromosomes
#----------------------------------------------------------

cd ${WORKDIR}
CHRLIST=''
for i in {1..22}; do
CHRLIST="$CHRLIST chr$i"
done
CHRLIST="$CHRLIST chrX chrY"
echo $CHRLIST
awk -v var="${CHRLIST}" 'BEGIN{split(var, arr); for (i in arr) names[arr[i]] } $1 in names'
PU.1_long_hg19_all.bed > ${TMPDIR}/PU.1_long_hg19_refChr.bed
awk -v OFS='\t' '{ if (word == $4) { counter++ } else { counter = 1; word = $4 };print
$1,$2,$3,$4"_"counter,$5,$6 }' ${TMPDIR}/PU.1_long_hg19_refChr.bed > PU.1_long_hg19_refChr.bed
$BEDTOOLS intersect -a PU.1_long_hg19_refChr.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.1.bed
$BEDTOOLS slop -i ${TMPDIR}/tmp.1.bed -g $CHROMSIZES_HG19 -b 94 > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Done. Filtered out 322829 regions with mappability scores below 0.8, leaving 1879437
regions.
pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt > ${TMPDIR}/tmp.3.bed
# add back motif scores
join -1 4 -2 4 -t $'\t' <(sort -k4,4 ${TMPDIR}/tmp.3.bed) <(sort -k4,4
PU.1_long_hg19_refChr.bed) > ${TMPDIR}/tmp.3b.bed
awk -v OFS='\t' '{print $7,$8,$9,$1,$10,$11 }' ${TMPDIR}/tmp.3b.bed > ${TMPDIR}/tmp.4.bed
sort -k1,1 -k2,2n ${TMPDIR}/tmp.4.bed >PU.1_long_hg19_refChr.filtered.bed
filterMotifClusters.py PU.1_long_hg19_refChr.filtered.bed PU.1_long_hg19_refChr.filtered

${WORKDIR}
# Finding homotypic clusters with a maximum distance of 150 bp
# Done. Filtered 455605 motifs/208007 clusters in 7 seconds.
# reformating pairs files to bed
cut -f 1-4 ${WORKDIR}/PU.1_long_hg19_all.pairs.txt | tail -n +2 >
${WORKDIR}/PU.1_long_hg19_all.pairs.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}' ${WORKDIR}/PU.1_long_hg19_all.pairs.4bed > \
${WORKDIR}/PU.1_long_hg19_all.pairs.pos.bed
bed2pos.pl ${WORKDIR}/PU.1_long_hg19_all.pairs.pos.bed >
${WORKDIR}/PU.1_long_hg19_all.pairs.pos.txt
cut -f 1-4 ${WORKDIR}/PU.1_long_hg19_rm.pairs.txt | tail -n +2 >
${WORKDIR}/PU.1_long_hg19_rm.pairs.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}' ${WORKDIR}/PU.1_long_hg19_rm.pairs.4bed >
${WORKDIR}/PU.1_long_hg19_rm.pairs.pos.bed
bed2pos.pl ${WORKDIR}/PU.1_long_hg19_rm.pairs.pos.bed >
${WORKDIR}/PU.1_long_hg19_rm.pairs.pos.txt
cut -f 1-4 ${WORKDIR}/PU.1_long_hg19_refChr.filtered.pairs.txt | tail -n +2 >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.pairs.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}'
${WORKDIR}/PU.1_long_hg19_refChr.filtered.pairs.4bed >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.pos.bed
bed2pos.pl ${WORKDIR}/PU.1_long_hg19_refChr.filtered.pos.bed >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.pairs.pos.txt
# filter out low distance pairs (<40bp)
myFilterFile.pl ${WORKDIR}/PU.1_long_hg19_refChr.filtered.pairs.txt -column 9 -upperlimit 40
> ${WORKDIR}/PU.1_long_hg19_refChr.filtered.shortpairs.txt
cut -f 1-4 ${WORKDIR}/PU.1_long_hg19_refChr.filtered.shortpairs.txt | tail -n +2 >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.shortpairs.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}'
${WORKDIR}/PU.1_long_hg19_refChr.filtered.shortpairs.4bed >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.short.pos.bed
bed2pos.pl ${WORKDIR}/PU.1_long_hg19_refChr.filtered.short.pos.bed >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.pairs.short.pos.txt

# filter out 6bp distance pairs
myFilterFile.pl ${WORKDIR}/PU.1_long_hg19_refChr.filtered.pairs.txt -column 9 -upperlimit 6
> ${WORKDIR}/PU.1_long_hg19_refChr.filtered.6-bp.txt
cut -f 1-4 ${WORKDIR}/PU.1_long_hg19_refChr.filtered.6-bp.txt | tail -n +2 >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.6-bp.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}'
${WORKDIR}/PU.1_long_hg19_refChr.filtered.6-bp.4bed >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.6-bp.pos.bed
#bed2pos.pl ${WORKDIR}/PU.1_long_hg19_refChr.filtered.6-bp.pos.bed >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.pairs.6-bp.pos.txt
# filter pairs with a distance between 25-75 bp
myFilterFile.pl ${WORKDIR}/PU.1_long_hg19_refChr.filtered.pairs.txt -column 9 -upperlimit 75
> ${TMPDIR}/tmp.PU.1_long_hg19_refChr.filtered.txt
myFilterFile.pl ${TMPDIR}/tmp.PU.1_long_hg19_refChr.filtered.txt -column 9 -lowerlimit 25 >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.25-75-bp.txt
cut -f 1-4 ${WORKDIR}/PU.1_long_hg19_refChr.filtered.25-75-bp.txt | tail -n +2 >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.25-75-bp.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}'
${WORKDIR}/PU.1_long_hg19_refChr.filtered.25-75-bp.4bed >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.25-75-bp.pos.bed
bed2pos.pl ${WORKDIR}/PU.1_long_hg19_refChr.filtered.25-75-bp.pos.bed >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.pairs.25-75-bp.pos.txt

# closeby sites are enriched at 6bp, 8bp and 12bp distance

# search longer motifs in pairs
findMotifsGenome.pl ${WORKDIR}/PU.1_long_hg19_refChr.filtered.pairs.short.pos.txt hg19
${MOTIFDIR}/PU1_pairs -size 100 -len 18,19,20,21,22,23,24 -p 10 -h
```

```
# reformating output with own parameters
compareMotifs.pl ${MOTIFDIR}/PU1_pairs/homerMotifs.all.motifs ${MOTIFDIR}/PU1_pairs/final
-reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 12
# extracting sequences for motifs

# all motifs
bed2pos.pl ${WORKDIR}/PU.1_long_hg19_refChr.bed > ${WORKDIR}/PU.1_long_hg19_refChr.pos.txt
homerTools extract ${WORKDIR}/PU.1_long_hg19_refChr.pos.txt ${GENOME_HG19}
>${WORKDIR}/PU.1_long_hg19_refChr.seq.txt

# motifs in mappable regions
homerTools extract ${WORKDIR}/PU.1_long_hg19_refChr.filtered.bed ${GENOME_HG19}
>${WORKDIR}/PU.1_long_hg19_refChr.filtered.seq.txt
join -1 4 -2 1 -t $'\t' <(sort -k4,4 ${WORKDIR}/PU.1_long_hg19_refChr.filtered.bed) <(sort
-k1,1 ${WORKDIR}/PU.1_long_hg19_refChr.filtered.seq.txt) >
/misc/data/tmp/PU.1_long_hg19_refChr.filtered.bed
awk -v OFS='\t' '{print $2,$3,$4,($1 "%" $7),$5,$6,$7 }'
/misc/data/tmp/PU.1_long_hg19_refChr.filtered.bed >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.seq.bed
for w in `cut -f 7 ${WORKDIR}/PU.1_long_hg19_refChr.filtered.seq.bed`; do echo $w;
done|sort|uniq -c|sort -k1,1nr|sed 's/^ *//g'|tr [:blank:] \\t >
${WORKDIR}/PU.1_long_hg19_refChr.filtered.seq.count.txt

# annotation of homotypic clusters

# prepare bound peaks overlapping and non-overlapping clusters for histograms etc.
cd ${PEAKDIR}
$BEDTOOLS intersect -a ${PU1CTV1PEAKS} -b <(cut -f1-3
${HOMOCLUSTERDIR}/PU.1_long_hg19_all.cluster.txt) -u >PU1.ntag15.filtered.cluster.bed
$BEDTOOLS intersect -a ${PU1CTV1PEAKS} -b ${HOMOCLUSTERDIR}/PU.1_long_hg19_all.bed -u
>PU1.ntag15.filtered.motif.bed
$BEDTOOLS intersect -a ${PU1CTV1PEAKS} -b ${HOMOCLUSTERDIR}/PU.1_long_hg19_all.bed -v
>PU1.ntag15.filtered.noMotif.bed
$BEDTOOLS intersect -a PU1.ntag15.filtered.motif.bed -b PU1.ntag15.filtered.cluster.bed -v
>PU1.ntag15.filtered.singleMotif.bed
bed2pos.pl PU1.ntag15.filtered.cluster.bed > PU1.ntag15.filtered.cluster.pos.txt
bed2pos.pl PU1.ntag15.filtered.singleMotif.bed > PU1.ntag15.filtered.singleMotif.pos.txt
bed2pos.pl PU1.ntag15.filtered.noMotif.bed > PU1.ntag15.filtered.noMotif.pos.txt
```

## 10.1.7 Analysis of Heterotypic PU.1 Clusters

The following script was used to analyze the occurrence of heterotypic PU.1 clusters in CTV-1 cells.

```
#!/bin/bash
# setting homer environment
DIR_PKG="/misc/software/ngs"
PATH_PERL=/misc/software/package/perl/perl-5.26.1/bin
PATH_SAMTOOLS=${DIR_PKG}/samtools/samtools-1.6/bin
PATH_HOMER=${DIR_PKG}/homer/v4.9/bin
PATH_R=/misc/software/package/RBioC/3.4.3/bin
export PATH=${PATH_R}:${PATH_PERL}:${PATH_SAMTOOLS}:${PATH_HOMER}:${PATH}
export PATH

BEDTOOLS="/misc/software/ngs/bedtools/bedtools2-2.27.1/bin/bedtools"
BLACKLIST_HG19="/misc/data/analysis/generalStuff/annotation/hg19/hg19.blacklist.bed"
WORKDIR="/misc/data/analysis/project_PU1/heterotypicClusters/"
CHROMSIZES_HG19="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/hg19.chrom.sizes"
GENOME_HG19="/misc/software/ngs/genome/sequence/hg19/hg19.fa"
MOTIFDIR="${WORKDIR}/motifs"
TMPDIR="/loctmp"

PU1POS="/misc/data/analysis/project_PU1/homotypicClusters/PU.1_long_hg19_refChr.filtered.bed"
PU1MOTIF="/misc/data/analysis/project_PU1/PU1long.motif"
PU1MOTIFR="/misc/data/analysis/project_PU1/PU1long.rev.motif"
CLASS1ETSMOTIF="/misc/data/analysis/project_PU1/class1ETSlong.motif"
```

```
# ANALYSIS OF HETEROTYPIC PU1 CLUSTERS
# first, define homotypic ETS clusters
# HG19 ETS motifs repeat masked and wo
# all ETS motifs across the genome
#--------------------------------
cd ${WORKDIR}
scanMotifGenomeWide.pl ${CLASS1ETSMOTIF} \
hg19 -keepAll -p 24 > ${WORKDIR}/class1ETSspecific_hg19_all.pos.txt
pos2bed.pl ${WORKDIR}/class1ETSspecific_hg19_all.pos.txt >
${TMPDIR}/tmp.class1ETSspecific_hg19_all.bed

# add score again
join -1 1 -2 4 -t $'\t' ${WORKDIR}/class1ETSspecific_hg19_all.pos.txt
${TMPDIR}/tmp.class1ETSspecific_hg19_all.bed > ${TMPDIR}/tmp2.class1ETSspecific_hg19_all.bed
awk -F '\t' '{print $8,$9,$10,$1,$6,$5}' ${TMPDIR}/tmp2.class1ETSspecific_hg19_all.bed >
${WORKDIR}/class1ETSspecific_hg19_all.bed
wc -l class1ETSspecific_hg19_all.bed #2141103
filterMotifClusters.py class1ETSspecific_hg19_all.bed class1ETSspecific_hg19_all ${WORKDIR}

# Finding homotypic clusters with a maximum distance of 150 bp
# Done. Filtered 519882 motifs/238687 clusters in 8 seconds.

# equivalent set for peaks and reduced to reference chromosomes
#---------------------------------------------------------
cd ${WORKDIR}
CHRLIST=''
for i in {1..22}; do
CHRLIST="$CHRLIST chr$i"
done
CHRLIST="$CHRLIST chrX chrY"
echo $CHRLIST
awk -v var="${CHRLIST}" 'BEGIN{split(var, arr); for (i in arr) names[arr[i]] } $1 in names'
class1ETSspecific_hg19_all.bed > ${TMPDIR}/class1ETSspecific_hg19_refChr.bed
awk -v OFS='\t' '{ if (word == $4) { counter++ } else { counter = 1; word = $4 };print
$1,$2,$3,$4"_"counter,$5,$6 }' ${TMPDIR}/class1ETSspecific_hg19_refChr.bed >
class1ETSspecific_hg19_refChr.bed
$BEDTOOLS intersect -a class1ETSspecific_hg19_refChr.bed -b $BLACKLIST_HG19 -v >
${TMPDIR}/tmp.1.bed
$BEDTOOLS slop -i ${TMPDIR}/tmp.1.bed -g $CHROMSIZES_HG19 -b 94 > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
# Done. Filtered out 368184 regions with mappability scores below 0.8, leaving 1734253
regions.
pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt > ${TMPDIR}/tmp.3.bed

# add back motif scores
join -1 4 -2 4 -t $'\t' <(sort -k4,4 ${TMPDIR}/tmp.3.bed) <(sort -k4,4
class1ETSspecific_hg19_refChr.bed) > ${TMPDIR}/tmp.3b.bed
awk -v OFS='\t' '{print $7,$8,$9,$1,$10,$11 }' ${TMPDIR}/tmp.3b.bed > ${TMPDIR}/tmp.4.bed
awk '{gsub(/,BestGuess:Etv2\(ETS\)\/ES-ER71-ChIP-Seq\(GSE59402\)\/Homer\(0.967\)\(0.981\)/,
"");print}' ${TMPDIR}/tmp.4.bed > ${TMPDIR}/tmp.5.bed
head ${TMPDIR}/tmp.5.bed
sort -k1,1 -k2,2n ${TMPDIR}/tmp.5.bed >class1ETSspecific_hg19_refChr.filtered.bed
filterMotifClusters.py class1ETSspecific_hg19_refChr.filtered.bed
class1ETSspecific_hg19_refChr.filtered ${WORKDIR}
# Finding homotypic clusters with a maximum distance of 150 bp
# Done. Filtered 406250 motifs/186724 clusters clusters in 7 seconds.

# reformating pairs files to bed
cut -f 1-4 ${WORKDIR}/class1ETSspecific_hg19_all.pairs.txt | tail -n +2 >
${WORKDIR}/class1ETSspecific_hg19_all.pairs.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}'
${WORKDIR}/class1ETSspecific_hg19_all.pairs.4bed > \
${WORKDIR}/class1ETSspecific_hg19_all.pairs.pos.bed
bed2pos.pl ${WORKDIR}/class1ETSspecific_hg19_all.pairs.pos.bed >
${WORKDIR}/class1ETSspecific_hg19_all.pairs.pos.txt
cut -f 1-4 ${WORKDIR}/class1ETSspecific_hg19_rm.pairs.txt | tail -n +2 >
${WORKDIR}/class1ETSspecific_hg19_rm.pairs.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}'
${WORKDIR}/class1ETSspecific_hg19_rm.pairs.4bed >
${WORKDIR}/class1ETSspecific_hg19_rm.pairs.pos.bed
bed2pos.pl ${WORKDIR}/class1ETSspecific_hg19_rm.pairs.pos.bed >
${WORKDIR}/class1ETSspecific_hg19_rm.pairs.pos.txt
cut -f 1-4 ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.txt | tail -n +2 >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}'
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.4bed >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pos.bed
bed2pos.pl ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pos.bed >
```

```
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.pos.txt

# filter out low distance pairs (<40bp)
myFilterFile.pl ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.txt -column 9
-upperlimit 40 > ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.shortpairs.txt
cut -f 1-4 ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.shortpairs.txt | tail -n +2 >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.shortpairs.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}'
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.shortpairs.4bed >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.short.pos.bed
bed2pos.pl ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.short.pos.bed >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.short.pos.txt

# filter out 6bp distance pairs
myFilterFile.pl ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.txt -column 9
-upperlimit 6 > ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.6-bp.txt
cut -f 1-4 ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.6-bp.txt | tail -n +2 >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.6-bp.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}'
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.6-bp.4bed >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.6-bp.pos.bed
# bed2pos.pl ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.6-bp.pos.bed >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.6-bp.pos.txt

# filter pairs with a distance between 25-75 bp
myFilterFile.pl ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.txt -column 9
-upperlimit 75 > ${TMPDIR}/tmp.class1ETSspecific_hg19_refChr.filtered.txt
myFilterFile.pl ${TMPDIR}/tmp.class1ETSspecific_hg19_refChr.filtered.txt -column 9
-lowerlimit 25 > ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.25-75-bp.txt
cut -f 1-4 ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.25-75-bp.txt | tail -n +2 >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.25-75-bp.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}'
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.25-75-bp.4bed >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.25-75-bp.pos.bed
bed2pos.pl ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.25-75-bp.pos.bed >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.25-75-bp.pos.txt

# filter pairs with a distance between 12-50 bp
myFilterFile.pl ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.txt -column 9
-upperlimit 50 > ${TMPDIR}/tmp.class1ETSspecific_hg19_refChr.filtered.txt
myFilterFile.pl ${TMPDIR}/tmp.class1ETSspecific_hg19_refChr.filtered.txt -column 9
-lowerlimit 12 > ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.12-50-bp.txt
cut -f 1-4 ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.12-50-bp.txt | tail -n +2 >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.12-50-bp.4bed
awk -v OFS='\t' '{print $1,$2,$3,$4,"0","."}'
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.12-50-bp.4bed >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.12-50-bp.pos.bed
bed2pos.pl ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.12-50-bp.pos.bed >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.12-50-bp.pos.txt
# closeby sites are enriched at 6bp, 8bp and 12bp distance

# search longer motifs in pairs
findMotifsGenome.pl ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.pairs.short.pos.txt
hg19 ${MOTIFDIR}/PU1_pairs -size 100 -len 18,19,20,21,22,23,24 -p 10 -h
# reformating output with own parameters
compareMotifs.pl ${MOTIFDIR}/PU1_pairs/homerMotifs.all.motifs ${MOTIFDIR}/PU1_pairs/final
-reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info 1.5 -cpu 12
# extracting sequences for motifs

# all motifs
bed2pos.pl ${WORKDIR}/class1ETSspecific_hg19_refChr.bed >
${WORKDIR}/class1ETSspecific_hg19_refChr.pos.txt
homerTools extract ${WORKDIR}/class1ETSspecific_hg19_refChr.pos.txt ${GENOME_HG19}
>${WORKDIR}/class1ETSspecific_hg19_refChr.seq.txt

# motifs in mappable regions
homerTools extract ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.bed ${GENOME_HG19}
>${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.seq.txt
join -1 4 -2 1 -t $'\t' <(sort -k4,4 ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.bed)
<(sort -k1,1 ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.seq.txt) >
/misc/data/tmp/class1ETSspecific_hg19_refChr.filtered.bed
awk -v OFS='\t' '{print $2,$3,$4,($1 "%" $7),$5,$6,$7 }'
/misc/data/tmp/class1ETSspecific_hg19_refChr.filtered.bed >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.seq.bed
for w in `cut -f 7 ${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.seq.bed`; do echo $w;
done|sort|uniq -c|sort -k1,1nr|sed 's/^ *//g'|tr [:blank:] \\t >
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.seq.count.txt
```

```
# definition of heterotypic PU.1/ETS clusters
# HG19 PU.1 ETS motif distribution
annotatePeaks.pl ${PU1POS} hg19 -size 300 -m ${CLASS1ETSMOTIF} -noann -nogene -cpu 8 >
${WORKDIR}/PU1motif_overlapclass1ETSspecific_hg19.ann.txt
myFilterFile.pl ${WORKDIR}/PU1motif_overlapclass1ETSspecific_hg19.ann.txt -column 10
-removeempty >${TMPDIR}/tmp.PU1motif_overlapclass1ETSspecific_hg19.ann.txt

# finding overlaps around PU.1 sites in 300 bp windows
${BEDTOOLS} slop -i ${PU1POS} -b 144 -g $CHROMSIZES_HG19 > ${TMPDIR}/tmp.PU1.bed
${BEDTOOLS} intersect -a ${TMPDIR}/tmp.PU1.bed -b
${WORKDIR}/class1ETSspecific_hg19_refChr.filtered.bed -loj >
${TMPDIR}/tmp.PU1class1ETSspecific_intersect.bed
myFilterFile.pl ${TMPDIR}/tmp.PU1class1ETSspecific_intersect.bed -column 7 -substring . >
${WORKDIR}/PU1class1ETSspecific_intersect.bed
${BEDTOOLS} intersect -a ${TMPDIR}/tmp.PU1.bed -b <(cut -f1-3
${WORKDIR}/PU1class1ETSspecific_intersect.bed) -v > ${TMPDIR}/tmp.no_intersect.bed
${BEDTOOLS} slop -i ${TMPDIR}/tmp.no_intersect.bed -b -144 -g $CHROMSIZES_HG19 >
${WORKDIR}/PU1class1ETSspecific_NOintersect.bed
wc -l ${WORKDIR}/PU1class1ETSspecific_NOintersect.bed # non overlapping sites: 1056791

# annotate motif distances
# first shift position of ETSmotif to match PU.1
awk -v OFS='\t' '{if($12 ~ "+") {$8=($8-2);$9=($9-2); print} else {$8=($8+2);$9=($9+2);
print} }' ${WORKDIR}/PU1class1ETSspecific_intersect.bed > ${TMPDIR}/tmp.cor.intersect.bed
# calculate distance
awk -v OFS='\t' '{print $0, ($2+144-$8) }' ${TMPDIR}/tmp.cor.intersect.bed >
${TMPDIR}/tmp.intersect.bed
# resize position of PU1 motif
awk -v OFS='\t' '{$2=($2+144); $3=($2+12); print}' ${TMPDIR}/tmp.intersect.bed >
${TMPDIR}/tmp.motifs.intersect.bed
# replace distance with absolute distance
awk -v OFS='\t' '{if ($13 < 0) {($13=-$13); print} else {print $0} }'
${TMPDIR}/tmp.motifs.intersect.bed > ${TMPDIR}/tmp.cor.motifs.intersect.bed
# check pair orientation
awk -v OFS='\t' '{if($12 ~ $6) {print $0, "0"} else {print $0, "1"} }'
${TMPDIR}/tmp.cor.motifs.intersect.bed > ${TMPDIR}/tmp.format.intersect.bed
# filter out exact overlaps
awk -v OFS='\t' '{if($13!=0) print}' ${TMPDIR}/tmp.format.intersect.bed >
${WORKDIR}/PU1class1ETSspecific_intersect.filtered.bed
wc -l ${WORKDIR}/PU1class1ETSspecific_intersect.filtered.bed # 495877 paired sites
# print exact overlaps
awk -v OFS='\t' '{if($13 < 1) print $0}' ${TMPDIR}/tmp.format.intersect.bed >
${WORKDIR}/PU1class1ETSspecific_intersect.ident.bed # 408205 overlapping motifs
```

### 10.1.8 Analysis of NGS Data of PU.1-deletion Constructs

The following script was used to analyze the binding properties of PU.1-deletion constructs in CTV-1 cells. The script includes parts of edgeR (Robinson et al. 2010), the HOMER suite (Heinz et al. 2010) as well as parts of the R software (R Development Core Team 2008).

```
#!/bin/bash
#setting homer environment
DIR_PKG="/misc/software/ngs"
PATH_PERL=/misc/software/package/perl/perl-5.26.1/bin
PATH_SAMTOOLS=${DIR_PKG}/samtools/samtools-1.6/bin
PATH_HOMER=${DIR_PKG}/homer/v4.9/bin
PATH_R=/misc/software/package/RBioC/3.4.3/bin
export PATH=${PATH_R}:${PATH_PERL}:${PATH_SAMTOOLS}:${PATH_HOMER}:${PATH}
export PATH

WORKDIR="/misc/data/analysis/project_PU1/CTV1/ChIP"
ATACDIR="/misc/data/processedData/tagDir/chromatin/hg19/ATAC/RNAtransfection/PU1"
TAGDIR="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/RNAtransfection/PU1"
BIGWIGDIR="/misc/data/processedData/bigWig/chromatin/hg19/ChIP/RNAtransfection/PU1/"
PEAKDIR="${WORKDIR}/peaksCNVcorr"
CHROMSIZES="/misc/software/viewer/IGV/IGVTools_2.3.98/genomes/hg19.chrom.sizes"
DIFFDIR="${WORKDIR}/diffPeaksCNVcorr"
BLACKLIST_HG19="/misc/data/analysis/generalStuff/annotation/hg19/hg19.blacklist.bed"
FIGURESDIR="/misc/data/analysis/project_PU1/CTV1/figures/deletionConstructs"
BEDTOOLS="/misc/software/ngs/bedtools/bedtools2-2.27.1/bin/bedtools"
MOTIFDIR="/misc/data/analysis/project_PU1/CTV1/motifs/deletionConstructs"
```

```
GHISTFILEDIR="${WORKDIR}/ghists/del"
CNVFILECTV1="/misc/data/processedData/mapping/DNA/hg19/Input/CellLines/CNVdata/CTV1_genInput_6
20.sam_CNVs"
PU1CTV1PEAKS="${PEAKDIR}/PU1_Flag_merged.factor.fdr05.ntag15.filtered.bed"
HG19GTF="/misc/software/ngs/genome/annotation/hg19/gencode.v19.annotation.gtf"
PU1MOTIF="/misc/data/analysis/project_PU1/PU1long.motif"
HCDIR="/misc/data/analysis/project_PU1/homotypicClusters"
BASICDIR="/misc/data/analysis/project_PU1/CTV1/analysis/basic"
ANALYSISDIR="/misc/data/analysis/project_PU1/CTV1/analysis/deletionConstructs"
TMPDIR="/loctmp"

# defining the colors
colPU1="blue"
colPU16u="dodgerblue4"
colPU13u="dodgerblue3"
colPU1p9u="dodgerblue1"
colDelP="blueviolet"
colDelQ="forestgreen"
colDelA="orange4"
colDelAQP="firebrick"
colMUT="gray50"
colClus="limegreen"
colSmot="mediumslateblue"
colNmot="black"

mkdir -p ${FIGURESDIR}/R
mkdir -p ${MOTIFDIR}
mkdir -p ${PEAKDIR}
mkdir -p ${GHISTFILEDIR}
mkdir -p ${FIGURESDIR}/hist/
mkdir -p ${DIFFDIR}
mkdir -p ${ANALYSISDIR}

# CNV normalization of tagDirs
# normalize & reduce tagDir for better comparability with other data sets
declare -a oCHIPDIRS=("${TAGDIR}/CTV1_PU1_Flag_3ug" "${TAGDIR}/CTV1_PU1_Flag_0.9ug"
"${TAGDIR}/CTV1_PU.1delP_Flag_R1" "${TAGDIR}/CTV1_PU.1delP_Flag_R2"
"${TAGDIR}/CTV1_PU.1delQ_Flag_R1" "${TAGDIR}/CTV1_PU.1delQ_Flag_R2"
"${TAGDIR}/CTV1_PU.1delA_Flag_R1" "${TAGDIR}/CTV1_PU.1delA_Flag_R2"
"${TAGDIR}/CTV1_PU.1delAQP_Flag_R1" "${TAGDIR}/CTV1_PU.1delAQP_Flag_R2")

declare -a oINPUTDIRS=("${TAGDIR}/CTV1_PU1mut_Flag" "${TAGDIR}/CTV1_PU1mut_Flag"
"${TAGDIR}/CTV1_Mock_Flag_R1" "${TAGDIR}/CTV1_PU1mut_Flag" "${TAGDIR}/CTV1_Mock_Flag_R1"
"${TAGDIR}/CTV1_PU1mut_Flag" "${TAGDIR}/CTV1_Mock_Flag_R1" "${TAGDIR}/CTV1_PU1mut_Flag"
"${TAGDIR}/CTV1_Mock_Flag_R1" "${TAGDIR}/CTV1_PU1mut_Flag")

COUNT=0
for SAMPLE in ${oCHIPDIRS[@]}; do
INPUT="${oINPUTDIRS[${COUNT}]}"
normalizeTagDirByCopyNumber.pl ${SAMPLE} -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl ${INPUT} -cnv "${CNVFILECTV1}" -remove
COUNT=$((COUNT+=1))
Done

cd ${TAGDIR}
makeTagDirectory CTV1_PU1mut_Flag_CNVnormRefChr_merged -d CTV1_PU1mut_Flag_R1_CNVnormRefChr
CTV1_PU1mut_Flag_R2_CNVnormRefChr CTV1_PU1mut_Flag_R3_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_PU1delA_Flag_CNVnormRefChr_merged -d
CTV1_PU.1delA_Flag_R1_CNVnormRefChr CTV1_PU.1delA_Flag_R2_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_PU1delQ_Flag_CNVnormRefChr_merged -d
CTV1_PU.1delQ_Flag_R1_CNVnormRefChr CTV1_PU.1delQ_Flag_R2_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_PU1delP_Flag_CNVnormRefChr_merged -d
CTV1_PU.1delP_Flag_R1_CNVnormRefChr CTV1_PU.1delP_Flag_R2_CNVnormRefChr -genome hg19 -checkGC
makeTagDirectory CTV1_PU1delAQP_Flag_CNVnormRefChr_merged -d
CTV1_PU.1delAQP_Flag_R1_CNVnormRefChr CTV1_PU.1delAQP_Flag_R2_CNVnormRefChr -genome hg19
-checkGC

# normalization BRG1-Inhibitor treated samples
cd ${TAGDIR}
normalizeTagDirByCopyNumber.pl CTV1_PU1_FLAG_8h_1uM_PFI3 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1_FLAG_8h_5uM_PFI3 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1_FLAG_8h_10uM_PFI3 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU1_FLAG_8h_DMSO -cnv "${CNVFILECTV1}" -remove

cd ${TAGDIR}
findPeaks CTV1_PU1delAQP_Flag_CNVnormRefChr_merged -i CTV1_PU1mut_CNVnormRefChr_Flag_merged
-style factor -fdr 0.00001 -o ${PEAKDIR}/PU1delAQP_Flag_merged.factor.fdr05.peaks.txt
findPeaks CTV1_PU1delA_Flag_CNVnormRefChr_merged -i CTV1_PU1mut_CNVnormRefChr_Flag_merged
```

```
-style factor -fdr 0.00001 -o ${PEAKDIR}/PU1delA_Flag_merged.factor.fdr05.peaks.txt
findPeaks CTV1_PU1delQ_Flag_CNVnormRefChr_merged -i CTV1_PU1mut_CNVnormRefChr_Flag_merged
-style factor -fdr 0.00001 -o ${PEAKDIR}/PU1delQ_Flag_merged.factor.fdr05.peaks.txt
findPeaks CTV1_PU1delP_Flag_CNVnormRefChr_merged -i CTV1_PU1mut_CNVnormRefChr_Flag_merged
-style factor -fdr 0.00001 -o ${PEAKDIR}/PU1delP_Flag_merged.factor.fdr05.peaks.txt
findPeaks CTV1_PU1_Flag_0.9ug_CNVnormRefChr -i ${TAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr
-style factor -fdr 0.00001 -o ${PEAKDIR}/PU1_Flag_0.9ug.factor.fdr05.peaks.txt

# relevant peaks further filtered for peaks with at least 15 tags
myFilterFile.pl ${PEAKDIR}/PU1delA_Flag_merged.factor.fdr05.peaks.txt -column 6 -lowerlimit
15 > ${PEAKDIR}/PU1delA_Flag_merged.factor.fdr05.ntag15.peaks.txt
myFilterFile.pl ${PEAKDIR}/PU1delQ_Flag_merged.factor.fdr05.peaks.txt -column 6 -lowerlimit
15 > ${PEAKDIR}/PU1delQ_Flag_merged.factor.fdr05.ntag15.peaks.txt
myFilterFile.pl ${PEAKDIR}/PU1delP_Flag_merged.factor.fdr05.peaks.txt -column 6 -lowerlimit
15 > ${PEAKDIR}/PU1delP_Flag_merged.factor.fdr05.ntag15.peaks.txt
myFilterFile.pl ${PEAKDIR}/PU1delAQP_Flag_merged.factor.fdr05.peaks.txt -column 6
-lowerlimit 15 > ${PEAKDIR}/PU1delAQP_Flag_merged.factor.fdr05.ntag15.peaks.txt
myFilterFile.pl ${PEAKDIR}/PU1_Flag_0.9ug.factor.fdr05.peaks.txt -column 6 -lowerlimit 15 >
${PEAKDIR}/PU1_Flag_0.9ug.factor.fdr05.ntag15.peaks.txt

mergePeaks -d 100 ${PEAKDIR}/PU1_Flag_merged.factor.fdr05.ntag15.peaks.txt \
${PEAKDIR}/PU1delA_Flag_merged.factor.fdr05.ntag15.peaks.txt
${PEAKDIR}/PU1delQ_Flag_merged.factor.fdr05.ntag15.peaks.txt \
${PEAKDIR}/PU1delP_Flag_merged.factor.fdr05.ntag15.peaks.txt
${PEAKDIR}/PU1delAQP_Flag_merged.factor.fdr05.ntag15.peaks.txt > ${TMPDIR}/tmp.1.txt
pos2bed.pl ${TMPDIR}/tmp.1.txt > ${TMPDIR}/tmp.1.bed # 58571 peaks
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/PU1_allFlag.ntag15.merged.filtered.bed
bed2pos.pl ${PEAKDIR}/PU1_allFlag.ntag15.merged.filtered.bed >
${PEAKDIR}/PU1_allFlag.ntag15.merged.filtered.pos.txt
rm ${TMPDIR}/tmp.*
# 54749 peaks remaining

# BigWigs for normalized TagDirs
# generating individual bigwigs
TAGDIRSETS="CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr
CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr \
CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr CTV1_PU.1delA_Flag_R1_CNVnormRefChr
CTV1_PU.1delA_Flag_R2_CNVnormRefChr \
CTV1_PU.1delQ_Flag_R1_CNVnormRefChr CTV1_PU.1delQ_Flag_R2_CNVnormRefChr
CTV1_PU.1delP_Flag_R1_CNVnormRefChr \
CTV1_PU.1delP_Flag_R2_CNVnormRefChr CTV1_PU.1delAQP_Flag_R1_CNVnormRefChr
CTV1_PU.1delAQP_Flag_R2_CNVnormRefChr \
CTV1_PU1_Flag_0.9ug_CNVnormRefChr CTV1_PU1_Flag_3ug_CNVnormRefChr
CTV1_PU1_Flag_6ug_CNVnormRefChr"

for SAMPLE in ${TAGDIRSETS[@]}; do
makeUCSCfile ${TAGDIR}/$SAMPLE -bigWig $CHROMSIZES -o $BIGWIGDIR/$SAMPLE.bigwig
done

# average bigwigs from replicates
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_PU.1delA_Flag_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU.1delA_Flag_R2_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_aveFlag_PU1delA_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_PU.1delQ_Flag_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU.1delQ_Flag_R2_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_aveFlag_PU1delQ_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_PU.1delP_Flag_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU.1delP_Flag_R2_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_aveFlag_PU1delP_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_PU.1delAQP_Flag_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU.1delAQP_Flag_R2_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_aveFlag_PU1delAQP_CNVnormRefChr.bigwig

# analyses of replicates without background correction
# annotation of raw count
TAGDIRLIST="${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_3ug_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr ${TAGDIR}/CTV1_PU.1delP_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delP_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU.1delQ_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delQ_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU.1delA_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delA_Flag_R2_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delAQP_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delAQP_Flag_R2_CNVnormRefChr"
annotatePeaks.pl ${PEAKDIR}/PU1_allFlag.ntag15.merged.filtered.pos.txt hg19 -size 200 -d
```

```
${TAGDIRLIST} -raw -noann -nogene -cpu 12 > ${PEAKDIR}/PU1_allFlag.ntag15.rawAnn.txt

# rlog transformation of raw data with DEseq2
cd ${PEAKDIR}
declare -a NAMES=("PU1.rep1" "PU1.rep2" "PU1.rep3" "PU1.3ug" "PU1.6ug" "PU1.0.9ug"
"PU1delP.rep1" "PU1delP.rep2" "PU1delQ.rep1" "PU1delQ.rep2" "PU1delA.rep1" "PU1delA.rep2"
"PU1delAQP.rep1" "PU1delAQP.rep2")
declare -a TYPE=("wt" "wt" "wt" "wt" "wt" "wt" "P" "P" "Q" "Q" "A" "A" "AQP" "AQP")
TAB="$(printf '\t')"
cd ${PEAKDIR}
cat >"groups.tmp.txt" <<EOF
Experiment${TAB}CellType
EOF
COUNT=0
for SAMPLE in ${NAMES[@]}; do
GROUP="${TYPE[${COUNT}]}"
cat >>"groups.tmp.txt" <<EOF
${SAMPLE}${TAB}${GROUP}
EOF
COUNT=$((COUNT+=1))
Done

# DEseq2 rlog transformation
cd ${PEAKDIR}
_DATE=$(date +%s)
cat >"${TMPDIR}/R.rlog.${_DATE}.R" <<EOF
library(DESeq2)
#Read Data in
countData <- read.delim("${PEAKDIR}/PU1_allFlag.ntag15.rawAnn.txt")
countData[-1,-1] <- round(countData[-1,-1],0)
colData <- read.delim("groups.tmp.txt")
dds <- DESeqDataSetFromMatrix(countData, colData,design=~CellType,tidy=TRUE)
norm <- rlog(dds,blind=FALSE)
norm_matrix <- assay(norm)
norm_df <- data.frame(Gene=rownames(norm_matrix), norm_matrix)
write.table(norm_df, "${PEAKDIR}/PU1_allFlag.ntag15.rlogAnn.txt", row.names =
FALSE,sep="\t", quote=FALSE)
EOF
chmod 750 "${TMPDIR}/R.rlog.${_DATE}.R"
R < ${TMPDIR}/R.rlog.${_DATE}.R --no-save
rm ${TMPDIR}/R.rlog.${_DATE}.R
# change header
tail -n +2 ${PEAKDIR}/PU1_allFlag.ntag15.rlogAnn.txt >
${PEAKDIR}/tmp2.PU1_allFlag.ntag15.allAnn.txt
echo
$'ID\tPU1.rep1\tPU1.rep2\tPU1.rep3\tPU1.3ug\tPU1.6ug\tPU1.0.9ug\tPU1delP.rep1\tPU1delP.rep2\tP
U1delQ.rep1\tPU1delQ.rep2\tPU1delA.rep1\tPU1delA.rep2\tPU1delAQP.rep1\tPU1delAQP.rep2' \
| cat - ${PEAKDIR}/tmp2.PU1_allFlag.ntag15.allAnn.txt >
${PEAKDIR}/PU1_allFlag.ntag15.rlogAnn.txt
# cleanup
rm ${PEAKDIR}/tmp2.PU1_allFlag.ntag15.allAnn.txt

# tSNE for rlog transformed data
declare -a TABLE=("${PEAKDIR}/PU1_allFlag.ntag15.rlogAnn.txt")
declare -a OUTPUT=("${FIGURESDIR}/tSNE_PU1_allFlag.ntag15.rlogAnn.pdf")
declare -a EMBEDDING=("${PEAKDIR}/PU1_allFlag.ntag15.rlogAnn.tSNEembedding.txt")

# tSNE plot
COUNT=0
_DATE=$(date +%s)
for TABLE in ${TABLES[@]}; do
OUTPUT="${OUTPUT[${COUNT}]}"
cd ${PEAKDIR}
cat >"${TMPDIR}/R.PCA.${_DATE}.R" <<EOF
library(ggplot2)
library(ggrepel)
library(Rtsne)
logcpm <- read.delim("${TABLE}", row.names="ID")
celltype <- factor(c("PU1.1", "PU1.2", "PU1.3", "PU1.50%", "PU1.4", "PU1.15%", "PU1delP.1",
"PU1delP.2", "PU1delQ.1", "PU1delQ.2", "PU1delA.1", "PU1delA.2", "PU1delAQP.1",
"PU1delAQP.2"))
colors2 <-
c("${colPU1}","${colPU1}","${colPU1}","${colPU13u}","${colPU1}","${colPU1p9u}","${colDelP}",
"${colDelP}", "${colDelQ}", "${colDelQ}", "${colDelA}", "${colDelA}", "${colDelAQP}",
"${colDelAQP}")
mydata <- data.matrix(t(logcpm))
rtsne_out <- Rtsne(mydata, check_duplicates = FALSE, pca = TRUE, perplexity=4, theta=0.225,
dims=2, max_iter = 10000)
```

```
embedding <- as.data.frame(rtsne_out\$Y)
write.table(embedding, file = "${EMBEDDING}", sep = "\t", col.names=NA, quote=FALSE)
#embedding <- read.delim("${EMBEDDING}")
embedding\$Class <- as.factor(celltype)
embedding\$Color <- as.factor(colors2)
p <- ggplot(embedding, aes(x=V1, y=V2, label=celltype)) +
geom_point(size=2, col=embedding\$Color) +
geom_text_repel(aes(label=celltype), col=embedding\$Color, size=1.5,segment.size=0.2,
min.segment.length=0.2, point.padding=.15, segment.alpha=0.5) +
guides(colour = guide_legend(override.aes = list(size=5))) +
xlab("tSNE-X") + ylab("tSNE-Y") +
ggtitle("t-SNE 2D Embedding\nfor all Cell Types") +
#xlim(-200, 200) + ylim(-200, 200) +
theme_light(base_size=8) + theme(plot.title = element_text(size = 10, face = "bold"))
pdf(file="${OUTPUT}", height=2.5, width=2.3)
plot(p, labels=TRUE)
dev.off()
EOF
chmod 750 "${TMPDIR}/R.PCA.${_DATE}.R"
R < ${TMPDIR}/R.PCA.${_DATE}.R --no-save
rm ${TMPDIR}/R.PCA.${_DATE}.R
COUNT=$((COUNT+=1))
done

# detection of differential peaks
# test for WT vs others using all the initial peak set but all four replicates for 3ug RNA

# WT vs delP
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
-b ${TAGDIR}/CTV1_PU.1delP_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delP_Flag_R2_CNVnormRefChr -genome hg19 \
-p ${PU1CTV1PEAKS} > ${DIFFDIR}/WTvsDelP.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 44629
# Total Up-regulated in target vs. bg: 40 (0.090%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 1 (0.002%) [log2fold<-1, FDR<0.05]
# NOT different!

# WT vs delQ
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
-b ${TAGDIR}/CTV1_PU.1delQ_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delQ_Flag_R2_CNVnormRefChr -genome hg19 \
-p ${PU1CTV1PEAKS} > ${DIFFDIR}/WTvsDelQ.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 44629
# Total Up-regulated in target vs. bg: 23249 (52.094%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 64 (0.143%) [log2fold<-1, FDR<0.05]

# WT vs delA
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
-b ${TAGDIR}/CTV1_PU.1delA_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delA_Flag_R2_CNVnormRefChr -genome hg19 \
-p ${PU1CTV1PEAKS} > ${DIFFDIR}/WTvsDelA.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 44629
# Total Up-regulated in target vs. bg: 32005 (71.713%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 10 (0.022%) [log2fold<-1, FDR<0.05]

# WT vs delAQP
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
-b ${TAGDIR}/CTV1_PU.1delAQP_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delAQP_Flag_R2_CNVnormRefChr -genome hg19 \
-p ${PU1CTV1PEAKS} > ${DIFFDIR}/WTvsDelAQP.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 44629
# Total Up-regulated in target vs. bg: 34414 (77.111%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 0 (0.000%) [log2fold<-1, FDR<0.05]

# WT vs 3ug (50%)
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
-b ${TAGDIR}/CTV1_PU1_Flag_3ug_CNVnormRefChr -genome hg19 \
-p ${PU1CTV1PEAKS} > ${DIFFDIR}/WTvshalfPU1.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 44629
# Total Up-regulated in target vs. bg: 0 (0.000%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 0 (0.000%) [log2fold<-1, FDR<0.05]

# WT vs 0.9ug (15%)
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
-b ${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr -genome hg19 \
-p ${PU1CTV1PEAKS} > ${DIFFDIR}/WTvsLessPU1.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 44629
# Total Up-regulated in target vs. bg: 13001 (29.131%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 0 (0.000%) [log2fold<-1, FDR<0.05]

# delP vs WT
pos2bed.pl ${PEAKDIR}/PU1delP_Flag_merged.factor.fdr05.ntag15.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/PU1delP_Flag_merged.ntag15.filtered.bed
bed2pos.pl ${PEAKDIR}/PU1delP_Flag_merged.ntag15.filtered.bed >
${PEAKDIR}/PU1delP_Flag_merged.ntag15.filtered.pos.txt
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -b ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
-t ${TAGDIR}/CTV1_PU.1delP_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delP_Flag_R2_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/PU1delP_Flag_merged.ntag15.filtered.pos.txt > ${DIFFDIR}/DelPvsWT.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 35176
# Total Up-regulated in target vs. bg: 13 (0.037%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 9 (0.026%) [log2fold<-1, FDR<0.05]
# NOT different!

# delQ vs WT
pos2bed.pl ${PEAKDIR}/PU1delQ_Flag_merged.factor.fdr05.ntag15.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/PU1delQ_Flag_merged.ntag15.filtered.bed
bed2pos.pl ${PEAKDIR}/PU1delQ_Flag_merged.ntag15.filtered.bed >
${PEAKDIR}/PU1delQ_Flag_merged.ntag15.filtered.pos.txt
getDifferentialPeaksReplicates.pl -b ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
-t ${TAGDIR}/CTV1_PU.1delQ_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delQ_Flag_R2_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/PU1delQ_Flag_merged.ntag15.filtered.pos.txt > ${DIFFDIR}/DelQvsWT.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 19741
# Total Up-regulated in target vs. bg: 757 (3.835%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 2340 (11.854%) [log2fold<-1, FDR<0.05]

# delQ vs delA
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -b ${TAGDIR}/CTV1_PU.1delA_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delA_Flag_R2_CNVnormRefChr \
-t ${TAGDIR}/CTV1_PU.1delQ_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delQ_Flag_R2_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/PU1delQ_Flag_merged.ntag15.filtered.pos.txt > ${DIFFDIR}/DelQvsDelA.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 19741
# Total Up-regulated in target vs. bg: 4246 (21.509%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 0 (0.000%) [log2fold<-1, FDR<0.05]
# delQ vs 0.9ug
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -b ${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr \
-t ${TAGDIR}/CTV1_PU.1delQ_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delQ_Flag_R2_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/PU1delQ_Flag_merged.ntag15.filtered.pos.txt > ${DIFFDIR}/DelQvsLessPU1.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 19741
# Total Up-regulated in target vs. bg: 0 (0.000%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 0 (0.000%) [log2fold<-1, FDR<0.05]
#delA vs WT
pos2bed.pl ${PEAKDIR}/PU1delA_Flag_merged.factor.fdr05.ntag15.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/PU1delA_Flag_merged.ntag15.filtered.bed
bed2pos.pl ${PEAKDIR}/PU1delA_Flag_merged.ntag15.filtered.bed >
${PEAKDIR}/PU1delA_Flag_merged.ntag15.filtered.pos.txt
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -b ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
-t ${TAGDIR}/CTV1_PU.1delA_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delA_Flag_R2_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/PU1delA_Flag_merged.ntag15.filtered.pos.txt > ${DIFFDIR}/DelAvsWT.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 11569
# Total Up-regulated in target vs. bg: 359 (3.103%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 3313 (28.637%) [log2fold<-1, FDR<0.05]

# delA vs delQ
cd ${PEAKDIR}
getDifferentialPeaksReplicates.pl -t ${TAGDIR}/CTV1_PU.1delA_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delA_Flag_R2_CNVnormRefChr \
-b ${TAGDIR}/CTV1_PU.1delQ_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delQ_Flag_R2_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/PU1delA_Flag_merged.ntag15.filtered.pos.txt > ${DIFFDIR}/DelAvsDelQ.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 11569
# Total Up-regulated in target vs. bg: 1 (0.009%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 1353 (11.695%) [log2fold<-1, FDR<0.05]

# delAQP vs WT
pos2bed.pl ${PEAKDIR}/PU1delAQP_Flag_merged.factor.fdr05.ntag15.peaks.txt >
${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/PU1delAQP_Flag_merged.ntag15.filtered.bed
bed2pos.pl ${PEAKDIR}/PU1delAQP_Flag_merged.ntag15.filtered.bed >
${PEAKDIR}/PU1delAQP_Flag_merged.ntag15.filtered.pos.txt
getDifferentialPeaksReplicates.pl -b ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
-t ${TAGDIR}/CTV1_PU.1delAQP_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delAQP_Flag_R2_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/PU1delAQP_Flag_merged.ntag15.filtered.pos.txt > ${DIFFDIR}/DelAQPvsWT.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 7560
# Total Up-regulated in target vs. bg: 0 (0.000%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 4160 (55.026%) [log2fold<-1, FDR<0.05]
# no peaks enriched

# LessPU1 vs WT
${PEAKDIR}/PU1_Flag_0.9ug.factor.fdr05.ntag15.peaks.txt
pos2bed.pl ${PEAKDIR}/PU1_Flag_0.9ug.factor.fdr05.ntag15.peaks.txt > ${TMPDIR}/tmp.1.bed
$BEDTOOLS intersect -a ${TMPDIR}/tmp.1.bed -b $BLACKLIST_HG19 -v > ${TMPDIR}/tmp.2.bed
filter4Mappability.sh -p ${TMPDIR}/tmp.2.bed -g hg19 -f 0.8 -s 50
pos2bed.pl ${TMPDIR}/tmp.2.mapScoreFiltered.txt >
${PEAKDIR}/PU1_Flag_0.9ug.ntag15.filtered.bed
bed2pos.pl ${PEAKDIR}/PU1_Flag_0.9ug.ntag15.filtered.bed >
${PEAKDIR}/PU1_Flag_0.9ug.ntag15.filtered.pos.txt
getDifferentialPeaksReplicates.pl -b ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
-t ${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/PU1_Flag_0.9ug.ntag15.filtered.pos.txt > ${DIFFDIR}/LessPU1vsWT.peaks.txt
# Output Stats bg vs. target:
```

```
# Total Genes: 16198
# Total Up-regulated in target vs. bg: 0 (0.000%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 1282 (7.915%) [log2fold<-1, FDR<0.05]

# LessPU1 vs delQ
getDifferentialPeaksReplicates.pl -b ${TAGDIR}/CTV1_PU.1delQ_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delQ_Flag_R2_CNVnormRefChr \
-t ${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr -genome hg19 \
-p ${PEAKDIR}/PU1_Flag_0.9ug.ntag15.filtered.pos.txt > ${DIFFDIR}/LessPU1vsWT.peaks.txt
# Output Stats bg vs. target:
# Total Genes: 16198
# Total Up-regulated in target vs. bg: 20 (0.123%) [log2fold>1, FDR<0.05]
# Total Dn-regulated in target vs. bg: 0 (0.000%) [log2fold<-1, FDR<0.05]
# not much different when we only have one replicate

# Comparison between differential peak sets
cd ${DIFFDIR}
mergePeaks -d 100 WTvsLessPU1.peaks.txt WTvsDelQ.peaks.txt WTvsDelA.peaks.txt
DelQvsDelA.peaks.txt -venn WTenrichedVsDelADelQLessPU1.venn.txt >
WTenrichedVsDelADelQLessPU1.pos.txt
awk 'BEGIN{OFS="\t"} $7 == "DelQvsDelA.peaks.txt" {print $1, $2, $3, $4, $5, $6}'
WTenrichedVsDelADelQLessPU1.pos.txt > DelQvsDelA.specific.peaks.txt
awk 'BEGIN{OFS="\t"} $7 == "WTvsDelA.peaks.txt" {print $1, $2, $3, $4, $5, $6}'
WTenrichedVsDelADelQLessPU1.pos.txt > WTvsDelA.specific.peaks.txt
awk 'BEGIN{OFS="\t"}$7 == "WTvsLessPU1.peaks.txt|WTvsDelQ.peaks.txt|WTvsDelA.peaks.txt"
{print $1, $2, $3, $4, $5, $6}' WTenrichedVsDelADelQLessPU1.pos.txt >
WTenrichedVsDelADelQLessPU1.specific.peaks.txt

# characterization of differential peaks

# delA vs WT
# histograms across differentially bound regions
# generating necessary ghist files to create the plots
declare -a MERGEDTAGDIR=("${TAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delA_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr" "${TAGDIR}/CTV1_PU1_Flag_3ug_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_H3K27ac_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged"
"${ATACDIR}/CTV1_PU1_CNVnormRefChr_merged" "${ATACDIR}/CTV1_PU1mut_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr")

for SET in "${MERGEDTAGDIR[@]}"
do
NAME=${SET##*/}
_DATE=$(date +%s)
cat >"${TMPDIR}/ghist.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
if [ -e ${DIFFDIR}/WTvsDelA.peaks.txt ] ; then
annotatePeaks.pl ${DIFFDIR}/WTvsDelA.peaks.txt hg19 -size 2000 -hist 25 -ghist -d ${SET} >
${GHISTFILEDIR}/WTvsDelA.${NAME}.ann.ghist.txt ; fi
if [ -e ${DIFFDIR}/DelAvsWT.peaks.txt ] ; then
annotatePeaks.pl ${DIFFDIR}/DelAvsWT.peaks.txt hg19 -size 2000 -hist 25 -ghist -d ${SET} >
${GHISTFILEDIR}/DelAvsWT.${NAME}.ann.ghist.txt ; fi
EOF
chmod 750 "${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
echo "generating ghist for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
done
```

```
# loop to check when screen sessions are done
#--------------------------------------------
for SET in "${MERGEDTAGDIR[@]}" ; do
NAME=${SET##*/}
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

# generate histogram plots
# WTvsDelA
cd ${GHISTFILEDIR}
# PU.1
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelA.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1_Flag_CNVnormRefChr_merged.ann.ghist.txt" \
-s "mutPU1 PU1delA PU1" -c "${colMUT} ${colDelA} ${colPU1}" -x 1000 -y "0 25" -d
${FIGURESDIR}/hist -n WTvsDelA.PU1
# ATAC & H3K27ac
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelA.CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1mut_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1_CNVnormRefChr_merged.ann.ghist.txt" \
-s "H3K27ac H3K27ac-PU1 ATAC ATAC-PU1" -c "cyan3 darkcyan ${colMUT} ${colPU1}" -x 1000 -y "0
8" -d ${FIGURESDIR}/hist -n WTvsDelA.H3K27ac-ATAC
# BRGi
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelA.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delA PU1-DMSO PU1-1uM PU1-5uM PU1-10uM" -c "${colMUT} ${colDelA} ${colPU1}
orchid3 palevioletred2 plum1" -x 1000 -y "0 20" -d ${FIGURESDIR}/hist -n WTvsDelA.BRGi
# titration
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelA.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1_Flag_6ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1_Flag_3ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelA.CTV1_PU1_Flag_0.9ug_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delA PU1-100% PU1-50% PU1-15%" -c "${colMUT} ${colDelA} ${colPU16u}
${colPU13u} ${colPU1p9u}" -x 1000 -y "0 20" -d ${FIGURESDIR}/hist -n WTvsDelA.Titration

# DelAvsWT
cd ${GHISTFILEDIR}
# PU.1
plotHIST.sh \
-f "${GHISTFILEDIR}/DelAvsWT.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1_Flag_CNVnormRefChr_merged.ann.ghist.txt" \
-s "mutPU1 PU1delA PU1" -c "${colMUT} ${colDelA} ${colPU1}" -x 1000 -y "0 12" -d
${FIGURESDIR}/hist -n DelAvsWT.PU1
# ATAC & H3K27ac
plotHIST.sh \
-f "${GHISTFILEDIR}/DelAvsWT.CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1mut_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1_CNVnormRefChr_merged.ann.ghist.txt" \
-s "H3K27ac H3K27ac-PU1 ATAC ATAC-PU1" -c "cyan3 darkcyan ${colMUT} ${colPU1}" -x 1000 -y "0
25" -d ${FIGURESDIR}/hist -n DelAvsWT.H3K27ac-ATAC
# BRGi
plotHIST.sh \
-f "${GHISTFILEDIR}/DelAvsWT.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr.ann.ghist.txt
```

```
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delA PU1-DMSO PU1-1uM PU1-5uM PU1-10uM" -c "${colMUT} ${colDelA} ${colPU1}
orchid3 palevioletred2 plum1" -x 1000 -y "0 12" -d ${FIGURESDIR}/hist -n DelAvsWT.BRGi
# titration
plotHIST.sh \
-f "${GHISTFILEDIR}/DelAvsWT.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1_Flag_6ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1_Flag_3ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelAvsWT.CTV1_PU1_Flag_0.9ug_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delA PU1-100% PU1-50% PU1-15%" -c "${colMUT} ${colDelA} ${colPU16u}
${colPU13u} ${colPU1p9u}" -x 1000 -y "0 12" -d ${FIGURESDIR}/hist -n DelAvsWT.Titration

# motif searches in differentially bound regions
cd ${WORKDIR}
findMotifsGenome.pl ${DIFFDIR}/WTvsDelA.peaks.txt hg19r ${MOTIFDIR}/WTvsDelA.peaks -size 200
-len 7,8,9,10,11,12,13,14 -p 12 -h
findMotifsGenome.pl ${DIFFDIR}/DelAvsWT.peaks.txt hg19r ${MOTIFDIR}/DelAvsWT.peaks -size 200
-len 7,8,9,10,11,12,13,14 -p 12 -h

# reformating output with own parameters
compareMotifs.pl ${MOTIFDIR}/DelAvsWT.peaks/homerMotifs.all.motifs
${MOTIFDIR}/DelAvsWT.peaks/final -reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info 1.5
-cpu 12
compareMotifs.pl ${MOTIFDIR}/WTvsDelA.peaks/homerMotifs.all.motifs
${MOTIFDIR}/WTvsDelA.peaks/final -reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info 1.5
-cpu 12

# genome ontology
cd ${WORKDIR}
annotatePeaks.pl ${DIFFDIR}/WTvsDelA.peaks.txt hg19 -size given -gtf ${HG19GTF} -annStats
${DIFFDIR}/WTvsDelA.peaks.genome.stats.txt > ${DIFFDIR}/tmp.4.WTvsDelA.peaks.txt
annotatePeaks.pl ${DIFFDIR}/DelAvsWT.peaks.txt hg19 -size given -gtf ${HG19GTF} -annStats
${DIFFDIR}/DelAvsWT.peaks.genome.stats.txt > ${DIFFDIR}/tmp.4.DelAvsWT.peaks.txt
head -n 6 ${DIFFDIR}/WTvsDelA.peaks.genome.stats.txt | cut -f 1,2 > ${TMPDIR}/tmp1.txt
head -n 6 ${DIFFDIR}/DelAvsWT.peaks.genome.stats.txt | cut -f 1,2 > ${TMPDIR}/tmp3.txt
cd ${WORKDIR}
cat >"${FIGURESDIR}/R/R.piechartWTvsDelA.R" <<EOF
library(plotrix)
pdf(file="${FIGURESDIR}/genomeOntologyPieChartWTvsDelA.pdf", height=4, width=8.5)
par(fig=c(0,0.5,0,1))
d <- read.table("${TMPDIR}/tmp1.txt", header=T, sep="\t")
data <- d[c(2,5,4,1,3),]
colnames(data) <- c("lbls","slices")
slices <- data\$slices
lbls <- data\$lbls
colors = c("red","blue","yellow","black","green")
pct <- round(slices/sum(slices)*100, digits = 1)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=colors, main="WT vs. delA")
par(fig=c(0.5,1,0,1), new=TRUE)
d3 <- read.table("${TMPDIR}/tmp3.txt", header=T, sep="\t")
data3 <- d3[c(2,5,4,1,3),]
colnames(data3) <- c("lbls","slices")
slices <- data3\$slices
lbls <- data3\$lbls
colors = c("red","blue","yellow","black","green")
pct <- round(slices/sum(slices)*100, digits = 1)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=colors, main="delA vs WT")
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.piechartWTvsDelA.R"
R < "${FIGURESDIR}/R/R.piechartWTvsDelA.R" --no-save

# motif score distribution
annotatePeaks.pl ${DIFFDIR}/WTvsDelA.peaks.txt hg19 -size 200 -m ${PU1MOTIF} -mscore -nogene
-noann > ${TMPDIR}/tmp.5.WTvsDelA.motifScore.txt
annotatePeaks.pl ${DIFFDIR}/DelAvsWT.peaks.txt hg19 -size 200 -m ${PU1MOTIF} -mscore -nogene
-noann > ${TMPDIR}/tmp.5.DelAvsWT.motifScore.txt

# beanplot (without stringent peaks)
cd ${WORKDIR}
cat >"${FIGURESDIR}/R/R.motifScores.WTvsDelA.R" <<EOF
library(beanplot)
```

```
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.5.WTvsDelA.motifScore.txt", header=T, sep="\t")
data3 <- read.table("${TMPDIR}/tmp.5.DelAvsWT.motifScore.txt", header=T, sep="\t")
# combining the data columns
a <- data1[-1,10]
c <- data3[-1,10]
z <- c("a", "c")
x <- lapply(z, get, envir=environment())
names(x) <- z
# determining the length of each column
anum <- nrow(as.matrix(a))
cnum <- nrow(as.matrix(c))
# define labels
laba <- paste("WT (",anum,")",sep="")
labc <- paste("delA (",cnum,")",sep="")
# statistical testing
wilac <- wilcox.test(a,c)
if (wilac\$p.value < 0.001) {
wac <- "***"
} else if (wilac\$p.value < 0.01) {
wac <- "**"
} else if (wilac\$p.value < 0.05) {
wac <- "*"
} else {
wac <- "ns"
}
# defining colors
beancol <- list("${colPU1}","${colDelA}")
boxcol <- c("gray65","gray85")
# creating the plot
pdf(file="${FIGURESDIR}/motifScores.WTvsDelA.pdf", height=3, width=1.6)
par(mar=c(4.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(2,13),lwd=0.5)
# adding box plot on top
par(mar=c(4.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(2,13),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:2, labels=c(laba,
labc),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext(" Motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
#mtext("Cluster",family="Helvetica",ps=12,side=1,line=2,cex=1.2,padj=2.0)
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
# adding significance levels
text(1.5,12.5,wac)
segments(1,12.3,2,12.3)
EOF
chmod 750 "${FIGURESDIR}/R/R.motifScores.WTvsDelA.R"
R < "${FIGURESDIR}/R/R.motifScores.WTvsDelA.R" --no-save

# position of differential peaks in K14CLUSTER
pos2bed.pl ${DIFFDIR}/WTvsDelA.peaks.txt > ${DIFFDIR}/WTvsDelA.peaks.bed
$BEDTOOLS intersect -a ${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed -b
${DIFFDIR}/WTvsDelA.peaks.bed -c > ${TMPDIR}/WTvsDelA.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}'
${TMPDIR}/WTvsDelA.peaks.bound.bed
${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.WTvsDelA.sorted.txt

# "ghistplot" for presence of paired motifs
_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.WTvsDelA.sorted.txt", header=T)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
mycol <- colorRampPalette(c("white","${colDelA}"))(3)
png(filename="WTvsDelAinK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
EOF
```

```
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# "ghistplot" for ratio of WTvsDelA
annotatePeaks.pl ${PU1CTV1PEAKS} hg19 -size 200 -d ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
${TAGDIR}/CTV1_PU.1delA_Flag_R1_CNVnormRefChr ${TAGDIR}/CTV1_PU.1delA_Flag_R2_CNVnormRefChr
-nogene -noann >${DIFFDIR}/WT_DelAann.txt
awk 'BEGIN{OFS="\t"} {print $1,(($8+$9+$10+$11+2)/2)/($12+$13+1)}' ${DIFFDIR}/WT_DelAann.txt
> ${TMPDIR}/WT_DelAratio.txt
awk 'BEGIN{OFS="\t"} NR==FNR{a[$1]=$2;next}{print $0,a[$1]}' ${TMPDIR}/WT_DelAratio.txt
${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/WT_DelAratio.sorted.txt

_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/WT_DelAratio.sorted.txt", header=T)
head(data, n=50)
d <- log2(data.matrix(data[,7]))
q <- apply(d, 2, function(x) ifelse(x > 2, 2 ,x))
p <- apply(q, 2, function(x) ifelse(x < -2, -2 ,x))
o <- p[order(nrow(q):1),]
mycol <- colorRampPalette(c("dodgerblue3","floralwhite","firebrick2"))(4)
png(filename="WTvsDelAinK14cluster_ratio.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(-2,2)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# delAQP vs WT

# histograms across differentially bound regions
# generating necessary ghist files to create the plots
declare -a MERGEDTAGDIR=("${TAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delAQP_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr" "${TAGDIR}/CTV1_PU1_Flag_3ug_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_H3K27ac_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged"
"${ATACDIR}/CTV1_PU1_CNVnormRefChr_merged" "${ATACDIR}/CTV1_PU1mut_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr")

for SET in "${MERGEDTAGDIR[@]}"
do
NAME=${SET##*/}
_DATE=$(date +%s)
cat >"${TMPDIR}/ghist.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
if [ -e ${DIFFDIR}/WTvsDelAQP.peaks.txt ] ; then
annotatePeaks.pl ${DIFFDIR}/WTvsDelAQP.peaks.txt hg19 -size 2000 -hist 25 -ghist -d ${SET} >
${GHISTFILEDIR}/WTvsDelAQP.${NAME}.ann.ghist.txt ; fi
EOF
chmod 750 "${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
echo "generating ghist for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
done
```

```
# loop to check when screen sessions are done
#--------------------------------------------
for SET in "${MERGEDTAGDIR[@]}" ; do
NAME=${SET##*/}
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

# generate histogram plots
# WTvsDelAQP
cd ${GHISTFILEDIR}
# PU.1
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1delAQP_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1_Flag_CNVnormRefChr_merged.ann.ghist.txt" \
-s "mutPU1 PU1delAQP PU1" -c "${colMUT} ${colDelAQP} ${colPU1}" -x 1000 -y "0 25" -d
${FIGURESDIR}/hist -n WTvsDelAQP.PU1
# ATAC & H3K27ac
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1mut_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1_CNVnormRefChr_merged.ann.ghist.txt" \
-s "H3K27ac H3K27ac-PU1 ATAC ATAC-PU1" -c "cyan3 darkcyan ${colMUT} ${colPU1}" -x 1000 -y "0
10" -d ${FIGURESDIR}/hist -n WTvsDelAQP.H3K27ac-ATAC
# BRGi
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1delAQP_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delAQP PU1-DMSO PU1-1uM PU1-5uM PU1-10uM" -c "${colMUT} ${colDelAQP} ${colPU1}
orchid3 palevioletred2 plum1" -x 1000 -y "0 20" -d ${FIGURESDIR}/hist -n WTvsDelAQP.BRGi
# titration
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1delAQP_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1_Flag_6ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1_Flag_3ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelAQP.CTV1_PU1_Flag_0.9ug_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delAQP PU1-100% PU1-50% PU1-15%" -c "${colMUT} ${colDelAQP} ${colPU16u}
${colPU13u} ${colPU1p9u}" -x 1000 -y "0 20" -d ${FIGURESDIR}/hist -n WTvsDelAQP.Titration

# motif searches in differentially bound regions
cd ${WORKDIR}
findMotifsGenome.pl ${DIFFDIR}/WTvsDelAQP.peaks.txt hg19r ${MOTIFDIR}/WTvsDelAQP.peaks -size
200 -len 7,8,9,10,11,12,13,14 -p 12 -h

# reformating output with own parameters
compareMotifs.pl ${MOTIFDIR}/WTvsDelAQP.peaks/homerMotifs.all.motifs
${MOTIFDIR}/WTvsDelAQP.peaks/final -reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info 1.5
-cpu 12

# genome ontology
cd ${WORKDIR}
annotatePeaks.pl ${DIFFDIR}/WTvsDelAQP.peaks.txt hg19 -size given -gtf ${HG19GTF} -annStats
${DIFFDIR}/WTvsDelAQP.peaks.genome.stats.txt > ${DIFFDIR}/tmp.4.WTvsDelAQP.peaks.txt
head -n 6 ${DIFFDIR}/WTvsDelAQP.peaks.genome.stats.txt | cut -f 1,2 > ${TMPDIR}/tmp1.txt
cd ${WORKDIR}
cat >"${FIGURESDIR}/R/R.piechartWTvsDelAQP.R" <<EOF
library(plotrix)
pdf(file="${FIGURESDIR}/genomeOntologyPieChartWTvsDelAQP.pdf", height=4, width=4.5)
d <- read.table("${TMPDIR}/tmp1.txt", header=T, sep="\t")
data <- d[c(2,5,4,1,3),]
colnames(data) <- c("lbls","slices")
slices <- data\$slices
lbls <- data\$lbls
```

```
colors = c("red","blue","yellow","black","green")
pct <- round(slices/sum(slices)*100, digits = 1)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=colors, main="WT vs. delAQP")
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.piechartWTvsDelAQP.R"
R < "${FIGURESDIR}/R/R.piechartWTvsDelAQP.R" --no-save

# motif score distribution
annotatePeaks.pl ${DIFFDIR}/WTvsDelAQP.peaks.txt hg19 -size 200 -m ${PU1MOTIF} -mscore
-nogene -noann > ${TMPDIR}/tmp.5.WTvsDelAQP.motifScore.txt

# beanplot (without stringent peaks)
cd ${WORKDIR}
cat >"${FIGURESDIR}/R/R.motifScores.WTvsDelAQP.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.5.WTvsDelAQP.motifScore.txt", header=T, sep="\t")
x <- data1[-1,10]
# determining the length of each column
anum <- nrow(as.matrix(x))
# define labels
laba <- paste("WT (",anum,")",sep="")
beancol <- "${colPU1}"
boxcol <- "gray65"
# creating the plot
pdf(file="${FIGURESDIR}/motifScores.WTvsDelAQP.pdf", height=3, width=1.2)
par(mar=c(4.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(2,13),lwd=0.5)
# adding box plot on top
par(mar=c(4.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.5, staplewex = 0.5, ylim=c(2,13),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:1, labels=laba,las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext(" Motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.motifScores.WTvsDelAQP.R"
R < "${FIGURESDIR}/R/R.motifScores.WTvsDelAQP.R" --no-save

# position of differential peaks in K14CLUSTER
pos2bed.pl ${DIFFDIR}/WTvsDelAQP.peaks.txt > ${DIFFDIR}/WTvsDelAQP.peaks.bed
$BEDTOOLS intersect -a ${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed -b
${DIFFDIR}/WTvsDelAQP.peaks.bed -c > ${TMPDIR}/WTvsDelAQP.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}'
${TMPDIR}/WTvsDelAQP.peaks.bound.bed
${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.WTvsDelAQP.sorted.txt

# "ghistplot" for presence of paired motifs
_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.WTvsDelAQP.sorted.txt", header=T)
head(data, n=50)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
mycol <- colorRampPalette(c("white","${colDelAQP}"))(3)
png(filename="WTvsDelAQPinK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R
```

```
# "ghistplot" for ratio of WTvsDelAQP
annotatePeaks.pl ${PU1CTV1PEAKS} hg19 -size 200 -d ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
${TAGDIR}/CTV1_PU.1delAQP_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU.1delAQP_Flag_R2_CNVnormRefChr -nogene -noann >${DIFFDIR}/WT_DelAQPann.txt
awk 'BEGIN{OFS="\t"} {print $1,(($8+$9+$10+$11+2)/2)/($12+$13+1)}'
${DIFFDIR}/WT_DelAQPann.txt > ${TMPDIR}/WT_DelAQPratio.txt
awk 'BEGIN{OFS="\t"} NR==FNR{a[$1]=$2;next}{print $0,a[$1]}' ${TMPDIR}/WT_DelAQPratio.txt
${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/WT_DelAQPratio.sorted.txt
_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/WT_DelAQPratio.sorted.txt", header=T)
head(data, n=50)
d <- log2(data.matrix(data[,7]))
q <- apply(d, 2, function(x) ifelse(x > 2, 2 ,x))
p <- apply(q, 2, function(x) ifelse(x < -2, -2 ,x))
o <- p[order(nrow(q):1),]
mycol <- colorRampPalette(c("dodgerblue3","floralwhite","firebrick2"))(4)
png(filename="WTvsDelAQPinK14cluster_ratio.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(-2,2)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# delQ vs WT

# histograms across differentially bound regions
# generating necessary ghist files to create the plots
declare -a MERGEDTAGDIR=("${TAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delQ_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr" "${TAGDIR}/CTV1_PU1_Flag_3ug_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_H3K27ac_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged"
"${ATACDIR}/CTV1_PU1_CNVnormRefChr_merged" "${ATACDIR}/CTV1_PU1mut_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr")

for SET in "${MERGEDTAGDIR[@]}"
do
NAME=${SET##*/}
_DATE=$(date +%s)
cat >"${TMPDIR}/ghist.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
if [ -e ${DIFFDIR}/WTvsDelQ.peaks.txt ] ; then
annotatePeaks.pl ${DIFFDIR}/WTvsDelQ.peaks.txt hg19 -size 2000 -hist 25 -ghist -d ${SET} >
${GHISTFILEDIR}/WTvsDelQ.${NAME}.ann.ghist.txt ; fi
if [ -e ${DIFFDIR}/DelQvsWT.peaks.txt ] ; then
annotatePeaks.pl ${DIFFDIR}/DelQvsWT.peaks.txt hg19 -size 2000 -hist 25 -ghist -d ${SET} >
${GHISTFILEDIR}/DelQvsWT.${NAME}.ann.ghist.txt ; fi
EOF
chmod 750 "${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
echo "generating ghist for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
done
```

```
# loop to check when screen sessions are done
#--------------------------------------------
for SET in "${MERGEDTAGDIR[@]}" ; do
NAME=${SET##*/}
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#--------------------------------------------

# generate histogram plots
# WTvsDelQ
cd ${GHISTFILEDIR}
# PU.1
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1_Flag_CNVnormRefChr_merged.ann.ghist.txt" \
-s "mutPU1 PU1delQ PU1" -c "${colMUT} ${colDelQ} ${colPU1}" -x 1000 -y "0 20" -d
${FIGURESDIR}/hist -n WTvsDelQ.PU1
# ATAC & H3K27ac
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1mut_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1_CNVnormRefChr_merged.ann.ghist.txt" \
-s "H3K27ac H3K27ac-PU1 ATAC ATAC-PU1" -c "cyan3 darkcyan ${colMUT} ${colPU1}" -x 1000 -y "0
5" -d ${FIGURESDIR}/hist -n WTvsDelQ.H3K27ac-ATAC
# BRGi
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delQ PU1-DMSO PU1-1uM PU1-5uM PU1-10uM" -c "${colMUT} ${colDelQ} ${colPU1}
orchid3 palevioletred2 plum1" -x 1000 -y "0 20" -d ${FIGURESDIR}/hist -n WTvsDelQ.BRGi
# titration
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1_Flag_6ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1_Flag_3ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsDelQ.CTV1_PU1_Flag_0.9ug_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delQ PU1-100% PU1-50% PU1-15%" -c "${colMUT} ${colDelQ} ${colPU16u}
${colPU13u} ${colPU1p9u}" -x 1000 -y "0 15" -d ${FIGURESDIR}/hist -n WTvsDelQ.Titration
# DelQvsWT
cd ${GHISTFILEDIR}
# PU.1
plotHIST.sh \
-f "${GHISTFILEDIR}/DelQvsWT.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1_Flag_CNVnormRefChr_merged.ann.ghist.txt" \
-s "mutPU1 PU1delQ PU1" -c "${colMUT} ${colDelQ} ${colPU1}" -x 1000 -y "0 15" -d
${FIGURESDIR}/hist -n DelQvsWT.PU1
# ATAC & H3K27ac
plotHIST.sh \
-f "${GHISTFILEDIR}/DelQvsWT.CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1mut_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1_CNVnormRefChr_merged.ann.ghist.txt" \
-s "H3K27ac H3K27ac-PU1 ATAC ATAC-PU1" -c "cyan3 darkcyan ${colMUT} ${colPU1}" -x 1000 -y "0
30" -d ${FIGURESDIR}/hist -n DelQvsWT.H3K27ac-ATAC
# BRGi
plotHIST.sh \
-f "${GHISTFILEDIR}/DelQvsWT.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delQ PU1-DMSO PU1-1uM PU1-5uM PU1-10uM" -c "${colMUT} ${colDelQ} ${colPU1}
orchid3 palevioletred2 plum1" -x 1000 -y "0 15" -d ${FIGURESDIR}/hist -n DelQvsWT.BRGi
# titration
plotHIST.sh \
-f "${GHISTFILEDIR}/DelQvsWT.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1_Flag_6ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1_Flag_3ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelQvsWT.CTV1_PU1_Flag_0.9ug_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delQ PU1-100% PU1-50% PU1-15%" -c "${colMUT} ${colDelQ} ${colPU16u}
${colPU13u} ${colPU1p9u}" -x 1000 -y "0 15" -d ${FIGURESDIR}/hist -n DelQvsWT.Titration

# motif searches in differentially bound regions
cd ${WORKDIR}
findMotifsGenome.pl ${DIFFDIR}/WTvsDelQ.peaks.txt hg19r ${MOTIFDIR}/WTvsDelQ.peaks -size 200
-len 7,8,9,10,11,12,13,14 -p 12 -h
findMotifsGenome.pl ${DIFFDIR}/DelQvsWT.peaks.txt hg19r ${MOTIFDIR}/DelQvsWT.peaks -size 200
-len 7,8,9,10,11,12,13,14 -p 12 -h

# reformating output with own parameters
compareMotifs.pl ${MOTIFDIR}/DelQvsWT.peaks/homerMotifs.all.motifs
${MOTIFDIR}/DelQvsWT.peaks/final -reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info 1.5
-cpu 12
compareMotifs.pl ${MOTIFDIR}/WTvsDelQ.peaks/homerMotifs.all.motifs
${MOTIFDIR}/WTvsDelQ.peaks/final -reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info 1.5
-cpu 12

# genome ontology
cd ${WORKDIR}
annotatePeaks.pl ${DIFFDIR}/WTvsDelQ.peaks.txt hg19 -size given -gtf ${HG19GTF} -annStats
${DIFFDIR}/WTvsDelQ.peaks.genome.stats.txt > ${DIFFDIR}/tmp.4.WTvsDelQ.peaks.txt
annotatePeaks.pl ${DIFFDIR}/DelQvsWT.peaks.txt hg19 -size given -gtf ${HG19GTF} -annStats
${DIFFDIR}/DelQvsWT.peaks.genome.stats.txt > ${DIFFDIR}/tmp.4.DelQvsWT.peaks.txt
head -n 6 ${DIFFDIR}/WTvsDelQ.peaks.genome.stats.txt | cut -f 1,2 > ${TMPDIR}/tmp1.txt
head -n 6 ${DIFFDIR}/DelQvsWT.peaks.genome.stats.txt | cut -f 1,2 > ${TMPDIR}/tmp3.txt
cd ${WORKDIR}
cat >"${FIGURESDIR}/R/R.piechartWTvsDelQ.R" <<EOF
library(plotrix)
pdf(file="${FIGURESDIR}/genomeOntologyPieChartWTvsDelQ.pdf", height=4, width=8.5)
par(fig=c(0,0.5,0,1))
d <- read.table("${TMPDIR}/tmp1.txt", header=T, sep="\t")
data <- d[c(2,5,4,1,3),]
colnames(data) <- c("lbls","slices")
slices <- data\$slices
lbls <- data\$lbls
colors = c("red","blue","yellow","black","green")
pct <- round(slices/sum(slices)*100, digits = 1)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=colors, main="WT vs. delQ")
par(fig=c(0.5,1,0,1), new=TRUE)
d3 <- read.table("${TMPDIR}/tmp3.txt", header=T, sep="\t")
data3 <- d3[c(2,5,4,1,3),]
colnames(data3) <- c("lbls","slices")
slices <- data3\$slices
lbls <- data3\$lbls
colors = c("red","blue","yellow","black","green")
pct <- round(slices/sum(slices)*100, digits = 1)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=colors, main="delQ vs. WT")
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.piechartWTvsDelQ.R"
R < "${FIGURESDIR}/R/R.piechartWTvsDelQ.R" --no-save

# motif score distribution
annotatePeaks.pl ${DIFFDIR}/WTvsDelQ.peaks.txt hg19 -size 200 -m ${PU1MOTIF} -mscore -nogene
-noann > ${TMPDIR}/tmp.5.WTvsDelQ.motifScore.txt
annotatePeaks.pl ${DIFFDIR}/DelQvsWT.peaks.txt hg19 -size 200 -m ${PU1MOTIF} -mscore -nogene
-noann > ${TMPDIR}/tmp.5.DelQvsWT.motifScore.txt

# beanplot
cd ${WORKDIR}
cat >"${FIGURESDIR}/R/R.motifScores.WTvsDelQ.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.5.WTvsDelQ.motifScore.txt", header=T, sep="\t")
```

```
data3 <- read.table("${TMPDIR}/tmp.5.DelQvsWT.motifScore.txt", header=T, sep="\t")
# combining the data columns
a <- data1[-1,10]
c <- data3[-1,10]
z <- c("a", "c")
x <- lapply(z, get, envir=environment())
names(x) <- z
# determining the length of each column
anum <- nrow(as.matrix(a))
cnum <- nrow(as.matrix(c))
# define labels
laba <- paste("WT (",anum,")",sep="")
labc <- paste("delQ (",cnum,")",sep="")
# statistical testing
wilac <- wilcox.test(a,c)
if (wilac\$p.value < 0.001) {
wac <- "***"
} else if (wilac\$p.value < 0.01) {
wac <- "**"
} else if (wilac\$p.value < 0.05) {
wac <- "*"
} else {
wac <- "ns"
}
# defining colors
beancol <- list("${colPU1}","${colDelQ}")
boxcol <- c("gray65","gray85")
# creating the plot
pdf(file="${FIGURESDIR}/motifScores.WTvsDelQ.pdf", height=3, width=1.6)
par(mar=c(4.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(2,13),lwd=0.5)
# adding box plot on top
par(mar=c(4.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(2,13),lwd=0.6)
# axis and legends
# axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:3, labels=c(laba, labb,
labc),las=2,mgp=c(1,.6,0))
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:2, labels=c(laba,
labc),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext(" Motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
#mtext("Cluster",family="Helvetica",ps=12,side=1,line=2,cex=1.2,padj=2.0)
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
# adding significance levels
text(1.5,12.5,wac)
segments(1,12.3,2,12.3)
EOF
chmod 750 "${FIGURESDIR}/R/R.motifScores.WTvsDelQ.R"
R < "${FIGURESDIR}/R/R.motifScores.WTvsDelQ.R" --no-save

# position of differential peaks in K14CLUSTER
pos2bed.pl ${DIFFDIR}/WTvsDelQ.peaks.txt > ${DIFFDIR}/WTvsDelQ.peaks.bed
$BEDTOOLS intersect -a ${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed -b
${DIFFDIR}/WTvsDelQ.peaks.bed -c > ${TMPDIR}/WTvsDelQ.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}'
${TMPDIR}/WTvsDelQ.peaks.bound.bed
${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.WTvsDelQ.sorted.txt

# "ghistplot" for presence of paired motifs
_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.WTvsDelQ.sorted.txt", header=T)
head(data, n=50)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
mycol <- colorRampPalette(c("white","${colDelQ}"))(3)
png(filename="WTvsDelQinK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
```

```
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# "ghistplot" for ratio of WTvsDelQ
annotatePeaks.pl ${PU1CTV1PEAKS} hg19 -size 200 -d ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
${TAGDIR}/CTV1_PU.1delQ_Flag_R1_CNVnormRefChr ${TAGDIR}/CTV1_PU.1delQ_Flag_R2_CNVnormRefChr
-nogene -noann >${DIFFDIR}/WT_DelQann.txt
awk 'BEGIN{OFS="\t"} {print $1,(($8+$9+$10+$11+2)/2)/($12+$13+1)}' ${DIFFDIR}/WT_DelQann.txt
> ${TMPDIR}/WT_DelQratio.txt
awk 'BEGIN{OFS="\t"} NR==FNR{a[$1]=$2;next}{print $0,a[$1]}' ${TMPDIR}/WT_DelQratio.txt
${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/WT_DelQratio.sorted.txt
_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/WT_DelQratio.sorted.txt", header=T)
head(data, n=50)
d <- log2(data.matrix(data[,7]))
q <- apply(d, 2, function(x) ifelse(x > 2, 2 ,x))
p <- apply(q, 2, function(x) ifelse(x < -2, -2 ,x))
o <- p[order(nrow(q):1),]
mycol <- colorRampPalette(c("dodgerblue3","floralwhite","firebrick2"))(4)
png(filename="WTvsDelQinK14cluster_ratio.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(-2,2)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# delQ vs delA

# histograms across differentially bound regions
# generating necessary ghist files to create the plots
declare -a MERGEDTAGDIR=("${TAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delQ_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delA_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr" "${TAGDIR}/CTV1_PU1_Flag_3ug_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_H3K27ac_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged"
"${ATACDIR}/CTV1_PU1_CNVnormRefChr_merged" "${ATACDIR}/CTV1_PU1mut_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr")

_DATE=$(date +%s)
for SET in "${MERGEDTAGDIR[@]}"
do
NAME=${SET##*/}
cat >"${TMPDIR}/ghist.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
if [ -e ${DIFFDIR}/DelQvsDelA.peaks.txt ] ; then
annotatePeaks.pl ${DIFFDIR}/DelQvsDelA.peaks.txt hg19 -size 2000 -hist 25 -ghist -d ${SET} >
${GHISTFILEDIR}/DelQvsDelA.${NAME}.ann.ghist.txt ; fi
EOF
chmod 750 "${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
echo "generating ghist for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
done
```

```
# loop to check when screen sessions are done
#---------------------------------------------
for SET in "${MERGEDTAGDIR[@]}" ; do
NAME=${SET##*/}
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------

# generate histogram plots
# DelQvsDelA
cd ${GHISTFILEDIR}
# PU.1
plotHIST.sh \
-f "${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1_Flag_CNVnormRefChr_merged.ann.ghist.txt" \
-s "mutPU1 PU1delQ PU1delA PU1" -c "${colMUT} ${colDelQ} ${colDelA} ${colPU1}" -x 1000 -y "0
35" -d ${FIGURESDIR}/hist -n DelQvsDelA.PU1
# ATAC & H3K27ac
plotHIST.sh \
-f "${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1mut_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1_CNVnormRefChr_merged.ann.ghist.txt" \
-s "H3K27ac H3K27ac-PU1 ATAC ATAC-PU1" -c "cyan3 darkcyan ${colMUT} ${colPU1}" -x 1000 -y "0
20" -d ${FIGURESDIR}/hist -n DelQvsDelA.H3K27ac-ATAC
# BRGi
plotHIST.sh \
-f "${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delQ PU1delA PU1-DMSO PU1-1uM PU1-5uM PU1-10uM" -c "${colMUT} ${colDelQ}
${colDelA} ${colPU1} orchid3 palevioletred2 plum1" -x 1000 -y "0 30" -d ${FIGURESDIR}/hist
-n DelQvsDelA.BRGi
# titration
plotHIST.sh \
-f "${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1_Flag_6ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1_Flag_3ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/DelQvsDelA.CTV1_PU1_Flag_0.9ug_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1delQ PU1delA PU1-100% PU1-50% PU1-15%" -c "${colMUT} ${colDelQ} ${colDelA}
${colPU16u} ${colPU13u} ${colPU1p9u}" -x 1000 -y "0 30" -d ${FIGURESDIR}/hist -n
DelQvsDelA.Titration

# motif searches in differentially bound regions
cd ${WORKDIR}
findMotifsGenome.pl ${DIFFDIR}/DelQvsDelA.peaks.txt hg19r ${MOTIFDIR}/DelQvsDelA.peaks -size
200 -len 7,8,9,10,11,12,13,14 -p 12 -h

# reformating output with own parameters
compareMotifs.pl ${MOTIFDIR}/DelQvsDelA.peaks/homerMotifs.all.motifs
${MOTIFDIR}/DelQvsDelA.peaks/final -reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info 1.5
-cpu 12

# genome ontology
cd ${WORKDIR}
annotatePeaks.pl ${DIFFDIR}/DelQvsDelA.peaks.txt hg19 -size given -gtf ${HG19GTF} -annStats
${DIFFDIR}/DelQvsDelA.peaks.genome.stats.txt > ${DIFFDIR}/tmp.4.DelQvsDelA.peaks.txt
head -n 6 ${DIFFDIR}/DelQvsDelA.peaks.genome.stats.txt | cut -f 1,2 > ${TMPDIR}/tmp1.txt
cd ${WORKDIR}
cat >"${FIGURESDIR}/R/R.piechartDelQvsDelA.R" <<EOF
library(plotrix)
pdf(file="${FIGURESDIR}/genomeOntologyPieChartDelQvsDelA.pdf", height=4, width=4.5)
```

```
d <- read.table("${TMPDIR}/tmp1.txt", header=T, sep="\t")
data <- d[c(2,5,4,1,3),]
colnames(data) <- c("lbls","slices")
slices <- data\$slices
lbls <- data\$lbls
colors = c("red","blue","yellow","black","green")
pct <- round(slices/sum(slices)*100, digits = 1)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=colors, main="DelQ vs. DelA")
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.piechartDelQvsDelA.R"
R < "${FIGURESDIR}/R/R.piechartDelQvsDelA.R" --no-save

# motif score distribution
annotatePeaks.pl ${DIFFDIR}/WTvsDelA.peaks.txt hg19 -size 200 -m ${PU1MOTIF} -mscore -nogene
-noann > ${TMPDIR}/tmp.5.WTvsDelA.motifScore.txt
annotatePeaks.pl ${DIFFDIR}/DelQvsDelA.peaks.txt hg19 -size 200 -m ${PU1MOTIF} -mscore
-nogene -noann > ${TMPDIR}/tmp.5.DelQvsDelA.motifScore.txt

# beanplot
cd ${WORKDIR}
cat >"${FIGURESDIR}/R/R.motifScores.DelQvsDelA.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.5.WTvsDelQ.motifScore.txt", header=T, sep="\t")
data3 <- read.table("${TMPDIR}/tmp.5.DelQvsDelA.motifScore.txt", header=T, sep="\t")
# combining the data columns
a <- data1[-1,10]
c <- data3[-1,10]
z <- c("a", "c")
x <- lapply(z, get, envir=environment())
names(x) <- z
# determining the length of each column
anum <- nrow(as.matrix(a))
cnum <- nrow(as.matrix(c))
# define labels
laba <- paste("WT (",anum,")",sep="")
labc <- paste("delQ (",cnum,")",sep="")
# statistical testing
wilac <- wilcox.test(a,c)
if (wilac\$p.value < 0.001) {
wac <- "***"
} else if (wilac\$p.value < 0.01) {
wac <- "**"
} else if (wilac\$p.value < 0.05) {
wac <- "*"
} else {
wac <- "ns"
}
# defining colors
beancol <- list("${colPU1}","${colDelQ}")
boxcol <- c("gray65","gray85")
# creating the plot
pdf(file="${FIGURESDIR}/motifScores.DelQvsDelA.pdf", height=3, width=1.6)
par(mar=c(4.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(2,13),lwd=0.5)
# adding box plot on top
par(mar=c(4.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(2,13),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:2, labels=c(laba,
labc),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext(" Motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
#mtext("Cluster",family="Helvetica",ps=12,side=1,line=2,cex=1.2,padj=2.0)
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
# adding significance levels
text(1.5,12.5,wac)
segments(1,12.3,2,12.3)
EOF
chmod 750 "${FIGURESDIR}/R/R.motifScores.DelQvsDelA.R"
R < "${FIGURESDIR}/R/R.motifScores.DelQvsDelA.R" --no-save
```

```
# position of differential peaks in K14CLUSTER
pos2bed.pl ${DIFFDIR}/DelQvsDelA.peaks.txt > ${DIFFDIR}/DelQvsDelA.peaks.bed
$BEDTOOLS intersect -a ${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed -b
${DIFFDIR}/DelQvsDelA.peaks.bed -c > ${TMPDIR}/DelQvsDelA.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}'
${TMPDIR}/DelQvsDelA.peaks.bound.bed
${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.DelQvsDelA.sorted.txt

# "ghistplot" for presence of paired motifs
_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.DelQvsDelA.sorted.txt", header=T)
head(data, n=50)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
mycol <- colorRampPalette(c("white","${colDelA}"))(3)
png(filename="DelQvsDelAinK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# "ghistplot" for ratio of DelQ vs DelA
annotatePeaks.pl ${PU1CTV1PEAKS} hg19 -size 200 -d
${TAGDIR}/CTV1_PU.1delQ_Flag_R1_CNVnormRefChr ${TAGDIR}/CTV1_PU.1delQ_Flag_R2_CNVnormRefChr \
${TAGDIR}/CTV1_PU.1delA_Flag_R1_CNVnormRefChr ${TAGDIR}/CTV1_PU.1delA_Flag_R2_CNVnormRefChr
-nogene -noann >${DIFFDIR}/DelQ_DelAann.txt
awk 'BEGIN{OFS="\t"} {print $1,(($8+$9+1))/($10+$11+1)}' ${DIFFDIR}/DelQ_DelAann.txt >
${TMPDIR}/DelQ_DelAratio.txt
awk 'BEGIN{OFS="\t"} NR==FNR{a[$1]=$2;next}{print $0,a[$1]}' ${TMPDIR}/DelQ_DelAratio.txt
${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/DelQ_DelAratio.sorted.txt
_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/DelQ_DelAratio.sorted.txt", header=T)
head(data, n=50)
d <- log2(data.matrix(data[,7]))
q <- apply(d, 2, function(x) ifelse(x > 2, 2 ,x))
p <- apply(q, 2, function(x) ifelse(x < -2, -2 ,x))
o <- p[order(nrow(q):1),]
mycol <- colorRampPalette(c("dodgerblue3","floralwhite","firebrick2"))(4)
png(filename="DelQvsDelAinK14cluster_ratio.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(-2,2)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R
# less PU.1 vs WT

# histograms across differentially bound regions
# generating necessary ghist files to create the plots
declare -a MERGEDTAGDIR=("${TAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr" "${TAGDIR}/CTV1_PU1_Flag_3ug_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_H3K27ac_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged"
"${ATACDIR}/CTV1_PU1_CNVnormRefChr_merged" "${ATACDIR}/CTV1_PU1mut_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr"
"${TAGDIR}/CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr")
```

```
for SET in "${MERGEDTAGDIR[@]}"
do
NAME=${SET##*/}
_DATE=$(date +%s)
cat >"${TMPDIR}/ghist.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
if [ -e ${DIFFDIR}/WTvsLessPU1.peaks.txt ] ; then
annotatePeaks.pl ${DIFFDIR}/WTvsLessPU1.peaks.txt hg19 -size 2000 -hist 25 -ghist -d ${SET}
> ${GHISTFILEDIR}/WTvsLessPU1.${NAME}.ann.ghist.txt ; fi
EOF
chmod 750 "${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
echo "generating ghist for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#---------------------------------------------
for SET in "${MERGEDTAGDIR[@]}" ; do
NAME=${SET##*/}
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------

# generate histogram plots
# WTvsLessPU1
cd ${GHISTFILEDIR}
# PU.1
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_Flag_0.9ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_Flag_CNVnormRefChr_merged.ann.ghist.txt" \
-s "mutPU1 lessPU1 PU1" -c "${colMUT} ${ colPU1p9u} ${colPU1}" -x 1000 -y "0 20" -d
${FIGURESDIR}/hist -n WTvsLessPU1.PU1
# ATAC & H3K27ac
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_H3K27ac_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1mut_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_CNVnormRefChr_merged.ann.ghist.txt" \
-s "H3K27ac H3K27ac-PU1 ATAC ATAC-PU1" -c "cyan3 darkcyan ${colMUT} ${colPU1}" -x 1000 -y "0
6" -d ${FIGURESDIR}/hist -n WTvsLessPU1.H3K27ac-ATAC
# BRGi
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_Flag_0.9ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_FLAG_8h_DMSO_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_FLAG_8h_1uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_FLAG_8h_5uM_PFI3_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_FLAG_8h_10uM_PFI3_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 lessPU1 PU1-DMSO PU1-1uM PU1-5uM PU1-10uM" -c "${colMUT} ${colDelA} ${colPU1}
orchid3 palevioletred2 plum1" -x 1000 -y "0 20" -d ${FIGURESDIR}/hist -n WTvsLessPU1.BRGi
# titration
plotHIST.sh \
-f "${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_Flag_6ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_Flag_3ug_CNVnormRefChr.ann.ghist.txt
${GHISTFILEDIR}/WTvsLessPU1.CTV1_PU1_Flag_0.9ug_CNVnormRefChr.ann.ghist.txt" \
-s "mutPU1 PU1-100% PU1-50% PU1-15%" -c "${colMUT} ${colPU16u} ${colPU13u} ${colPU1p9u}" -x
1000 -y "0 20" -d ${FIGURESDIR}/hist -n WTvsLessPU1.Titration

# motif searches in differentially bound regions
cd ${WORKDIR}
findMotifsGenome.pl ${DIFFDIR}/WTvsLessPU1.peaks.txt hg19r ${MOTIFDIR}/WTvsLessPU1.peaks
-size 200 -len 7,8,9,10,11,12,13,14 -p 12 -h
```

```
# reformating output with own parameters
compareMotifs.pl ${MOTIFDIR}/WTvsLessPU1.peaks/homerMotifs.all.motifs
${MOTIFDIR}/WTvsLessPU1.peaks/final -reduceThresh .75 -matchThresh .6 -pvalue 1e-12 -info
1.5 -cpu 12

# genome ontology
cd ${WORKDIR}
annotatePeaks.pl ${DIFFDIR}/WTvsLessPU1.peaks.txt hg19 -size given -gtf ${HG19GTF} -annStats
${DIFFDIR}/WTvsLessPU1.peaks.genome.stats.txt > ${DIFFDIR}/tmp.4.WTvsLessPU1.peaks.txt
head -n 6 ${DIFFDIR}/WTvsLessPU1.peaks.genome.stats.txt | cut -f 1,2 > ${TMPDIR}/tmp1.txt
cd ${WORKDIR}
cat >"${FIGURESDIR}/R/R.piechartWTvsLessPU1.R" <<EOF
library(plotrix)
pdf(file="${FIGURESDIR}/genomeOntologyPieChartWTvsLessPU1.pdf", height=4, width=4.5)
d <- read.table("${TMPDIR}/tmp1.txt", header=T, sep="\t")
data <- d[c(2,5,4,1,3),]
colnames(data) <- c("lbls","slices")
slices <- data\$slices
lbls <- data\$lbls
colors = c("red","blue","yellow","black","green")
pct <- round(slices/sum(slices)*100, digits = 1)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=colors, main="WT vs. lessPU.1")
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.piechartWTvsLessPU1.R"
R < "${FIGURESDIR}/R/R.piechartWTvsLessPU1.R" --no-save

# motif score distribution
annotatePeaks.pl ${DIFFDIR}/WTvsLessPU1.peaks.txt hg19 -size 200 -m ${PU1MOTIF} -mscore
-nogene -noann > ${TMPDIR}/tmp.5.WTvsLessPU1.motifScore.txt

# beanplot
cd ${WORKDIR}
cat >"${FIGURESDIR}/R/R.motifScores.WTvsLessPU1.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.5.WTvsLessPU1.motifScore.txt", header=T, sep="\t")
x <- data1[-1,10]
# determining the length of each column
anum <- nrow(as.matrix(x))
# define labels
laba <- paste("WT (",anum,")",sep="")
beancol <- "${colPU1}"
boxcol <- "gray65"
# creating the plot
pdf(file="${FIGURESDIR}/motifScores.WTvsLessPU1.pdf", height=3, width=1.2)
par(mar=c(4.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(2,13),lwd=0.5)
# adding box plot on top
par(mar=c(4.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.5, staplewex = 0.5, ylim=c(2,13),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:1, labels=laba,las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext(" Motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.motifScores.WTvsLessPU1.R"
R < "${FIGURESDIR}/R/R.motifScores.WTvsLessPU1.R" --no-save

# position of differential peaks in K14CLUSTER
pos2bed.pl ${DIFFDIR}/WTvsLessPU1.peaks.txt > ${DIFFDIR}/WTvsLessPU1.peaks.bed
$BEDTOOLS intersect -a ${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed -b
${DIFFDIR}/WTvsLessPU1.peaks.bed -c > ${TMPDIR}/WTvsLessPU1.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}'
${TMPDIR}/WTvsLessPU1.peaks.bound.bed
${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.WTvsLessPU1.sorted.txt
```

```
# "ghistplot" for presence of paired motifs
_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.WTvsLessPU1.sorted.txt", header=T)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
mycol <- colorRampPalette(c("white","dodgerblue1"))(3)
png(filename="WTvsLessPU1inK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# "ghistplot" for ratio of WTvslessPU1
annotatePeaks.pl ${PU1CTV1PEAKS} hg19 -size 200 -d ${TAGDIR}/CTV1_PU1_Flag_R1_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_R2_CNVnormRefChr ${TAGDIR}/CTV1_PU1_Flag_R3_CNVnormRefChr
${TAGDIR}/CTV1_PU1_Flag_6ug_CNVnormRefChr \
${TAGDIR}/CTV1_PU1_Flag_0.9ug_CNVnormRefChr -nogene -noann >${DIFFDIR}/WT_lessPU1ann.txt
awk 'BEGIN{OFS="\t"} {print $1,(($8+$9+$10+$11+2)/4)/($12+0.5)}'
${DIFFDIR}/WT_lessPU1ann.txt > ${TMPDIR}/WT_lessPU1ratio.txt
awk 'BEGIN{OFS="\t"} NR==FNR{a[$1]=$2;next}{print $0,a[$1]}' ${TMPDIR}/WT_lessPU1ratio.txt
${BASICDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/WT_lessPU1ratio.sorted.txt
_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/WT_lessPU1ratio.sorted.txt", header=T)
head(data, n=50)
d <- log2(data.matrix(data[,7]))
q <- apply(d, 2, function(x) ifelse(x > 2, 2 ,x))
p <- apply(q, 2, function(x) ifelse(x < -2, -2 ,x))
o <- p[order(nrow(q):1),]
mycol <- colorRampPalette(c("dodgerblue3","floralwhite","firebrick2"))(4)
png(filename="WTvslessPU1inK14cluster_ratio.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(-2,2)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# cluster vs single or no motif
# histograms across differentially bound regions
# ==============================================

# generating necessary ghist files to create the plots
declare -a MERGEDTAGDIR=("${TAGDIR}/CTV1_PU1_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delP_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delQ_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delA_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1delAQP_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_Flag_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1_H3K27ac_CNVnormRefChr_merged"
"${TAGDIR}/CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged"
"${ATACDIR}/CTV1_PU1_CNVnormRefChr_merged" "${ATACDIR}/CTV1_PU1mut_CNVnormRefChr_merged")
_DATE=$(date +%s)
for SET in "${MERGEDTAGDIR[@]}"
do
NAME=${SET##*/}
cat >"${TMPDIR}/ghist.${NAME}.${_DATE}.sh" <<EOF
#!/bin/bash
#setting homer environment
export
PATH=/misc/software/package/RBioC/3.4.3/bin:/misc/software/package/perl/perl-5.26.1/bin:/misc/
software/ngs/samtools/samtools-1.6/bin:/misc/software/ngs/homer/v4.9/bin:${PATH}
export PATH
cd ${TMPDIR}
annotatePeaks.pl ${PEAKDIR}/PU1.ntag15.filtered.cluster.pos.txt hg19 -size 2000 -hist 25
-ghist -d ${SET} > ${GHISTFILEDIR}/PU1.ntag15.filtered.cluster.${NAME}.ann.ghist.txt
annotatePeaks.pl ${PEAKDIR}/PU1.ntag15.filtered.singleMotif.pos.txt hg19 -size 2000 -hist 25
```

```
-ghist -d ${SET} > ${GHISTFILEDIR}/PU1.ntag15.filtered.singleMotif.${NAME}.ann.ghist.txt
annotatePeaks.pl ${PEAKDIR}/PU1.ntag15.filtered.noMotif.pos.txt hg19 -size 2000 -hist 25
-ghist -d ${SET} > ${GHISTFILEDIR}/PU1.ntag15.filtered.noMotif.${NAME}.ann.ghist.txt
EOF
chmod 750 "${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
echo "generating ghist for ${NAME}"
screen -dm -S ${NAME} bash -c "bash ${TMPDIR}/ghist.${NAME}.${_DATE}.sh"
done

# loop to check when screen sessions are done
#---------------------------------------------
for SET in "${MERGEDTAGDIR[@]}" ; do
NAME=${SET##*/}
while [ true ]; do # Endless loop.
pid=`screen -S ${NAME} -Q echo '$PID'` # Get a pid.
if [[ $pid = *"session"* ]] ; then # If there is none,
echo -e "\tFinished sample ${NAME}"
break # Test next one.
else
sleep 10 # Else wait.
fi
done
done
#---------------------------------------------

# generate histogram plots

# single motifs
cd ${GHISTFILEDIR}
# PU.1
plotHIST.sh \
-f
"${GHISTFILEDIR}/PU1.ntag15.filtered.singleMotif.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghi
st.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.singleMotif.CTV1_PU1delP_Flag_CNVnormRefChr_merged.ann.ghi
st.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.singleMotif.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghi
st.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.singleMotif.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghi
st.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.singleMotif.CTV1_PU1delAQP_Flag_CNVnormRefChr_merged.ann.g
hist.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.singleMotif.CTV1_PU1_Flag_CNVnormRefChr_merged.ann.ghist.t
xt" \
-s "mutPU1 PU1delP PU1delQ PU1delA PU1delAQP PU1" -c "${colMUT} ${colDelP} ${colDelQ}
${colDelA} ${colDelAQP} ${colPU1}" -x 1000 -y "0 20" -d ${FIGURESDIR}/hist -n singleMotif.PU1
# ATAC & H3K27ac
plotHIST.sh \
-f
"${GHISTFILEDIR}/PU1.ntag15.filtered.singleMotif.CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged.ann.
ghist.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.singleMotif.CTV1_PU1_H3K27ac_CNVnormRefChr_merged.ann.ghis
t.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.singleMotif.CTV1_PU1mut_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.singleMotif.CTV1_PU1_CNVnormRefChr_merged.ann.ghist.txt" \
-s "H3K27ac H3K27ac-PU1 ATAC ATAC-PU1" -c "cyan3 darkcyan ${colMUT} ${colPU1}" -x 1000 -y "0
10" -d ${FIGURESDIR}/hist -n PU1.ntag15.filtered.singleMotif.H3K27ac-ATAC

# cluster
cd ${GHISTFILEDIR}
# PU.1
plotHIST.sh \
-f
"${GHISTFILEDIR}/PU1.ntag15.filtered.cluster.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.t
xt
${GHISTFILEDIR}/PU1.ntag15.filtered.cluster.CTV1_PU1delP_Flag_CNVnormRefChr_merged.ann.ghist.t
xt
${GHISTFILEDIR}/PU1.ntag15.filtered.cluster.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghist.t
xt
${GHISTFILEDIR}/PU1.ntag15.filtered.cluster.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghist.t
xt
${GHISTFILEDIR}/PU1.ntag15.filtered.cluster.CTV1_PU1delAQP_Flag_CNVnormRefChr_merged.ann.ghist
.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.cluster.CTV1_PU1_Flag_CNVnormRefChr_merged.ann.ghist.txt"
\
-s "mutPU1 PU1delP PU1delQ PU1delA PU1delAQP PU1" -c "${colMUT} ${colDelP} ${colDelQ}
${colDelA} ${colDelAQP} ${colPU1}" -x 1000 -y "0 22" -d ${FIGURESDIR}/hist -n cluster.PU1
# ATAC & H3K27ac
```

```
plotHIST.sh \
-f
"${GHISTFILEDIR}/PU1.ntag15.filtered.cluster.CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged.ann.ghis
t.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.cluster.CTV1_PU1_H3K27ac_CNVnormRefChr_merged.ann.ghist.tx
t ${GHISTFILEDIR}/PU1.ntag15.filtered.cluster.CTV1_PU1mut_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.cluster.CTV1_PU1_CNVnormRefChr_merged.ann.ghist.txt" \
-s "H3K27ac H3K27ac-PU1 ATAC ATAC-PU1" -c "cyan3 darkcyan ${colMUT} ${colPU1}" -x 1000 -y "0
10" -d ${FIGURESDIR}/hist -n PU1.ntag15.filtered.cluster.H3K27ac-ATAC

# no motifs
cd ${GHISTFILEDIR}
# PU.1
plotHIST.sh \
-f
"${GHISTFILEDIR}/PU1.ntag15.filtered.noMotif.CTV1_PU1mut_Flag_CNVnormRefChr_merged.ann.ghist.t
xt
${GHISTFILEDIR}/PU1.ntag15.filtered.noMotif.CTV1_PU1delP_Flag_CNVnormRefChr_merged.ann.ghist.t
xt
${GHISTFILEDIR}/PU1.ntag15.filtered.noMotif.CTV1_PU1delQ_Flag_CNVnormRefChr_merged.ann.ghist.t
xt
${GHISTFILEDIR}/PU1.ntag15.filtered.noMotif.CTV1_PU1delA_Flag_CNVnormRefChr_merged.ann.ghist.t
xt
${GHISTFILEDIR}/PU1.ntag15.filtered.noMotif.CTV1_PU1delAQP_Flag_CNVnormRefChr_merged.ann.ghist
.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.noMotif.CTV1_PU1_Flag_CNVnormRefChr_merged.ann.ghist.txt"
\
-s "mutPU1 PU1delP PU1delQ PU1delA PU1delAQP PU1" -c "${colMUT} ${colDelP} ${colDelQ}
${colDelA} ${colDelAQP} ${colPU1}" -x 1000 -y "0 20" -d ${FIGURESDIR}/hist -n noMotif.PU1
# ATAC & H3K27ac
plotHIST.sh \
-f
"${GHISTFILEDIR}/PU1.ntag15.filtered.noMotif.CTV1_PU1mut_H3K27ac_CNVnormRefChr_merged.ann.
ghist.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.noMotif.CTV1_PU1_H3K27ac_CNVnormRefChr_merged.ann.ghist.tx
t ${GHISTFILEDIR}/PU1.ntag15.filtered.noMotif.CTV1_PU1mut_CNVnormRefChr_merged.ann.ghist.txt
${GHISTFILEDIR}/PU1.ntag15.filtered.noMotif.CTV1_PU1_CNVnormRefChr_merged.ann.ghist.txt" \
-s "H3K27ac H3K27ac-PU1 ATAC ATAC-PU1" -c "cyan3 darkcyan ${colMUT} ${colPU1}" -x 1000 -y "0
20" -d ${FIGURESDIR}/hist -n PU1.ntag15.filtered.noMotif.H3K27ac-ATAC

# motif searches in homotypic clusters versus single motifs
# ========================================================
cd ${WORKDIR}
findMotifsGenome.pl ${PEAKDIR}/PU1.ntag15.filtered.cluster.pos.txt hg19r
${MOTIFDIR}/PU1.ntag15.filtered.clusters -size 200 -len 7,8,9,10,11,12,13,14 -p 12 -h
findMotifsGenome.pl ${PEAKDIR}/PU1.ntag15.filtered.singleMotif.pos.txt hg19r
${MOTIFDIR}/PU1.ntag15.filtered.singleMotifs -size 200 -len 7,8,9,10,11,12,13,14 -p 12 -h
findMotifsGenome.pl ${PEAKDIR}/PU1.ntag15.filtered.noMotif.pos.txt hg19r
${MOTIFDIR}/PU1.ntag15.filtered.noMotifs -size 200 -len 7,8,9,10,11,12,13,14 -p 12 -h

# reformating output with own parameters
compareMotifs.pl ${MOTIFDIR}/PU1.ntag15.filtered.clusters/homerMotifs.all.motifs
${MOTIFDIR}/PU1.ntag15.filtered.clusters/final -reduceThresh .75 -matchThresh .6 -pvalue
1e-12 -info 1.5 -cpu 12
compareMotifs.pl ${MOTIFDIR}/PU1.ntag15.filtered.singleMotifs/homerMotifs.all.motifs
${MOTIFDIR}/PU1.ntag15.filtered.singleMotifs/final -reduceThresh .75 -matchThresh .6 -pvalue
1e-12 -info 1.5 -cpu 12
compareMotifs.pl ${MOTIFDIR}/PU1.ntag15.filtered.noMotifs/homerMotifs.all.motifs
${MOTIFDIR}/PU1.ntag15.filtered.noMotifs/final -reduceThresh .75 -matchThresh .6 -pvalue
1e-12 -info 1.5 -cpu 12

# genome ontology
# ===============
cd ${WORKDIR}
annotatePeaks.pl ${PEAKDIR}/PU1.ntag15.filtered.cluster.pos.txt hg19 -size given -gtf
${HG19GTF} -annStats ${DIFFDIR}/PU1.ntag15.filtered.clusters.genome.stats.txt >
${DIFFDIR}/tmp.PU1.ntag15.filtered.clusters.txt
annotatePeaks.pl ${PEAKDIR}/PU1.ntag15.filtered.singleMotif.pos.txt hg19 -gtf ${HG19GTF}
-size given -annStats ${DIFFDIR}/PU1.ntag15.filtered.singleMotifs.genome.stats.txt >
${DIFFDIR}/tmp.4.PU1.ntag15.filtered.singleMotifs.peaks.txt
annotatePeaks.pl ${PEAKDIR}/PU1.ntag15.filtered.noMotif.pos.txt hg19 -gtf ${HG19GTF} -size
given -annStats ${DIFFDIR}/PU1.ntag15.filtered.noMotifs.genome.stats.txt >
${DIFFDIR}/tmp.4.PU1.ntag15.filtered.noMotifs.peaks.txt
head -n 6 ${DIFFDIR}/PU1.ntag15.filtered.clusters.genome.stats.txt | cut -f 1,2 >
${TMPDIR}/tmp1.txt
head -n 6 ${DIFFDIR}/PU1.ntag15.filtered.singleMotifs.genome.stats.txt | cut -f 1,2 >
${TMPDIR}/tmp2.txt
head -n 6 ${DIFFDIR}/PU1.ntag15.filtered.noMotifs.genome.stats.txt | cut -f 1,2 >
```

```
${TMPDIR}/tmp3.txt

cat >"${FIGURESDIR}/R/R.piechartWTvsDelA.R" <<EOF
library(plotrix)
pdf(file="${FIGURESDIR}/genomeOntologyPieChartClustervsSingleorNoMotifs.pdf", height=4,
width=12.5)
par(fig=c(0,0.33,0,1))
d <- read.table("${TMPDIR}/tmp1.txt", header=T, sep="\t")
data <- d[c(2,5,4,1,3),]
colnames(data) <- c("lbls","slices")
slices <- data\$slices
lbls <- data\$lbls
colors = c("red","blue","yellow","black","green")
pct <- round(slices/sum(slices)*100, digits = 1)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=colors, main="Homotypic clusters")
par(fig=c(0.34,0.66,0,1), new=TRUE)
d2 <- read.table("${TMPDIR}/tmp2.txt", header=T, sep="\t")
data2 <- d2[c(2,5,4,1,3),]
colnames(data2) <- c("lbls","slices")
slices <- data2\$slices
lbls <- data2\$lbls
colors = c("red","blue","yellow","black","green")
pct <- round(slices/sum(slices)*100, digits = 1)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=colors, main="Single motifs")
par(fig=c(0.67,1,0,1), new=TRUE)
d3 <- read.table("${TMPDIR}/tmp3.txt", header=T, sep="\t")
data3 <- d3[c(2,5,4,1,3),]
colnames(data3) <- c("lbls","slices")
slices <- data3\$slices
lbls <- data3\$lbls
colors = c("red","blue","yellow","black","green")
pct <- round(slices/sum(slices)*100, digits = 1)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=colors, main="No motif")
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.piechartWTvsDelA.R"
R < "${FIGURESDIR}/R/R.piechartWTvsDelA.R" --no-save

# motif score distribution
# ========================
annotatePeaks.pl ${PEAKDIR}/PU1.ntag15.filtered.cluster.pos.txt hg19 -size 200 -m
${PU1MOTIF} -mscore -nogene -noann > ${TMPDIR}/tmp.5.cluster.motifScore.txt
annotatePeaks.pl ${PEAKDIR}/PU1.ntag15.filtered.singleMotif.pos.txt hg19 -size 200 -m
${PU1MOTIF} -mscore -nogene -noann > ${TMPDIR}/tmp.5.singleMotifs.motifScore.txt
# not needed since no motif there:
# annotatePeaks.pl ${PEAKDIR}/PU1.ntag15.filtered.noMotif.pos.txt hg19 -size 200 -m
${PU1MOTIF} -mscore -nogene -noann > ${TMPDIR}/tmp.5.noMotifs.motifScore.txt

# combined bean- and box-plots
cd ${WORKDIR}
cat >"${FIGURESDIR}/R/R.motifScores.cluster.R" <<EOF
library(beanplot)
# collecting the data columns from individual annotation files
data1 <- read.table("${TMPDIR}/tmp.5.cluster.motifScore.txt", header=T, sep="\t")
data3 <- read.table("${TMPDIR}/tmp.5.singleMotifs.motifScore.txt", header=T, sep="\t")
# combining the data columns
a <- data1[-1,10]
c <- data3[-1,10]
z <- c("a", "c")
x <- lapply(z, get, envir=environment())
names(x) <- z
# determining the length of each column
anum <- nrow(as.matrix(a))
cnum <- nrow(as.matrix(c))
# define labels
laba <- paste("cluster (",anum,")",sep="")
labc <- paste("single (",cnum,")",sep="")
# statistical testing
wilac <- wilcox.test(a,c)
if (wilac\$p.value < 0.001) {
wac <- "***"
} else if (wilac\$p.value < 0.01) {
```

```
wac <- "**"
} else if (wilac\$p.value < 0.05) {
wac <- "*"
} else {
wac <- "ns"
}
# defining colors
beancol <- list("${colClus}","${colSmot}")
boxcol <- c("gray65","gray85")
# creating the plot
pdf(file="${FIGURESDIR}/motifScores.ClustervsSingleMotifs.pdf", height=3, width=1.6)
par(mar=c(4.5,2.5,1,1))
# plotting the beans
beanplot(x,log="",bw="nrd", what=c(0,1,0,0),axes = FALSE, col = beancol , border = boxcol
,overallline = "median", method="jitter", boxwex = 1, beanlinewd = .5, maxstripline =
0.8,ylim=c(2,13),lwd=0.5)
# adding box plot on top
par(mar=c(4.5,2.5,1,1),new=TRUE)
boxplot(x,range=0,log="", style="quantile",axes = FALSE, col = boxcol, border = "black",
overallline = "median", notch=TRUE,boxwex = 0.3, staplewex = 0.5, ylim=c(2,13),lwd=0.6)
# axis and legends
axis(1,padj=0.4,family="Helvetica",cex.axis=.8,at=1:2, labels=c(laba,
labc),las=2,mgp=c(1,.6,0))
axis(2,padj=0.4,family="Helvetica",cex.axis=.8,las=1,mgp=c(2,.6,0))
mtext(" Motif log odds score",family="Helvetica",side=2,line=2,cex=1.2,padj=0.8)
abline(h=6.751641,col="darkblue",lty=3,lwd=1)
# adding significance levels
text(1.5,12.5,wac)
segments(1,12.0,2,12.0)
dev.off()
EOF
chmod 750 "${FIGURESDIR}/R/R.motifScores.cluster.R"
R < "${FIGURESDIR}/R/R.motifScores.cluster.R" --no-save

# position of homotypic clusters versus single motifs in K14CLUSTER

# "ghistplot" for presence of paired motifs
$BEDTOOLS intersect -a ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed
-b ${PEAKDIR}/PU1.ntag15.filtered.cluster.bed -c > ${TMPDIR}/cluster.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}'
${TMPDIR}/cluster.peaks.bound.bed
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.cluster.sorted.txt

_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.cluster.sorted.txt", header=T)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
mycol <- colorRampPalette(c("white","${colClus}"))(3)
png(filename="ClustersInK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# "ghistplot" for presence of single motifs
$BEDTOOLS intersect -a ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed
-b ${PEAKDIR}/PU1.ntag15.filtered.singleMotif.bed -c > ${TMPDIR}/singleMotif.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}'
${TMPDIR}/singleMotif.peaks.bound.bed
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.singleMotif.sorted.txt

_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.singleMotif.sorted.txt", header=T)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
```

```
mycol <- colorRampPalette(c("white","${colSmot}"))(3)
png(filename="SingleMotifsInK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# "ghistplot" for absence of motifs
$BEDTOOLS intersect -a ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed
-b ${PEAKDIR}/PU1.ntag15.filtered.noMotif.bed -c > ${TMPDIR}/noMotif.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}'
${TMPDIR}/noMotif.peaks.bound.bed
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.noMotif.sorted.txt

_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.noMotif.sorted.txt", header=T)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
mycol <- colorRampPalette(c("white","${colNmot}"))(3)
png(filename="NoMotifsInK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R

# are homotypic clusters enriched compared to entire genome?
# perform fischer exact, need
# a number of motifs in total
# b number of motif clusters in total
# c number of motifs in peaks
# d number of clusters overlapping motifs in peaks

rm ${ANALYSISDIR}/Enrichment.homotypicClustersKmeans.14.txt
declare -a PEAKMOTIFS=()
declare -a CLUSTERMOTIFS=()
TOTALMOTIFS=$(wc -l <"${HOMOCLUSTERDIR}/PU.1_long_hg19_all.bed")
TOTALCLUSTERS=$(wc -l <"${HOMOCLUSTERDIR}/PU.1_long_hg19_all.cluster.txt")
TAB=$(echo -e "\t")

for ((i=1;i<=14;i++));do
pos2bed.pl "${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt" >
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed"
$BEDTOOLS intersect -a
/misc/data/analysis/project_PU1/homotypicClusters/PU.1_long_hg19_all.bed -b
/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed -u >
/misc/data/tmp/motifs.kmeans.14.sorted.${i}.bed
PEAKMOTIFS[${i}]=$(wc -l <"/misc/data/tmp/motifs.kmeans.14.sorted.${i}.bed")
$BEDTOOLS intersect -a <(cut -f1-3 "${HOMOCLUSTERDIR}/PU.1_long_hg19_all.cluster.txt")
-b /misc/data/tmp/motifs.kmeans.14.sorted.${i}.bed -u >
/misc/data/tmp/clusters.kmeans.14.sorted.${i}.bed
CLUSTERMOTIFS[${i}]=$(wc -l <"/misc/data/tmp/clusters.kmeans.14.sorted.${i}.bed")
cat >>"${ANALYSISDIR}/Enrichment.homotypicClustersKmeans.14.txt" <<EOF
Cluster${i}$TAB$TOTALMOTIFS$TAB$TOTALCLUSTERS$TAB${PEAKMOTIFS[${i}]}$TAB${CLUSTERMOTIFS[${i}]}
EOF
done

hyperTab.pl ${ANALYSISDIR}/Enrichment.homotypicClustersKmeans.14.txt >
${ANALYSISDIR}/Enrichment.homotypicClustersKmeans.14.fisher.txt

# are homotypic clusters, single Motifs or the absence of motifs enriched in the 14 Kmeans
clusters compared to all peaks?

# homotypic clusters
rm ${ANALYSISDIR}/ClusterEnrichment.homotypicClustersKmeans.14.txt
declare -a PEAKS=()
declare -a CLUSTERPEAKS=()
TOTALPEAKS=$(wc -l <"${PU1CTV1PEAKS}")
```

```
TOTALCLUSTERS=$(wc -l <"${PEAKDIR}/PU1.ntag15.filtered.cluster.bed")
TAB=$(echo -e "\t")

for ((i=1;i<=14;i++));do
pos2bed.pl "${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.txt" >
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed"
PEAKS[${i}]=$(wc -l <"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed")
$BEDTOOLS intersect -a <(cut -f1-3 "${HOMOCLUSTERDIR}/PU.1_long_hg19_all.cluster.txt")
-b "/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed" -u >
/misc/data/tmp/clusters.kmeans.14.sorted.${i}.bed
CLUSTERPEAKS[${i}]=$(wc -l <"/misc/data/tmp/clusters.kmeans.14.sorted.${i}.bed")
cat >>"${ANALYSISDIR}/ClusterEnrichment.homotypicClustersKmeans.14.txt" <<EOF
homotypic.Cluster${i}$TAB$TOTALPEAKS$TAB$TOTALCLUSTERS$TAB${PEAKS[${i}]}$TAB${CLUSTERPEAKS[${i
}]}
EOF
done

hyperTab.pl ${ANALYSISDIR}/ClusterEnrichment.homotypicClustersKmeans.14.txt >
${ANALYSISDIR}/ClusterEnrichment.homotypicClustersKmeans.14.fisher.txt

# is there a preferred distance between PU.1 motifs in pairs?

# define all unbound pairs
$BEDTOOLS intersect -a <(cut -f1-4 "${HOMOCLUSTERDIR}/PU.1_long_hg19_all.pairs.txt") -b
"${PU1CTV1PEAKS}" -v > ${TMPDIR}/tmp1.unbound_pairs.txt
join -1 4 -2 4 -t $'\t' <(sort -k4,4 ${TMPDIR}/tmp1.unbound_pairs.txt) <(sort -k4,4
${HOMOCLUSTERDIR}/PU.1_long_hg19_all.pairs.txt) | cut -f 12,13 >
${TMPDIR}/tmp.unbound_pairs.txt
# 276629/302897 are not bound in CTV1

# analysis for all peaks
$BEDTOOLS intersect -a <(cut -f1-4 "${HOMOCLUSTERDIR}/PU.1_long_hg19_all.pairs.txt") -b
"${PU1CTV1PEAKS}" -u > ${TMPDIR}/pairs.all.bed
join -1 4 -2 4 -t $'\t' <(sort -k4,4 ${TMPDIR}/pairs.all.bed) <(sort -k4,4
${HOMOCLUSTERDIR}/PU.1_long_hg19_all.pairs.txt) | cut -f 12,13 > ${TMPDIR}/tmp.bound_pairs.txt

cat >"${TMPDIR}/R.distplot.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}/hist")
library(ggplot2)
library(grid)
# read in unbound data
udata <- read.table("${TMPDIR}/tmp.unbound_pairs.txt", header=F, sep="\t")
ud <- data.frame(udata)
colnames(ud) <- c("distance","orientation")
# seperating sense and antisense pairs
uantisense <- subset(ud, orientation==0)
usense <- subset(ud, orientation==1)
# generating count tables
uas_counttable <- as.data.frame(table(uantisense\$distance))
us_counttable <- as.data.frame(table(usense\$distance))
# read in bound data
bdata <- read.table("${TMPDIR}/tmp.bound_pairs.txt", header=F, sep="\t")
bd <- data.frame(bdata)
colnames(bd) <- c("distance","orientation")
# seperating sense and antisense pairs
bantisense <- subset(bd, orientation==0)
bsense <- subset(bd, orientation==1)
# generating count tables
bas_counttable <- as.data.frame(table(bantisense\$distance))
bs_counttable <- as.data.frame(table(bsense\$distance))
# merge sense & antisense tables and perform calculations (% frequency, hypergeom. test,...)
sense <- merge(bs_counttable, us_counttable, by.x = "Var1" , by.y = "Var1")
sense\$perc.x <- (sense\$Freq.x * 100 / sum(sense\$Freq.x))
sense\$perc.y <- (sense\$Freq.y * 100 / sum(sense\$Freq.y))
antisense <- merge(bas_counttable, uas_counttable, by.x = "Var1" , by.y = "Var1")
antisense\$perc.x <- (antisense\$Freq.x * 100 / sum(antisense\$Freq.x))
antisense\$perc.y <- (antisense\$Freq.y * 100 / sum(antisense\$Freq.y))
maxperc <- round(max(c(sense\$perc.x,sense\$perc.y,antisense\$perc.x,antisense\$perc.y)),0)
+ 1
# perform calculations for sense (% frequency, hypergeom. test,...)
sense\$ratio <- (sense\$Freq.x * 100 / sum(sense\$Freq.x)) / (sense\$Freq.y * 100 /
sum(sense\$Freq.y))
sense\$totpair <- (sum(sense\$Freq.x) + sum(sense\$Freq.y))
sense\$pair <- sum(sense\$Freq.x)
sense\$totbound <- sense\$Freq.x + sense\$Freq.y
sense\$bound <- sense\$Freq.x
sense\$dist <- as.numeric(levels(sense\$Var1))[sense\$Var1]
sense\$hyper <- phyper(sense\$bound,sense\$totbound,sense\$totpair,sense\$pair,lower.tail =
FALSE, log.p = FALSE)
sense\$sig <- factor(as.numeric(sense\$hyper < 0.05))
grobsense <- grobTree(textGrob("sense - sense", x=.05, y=.95, hjust=0,
gp=gpar(col="black", fontsize=6, fontface="italic")))
# generating the sense plot
p <- ggplot(data=sense, aes(x=dist,y=perc.x,fill=sig))
p <- p + geom_bar(stat="identity", position=position_dodge(width = 0),alpha=.5)
p <- p + geom_bar(data=sense, aes(x=dist,y=perc.y), stat="identity",
position=position_dodge(width = 0), fill="gray20", alpha=.3)
p <- p + xlab("Distance between motifs") + ylab("% motifs") + annotation_custom(grobsense)
p <- p + theme_light(base_size=8) + theme(panel.grid.major =
element_blank(),panel.grid.minor = element_blank())
p <- p + scale_x_continuous(expand = c(0,0),limits=c(0,150)) + scale_y_continuous(expand =
c(0,0),limits=c(0,maxperc)) + scale_x_reverse(expand = c(0,0),limits=c(150,0))
p <- p + scale_fill_manual(values=c("cornflowerblue", "red")) + theme(legend.position="none")
pdf(file="PairDist.allPeaks.s.hist.pdf", height=2, width=1.4)
plot(p)
dev.off()
# perform calculations for antisense (% frequency, hypergeom. test,...)
antisense\$ratio <- (antisense\$Freq.x * 100 / sum(antisense\$Freq.x)) / (antisense\$Freq.y
* 100 / sum(antisense\$Freq.y))
antisense\$totpair <- (sum(antisense\$Freq.x) + sum(antisense\$Freq.y))
antisense\$pair <- sum(antisense\$Freq.x)
antisense\$totbound <- antisense\$Freq.x + antisense\$Freq.y
antisense\$bound <- antisense\$Freq.x
antisense\$dist <- as.numeric(levels(antisense\$Var1))[antisense\$Var1]
antisense\$hyper <-
phyper(antisense\$bound,antisense\$totbound,antisense\$totpair,antisense\$pair,lower.tail =
FALSE, log.p = FALSE)
antisense\$sig <- factor(as.numeric(antisense\$hyper < 0.05))
grobantisense <- grobTree(textGrob("sense - antisense", x=0.95, y=0.95, hjust=1,
gp=gpar(col="black", fontsize=6, fontface="italic")))
# generating the antisense plot
p <- ggplot(data=antisense, aes(x=dist,y=perc.x,fill=sig))
p <- p + geom_bar(stat="identity", position=position_dodge(width = 0),alpha=.5)
p <- p + geom_bar(data=antisense, aes(x=dist,y=perc.y), stat="identity",
position=position_dodge(width = 0), fill="gray20", alpha=.3)
p <- p + xlab("Distance between motifs") + ylab("% motifs") + annotation_custom(grobantisense)
p <- p + theme_light(base_size=8) + theme(panel.grid.major =
element_blank(),panel.grid.minor = element_blank())
p <- p + scale_x_continuous(expand = c(0,0),limits=c(0,150)) + scale_y_continuous(expand =
c(0,0),limits=c(0,maxperc),position="right")
p <- p + scale_fill_manual(values=c("cornflowerblue", "red")) + theme(legend.position="none")
pdf(file="PairDist.allPeaks.as.hist.pdf", height=2, width=1.4)
plot(p)
dev.off()
EOF
chmod 750 "\${TMPDIR}/R.distplot.P.${_DATE}.R"
R < ${TMPDIR}/R.distplot.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.distplot.P.${_DATE}.R

# ---> across all peaks, mainly distances between 12-50 bp are enriched
# ---> should do further analyses with 12-50bp pairs

# analysis for individual clusters
declare -a COLORS=(gold goldenrod2 darkorange1 darkorange3 firebrick1 firebrick3 firebrick
deeppink darkmagenta deepskyblue2 dodgerblue1 dodgerblue3 blue2 blue4)
COUNT=0
_DATE=$(date +%s)
for ((i=1;i<15;i++));do
COLOR="${COLORS[${COUNT}]}"
$BEDTOOLS intersect -a <(cut -f1-4 "${HOMOCLUSTERDIR}/PU.1_long_hg19_all.pairs.txt") -b
"/misc/data/tmp/PU1_Flag_merged.ATACann.kmeans.14.sorted.${i}.bed" -u >
${TMPDIR}/pairs.kmeans.14.sorted.${i}.bed
join -1 4 -2 4 -t $'\t' <(sort -k4,4 ${TMPDIR}/pairs.kmeans.14.sorted.${i}.bed) <(sort
-k4,4 ${HOMOCLUSTERDIR}/PU.1_long_hg19_all.pairs.txt) | cut -f 12,13 >
${TMPDIR}/tmp.bound_pairs${i}.txt
cat >"${TMPDIR}/R.distplot${i}.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}/hist")
library(ggplot2)
library(grid)
# read in unbound data
udata <- read.table("${TMPDIR}/tmp.unbound_pairs.txt", header=F, sep="\t")
ud <- data.frame(udata)
colnames(ud) <- c("distance","orientation")
# seperating sense and antisense pairs
uantisense <- subset(ud, orientation==0)
usense <- subset(ud, orientation==1)
```

```
# generating count tables
uas_counttable <- as.data.frame(table(uantisense\$distance))
us_counttable <- as.data.frame(table(usense\$distance))
# read in bound data
bdata <- read.table("${TMPDIR}/tmp.bound_pairs${i}.txt", header=F, sep="\t")
bd <- data.frame(bdata)
colnames(bd) <- c("distance","orientation")
# seperating sense and antisense pairs
bantisense <- subset(bd, orientation==0)
bsense <- subset(bd, orientation==1)
# generating count tables
bas_counttable <- as.data.frame(table(bantisense\$distance))
bs_counttable <- as.data.frame(table(bsense\$distance))
# merge sense & antisense tables and perform calculations (% frequency, hypergeom. test,...)
sense <- merge(bs_counttable, us_counttable, by.x = "Var1" , by.y = "Var1")
sense\$perc.x <- (sense\$Freq.x * 100 / sum(sense\$Freq.x))
sense\$perc.y <- (sense\$Freq.y * 100 / sum(sense\$Freq.y))
antisense <- merge(bas_counttable, uas_counttable, by.x = "Var1" , by.y = "Var1")
antisense\$perc.x <- (antisense\$Freq.x * 100 / sum(antisense\$Freq.x))
antisense\$perc.y <- (antisense\$Freq.y * 100 / sum(antisense\$Freq.y))
maxperc <- round(max(c(sense\$perc.x,sense\$perc.y,antisense\$perc.x,antisense\$perc.y)),0)
+ 1
# perform calculations for sense (% frequency, hypergeom. test,...)
sense\$ratio <- (sense\$Freq.x * 100 / sum(sense\$Freq.x)) / (sense\$Freq.y * 100 /
sum(sense\$Freq.y))
sense\$totpair <- (sum(sense\$Freq.x) + sum(sense\$Freq.y))
sense\$pair <- sum(sense\$Freq.x)
sense\$totbound <- sense\$Freq.x + sense\$Freq.y
sense\$bound <- sense\$Freq.x
sense\$dist <- as.numeric(levels(sense\$Var1))[sense\$Var1]
sense\$hyper <- phyper(sense\$bound,sense\$totbound,sense\$totpair,sense\$pair,lower.tail =
FALSE, log.p = FALSE)
sense\$sig <- factor(as.numeric(sense\$hyper < 0.05))
grobsense <- grobTree(textGrob("sense - sense", x=.05, y=.95, hjust=0,
gp=gpar(col="black", fontsize=6, fontface="italic")))
# generating the sense plot
p <- ggplot(data=sense, aes(x=dist,y=perc.x,fill=sig))
p <- p + geom_bar(stat="identity", position=position_dodge(width = 0),alpha=.5)
p <- p + geom_bar(data=sense, aes(x=dist,y=perc.y), stat="identity",
position=position_dodge(width = 0), fill="gray20", alpha=.3)
p <- p + xlab("Distance between motifs") + ylab("% motifs") + annotation_custom(grobsense)
p <- p + theme_light(base_size=8) + theme(panel.grid.major =
element_blank(),panel.grid.minor = element_blank())
p <- p + scale_x_continuous(expand = c(0,0),limits=c(150,0)) + scale_y_continuous(expand =
c(0,0),limits=c(0,maxperc)) + scale_x_reverse()
p <- p + scale_fill_manual(values=c("$COLOR", "springgreen2")) +
theme(legend.position="none")
pdf(file="PairDist.Kmeans${i}.s.hist.pdf", height=2, width=2)
plot(p)
dev.off()
# perform calculations for antisense (% frequency, hypergeom. test,...)
antisense\$ratio <- (antisense\$Freq.x * 100 / sum(antisense\$Freq.x)) / (antisense\$Freq.y
* 100 / sum(antisense\$Freq.y))
antisense\$totpair <- (sum(antisense\$Freq.x) + sum(antisense\$Freq.y))
antisense\$pair <- sum(antisense\$Freq.x)
antisense\$totbound <- antisense\$Freq.x + antisense\$Freq.y
antisense\$bound <- antisense\$Freq.x
antisense\$dist <- as.numeric(levels(antisense\$Var1))[antisense\$Var1]
antisense\$hyper <-
phyper(antisense\$bound,antisense\$totbound,antisense\$totpair,antisense\$pair,lower.tail =
FALSE, log.p = FALSE)
antisense\$sig <- factor(as.numeric(antisense\$hyper < 0.05))
grobantisense <- grobTree(textGrob("sense - antisense", x=0.95, y=0.95, hjust=1,
gp=gpar(col="black", fontsize=6, fontface="italic")))
# generating the antisense plot
p <- ggplot(data=antisense, aes(x=dist,y=perc.x,fill=sig))
p <- p + geom_bar(stat="identity", position=position_dodge(width = 0),alpha=.5)
p <- p + geom_bar(data=antisense, aes(x=dist,y=perc.y), stat="identity",
position=position_dodge(width = 0), fill="gray20", alpha=.3)
p <- p + xlab("Distance between motifs") + ylab("% motifs") + annotation_custom(grobantisense)
p <- p + theme_light(base_size=8) + theme(panel.grid.major =
element_blank(),panel.grid.minor = element_blank())
p <- p + scale_x_continuous(expand = c(0,0),limits=c(0,150)) + scale_y_continuous(expand =
c(0,0),limits=c(0,maxperc))
p <- p + scale_fill_manual(values=c("$COLOR", "springgreen2")) +
theme(legend.position="none")
pdf(file="PairDist.Kmeans${i}.as.hist.pdf", height=2, width=2)
plot(p)
```

```
dev.off()
EOF
chmod 750 "\${TMPDIR}/R.distplot${i}.P.${_DATE}.R"
R < ${TMPDIR}/R.distplot${i}.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.distplot${i}.P.${_DATE}.R
COUNT=$((COUNT+=1))
done

# "ghistplot" for the presence of the cluster motif
CLUSTERMOTIF="${HOMOCLUSTERDIR}/motifs/PU1_pairs/final/homerResults/motif2.motif"
annotatePeaks.pl ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt hg19
-size 200 -center ${CLUSTERMOTIF} -nogene -noann > ${TMPDIR}/tmp.cluster.peaks.mot.ann.txt
pos2bed.pl ${TMPDIR}/tmp.cluster.peaks.mot.ann.txt >${TMPDIR}/tmp.cluster.peaks.mot.ann.bed
$BEDTOOLS intersect -a ${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.bed
-b ${TMPDIR}/tmp.cluster.peaks.mot.ann.bed -c > ${TMPDIR}/tmp.clusterMotif.peaks.bound.bed
awk 'BEGIN{OFS="\t"} NR==FNR{a[$4]=$7;next}{print $0,a[$1]}'
${TMPDIR}/tmp.clusterMotif.peaks.bound.bed
${ANALYSISDIR}/PU1_Flag_merged.ATACann.kmeans.14.sorted.cleaned.txt >
${TMPDIR}/tmp.ann.clusterMotif.sorted.txt
_DATE=$(date +%s)
cat >"${TMPDIR}/R.image.P.${_DATE}.R" <<EOF
setwd("${FIGURESDIR}")
library(RColorBrewer)
data <- read.delim("${TMPDIR}/tmp.ann.clusterMotif.sorted.txt", header=T)
d <- data.matrix(data[,7])
q <- apply(d, 2, function(x) ifelse(x > 1, 1 ,x))
o <- q[order(nrow(q):1),]
mycol <- colorRampPalette(c("white","${colCluM}"))(3)
png(filename="ClusterMotifInK14cluster.png", height=22315, width=500)
par(mar = rep(0, 4))
image(t(o),col=mycol, zlim=range(c(0,1)))
dev.off()
EOF
chmod 750 "${TMPDIR}/R.image.P.${_DATE}.R"
R < ${TMPDIR}/R.image.P.${_DATE}.R --no-save
rm ${TMPDIR}/R.image.P.${_DATE}.R
```

### 10.1.9 Analysis of the Binding Properties of PU.1-fusion Proteins and Corresponding Mass-spectrometry Data

The following script was used to analyze the binding properties of PU.1-fusion proteins in CTV-1 cells as well as their corresponding mass spectrometry data. The script includes parts of edgeR (Robinson et al. 2010), the HOMER suite (Heinz et al. 2010) as well as parts of the R software (R Development Core Team 2008).

```
#!/bin/bash
# setting homer environment
DIR_PKG="/misc/software/ngs"
PATH_PERL=/misc/software/package/perl/perl-5.26.1/bin
PATH_SAMTOOLS=${DIR_PKG}/samtools/samtools-1.6/bin
PATH_HOMER=${DIR_PKG}/homer/v4.9/bin
PATH_R=/misc/software/package/RBioC/3.4.3/bin
export PATH=${PATH_R}:${PATH_PERL}:${PATH_SAMTOOLS}:${PATH_HOMER}:${PATH}
export PATH

WORKDIR="/misc/data/analysis/project_PU1/CTV1/ChIP"
TAGDIR="/misc/data/processedData/tagDir/chromatin/hg19/ChIP/RNAtransfection/PU1"
PEAKDIR="${WORKDIR}/peaksNoCNVcorr"
BLACKLIST_HG19="/misc/data/analysis/generalStuff/annotation/hg19/hg19.blacklist.bed"
FIGURESDIR="/misc/data/analysis/project_PU1/CTV1/figures/BirAvsFlag"
BEDTOOLS="/misc/software/ngs/bedtools/bedtools2-2.27.1/bin/bedtools"
ANALYSISDIR="/misc/data/analysis/project_PU1/BioID"
TMPDIR="/loctmp"

mkdir -p ${FIGURESDIR}
mkdir -p ${PEAKDIR}
```

```
# Analysis of the binding properties of PU.1-fusion Proteins

# normalize & reduce tagDir for better comparability with other data sets
declare -a oCHIPDIRS=("${TAGDIR}/CTV1_PU1_PU1_R1" "${TAGDIR}/CTV1_PU1_PU1_R2"
"${TAGDIR}/CTV1_PU1_PU1_R3")
declare -a oINPUTDIRS=("${TAGDIR}/CTV1_PU1mut_PU1_R1" "${TAGDIR}/CTV1_PU1mut_PU1_R2"
"${TAGDIR}/CTV1_PU1mut_PU1_R3")

COUNT=0
for SAMPLE in ${oCHIPDIRS[@]}; do
INPUT="${oINPUTDIRS[${COUNT}]}"
normalizeTagDirByCopyNumber.pl ${SAMPLE} -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl ${INPUT} -cnv "${CNVFILECTV1}" -remove
COUNT=$((COUNT+=1))
done

cd ${TAGDIR}
makeTagDirectory CTV1_PU1_PU1_merged -d CTV1_PU1_PU1_R1 CTV1_PU1_PU1_R2 CTV1_PU1_PU1_R3
-genome hg19 -checkGC
makeTagDirectory CTV1_PU1mut_PU1_merged -d CTV1_PU1mut_PU1_R1 CTV1_PU1mut_PU1_R2
CTV1_PU1mut_PU1_R3 -genome hg19 -checkGC

# normalization of BirA tagDirs (only one replicate)
cd ${TAGDIR}
normalizeTagDirByCopyNumber.pl CTV1_PU.1-BirA_PU1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU.1delA-BirA_PU1 -cnv "${CNVFILECTV1}" -remove
normalizeTagDirByCopyNumber.pl CTV1_PU.1delQ-BirA_PU1 -cnv "${CNVFILECTV1}" -remove

# BigWigs for normalized TagDirs

# generating individual bigwigs
TAGDIRSETS="CTV1_PU1_PU1_R1_CNVnormRefChr CTV1_PU1_PU1_R2_CNVnormRefChr
CTV1_PU1_PU1_R3_CNVnormRefChr \
CTV1_PU1mut_PU1_R1_CNVnormRefChr CTV1_PU1mut_PU1_R2_CNVnormRefChr
CTV1_PU1mut_PU1_R3_CNVnormRefChr \
CTV1_PU.1-BirA_PU1_CNVnormRefChr CTV1_PU.1delA-BirA_PU1_CNVnormRefChr
CTV1_PU.1delQ-BirA_PU1_CNVnormRefChr "


for SAMPLE in ${TAGDIRSETS[@]}; do
makeUCSCfile ${CHIPTAGDIR}/$SAMPLE -bigWig $CHROMSIZES -o $BIGWIGDIR/$SAMPLE.bigwig
done

# average bigwigs from triplicates
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_PU1_PU1_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1_PU1_R2_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1_PU1_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_avePU1_PU1_CNVnormRefChr.bigwig
myAverageBigWig.pl -bw $BIGWIGDIR/CTV1_PU1mut_PU1_R1_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1mut_PU1_R2_CNVnormRefChr.bigwig \
$BIGWIGDIR/CTV1_PU1mut_PU1_R3_CNVnormRefChr.bigwig \
-chr ${CHROMSIZES} -o $BIGWIGDIR/CTV1_avePU1_PU1mut_CNVnormRefChr.bigwig

# comparison of PU.1-tags wt vs PU.1-BirA

cd ${PEAKDIR}
mergePeaks -d 100 CTV1_PU1_PU1_merged.factor.fdr05.peaks.txt
${PEAKDIR}/CTV1_PU.1-BirA_PU1.factor.fdr05.peaks.txt > /loctmp/tmp.1.txt
pos2bed.pl /loctmp/tmp.1.txt > /loctmp/tmp.1.bed
$BEDTOOLS intersect -a /loctmp/tmp.1.bed -b $BLACKLIST_HG19 -v > /loctmp/tmp.2.bed
filter4Mappability.sh -p /loctmp/tmp.2.bed -g hg19 -f 0.8 -s 50
pos2bed.pl /loctmp/tmp.2.mapScoreFiltered.txt > CTV1_PU1_BirA.merged.filtered.bed

annotatePeaks.pl CTV1_PU1_BirA.merged.filtered.bed hg19 -size 200 -d
${TAGDIR}/CTV1_PU1_PU1_merged ${TAGDIR}/CTV1_PU.1-BirA_PU1 -noann -nogene >
CTV1_PU1_BirA.merged.filtered.ann.txt

cd ${PEAKDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.scatter.${_DATE}.R" <<EOF
library(ggplot2)
library(MASS)
library(scales)
data <- read.delim("CTV1_PU1_BirA.merged.filtered.ann.txt", header=T)
data.red <- data[,c(8:9)]
colnames(data.red) <- c("PU1","PU1_BirA")
attach(data.red)
d <- data.frame(log10 (data.red + 0.1))
```

```
lm_eqn = function(d){
m = lm(PU1 ~ PU1_BirA, d);
eq <- substitute(italic(r)^2~"="~r2,
list(r2 = format(summary(m)\$r.squared, digits = 3)))
as.character(as.expression(eq)); }
xlabel = expression("Tag count PU.1")
ylabel = expression("Tag count PU.1-BirA")
p <- ggplot(data.red,aes(x=PU1, y=PU1_BirA)) + coord_trans(x="log10",y="log10")
p <- p + theme_bw(base_size = 12, base_family = "Helvetica") +
coord_cartesian(xlim=c(1,300),ylim=c(1,300))
p <- p + scale_y_continuous(trans = 'log10', breaks = c(1,10,100), labels = c(1,10,100))
p <- p + scale_x_continuous(trans = 'log10', breaks = c(1,10,100), labels = c(1,10,100))
p <- p + geom_jitter(size=.25,alpha=0.02,shape=20,fill="blue",color="blue",width=.1,height
=.1)
p <- p + annotate("text", x = 100, y = 1.4, label = lm_eqn(d), size = 4, colour="black",
parse = TRUE)
p <- p + annotation_logticks(base = 10, short = unit(0.05, "cm"), mid = unit(0.10, "cm"),
long = unit(0.15, "cm"))
p <- p + labs(x = xlabel, y = ylabel)
pdf(file="${FIGURESDIR}/Scatter.peaksPU1.PU1vsPU1_BirA.comb.pdf", height=3, width=3)
plot(p)
dev.off()
EOF
chmod 750 "/loctmp/R.scatter.${_DATE}.R"
R < /loctmp/R.scatter.${_DATE}.R --no-save
rm /loctmp/R.scatter.${_DATE}.R
```

```
# Analysis of obtained mass spectrometry data
```

```
# Volcano plot of MS data of THP-1 cells
cd ${ANALYSISDIR}
cut -f1-3,6 THP1_volcano_plot_Pu1_vs_NLS.txt > ${TMPDIR}/tmp.THP1_Pu1_vs_NLS_volcplot.txt

# PU1 vs. BirA generate volcano (only GO terms for myeloid differentiation and chromatin
remodeling were selected in Metascape to be displayed)
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.volcano.P.${_DATE}.R" <<EOF
library(ggplot2)
library(ggrepel)
res <- read.table("${TMPDIR}/tmp.THP1_Pu1_vs_NLS_volcplot.txt", header=T, sep="\t")
colnames(res) <- c("sig","Pvalue","logFC","GeneSymbol")
genelist <-
c("IKZF1","KMT2D","TET2","KMT2C","TCF12","SPI1","NFATC2","LGALS3","JMJD1C","EMSY","HCFC1","HNR
NPU","EP400","CHD8","SMARCA4","ARID1B","SMARCC2","SMARCE1","NUCKS1","ACTL6A","BRD7","BCORL1","
NUP133","MBD3","UBE2E1","ELP4","KDM6B","DMAP1","ANP32A","ANP32D","ANP32C","NCOA6","HCLS1","FLI
1","FOXP1","RPS17","KRAS")
head(res)
# Highlight genes that pass FDR
res\$threshold = as.factor(res\$sig == "+")
res\$namethresh = as.factor(res\$sig == "+")
res\$fCategory <- factor(res\$sig)
# Construct the plot object
p <- ggplot(data=res, aes(x=logFC, y=Pvalue, color=threshold))
p <- p + theme_bw(base_size = 8, base_family = "Helvetica") + theme(legend.position = "none")
p <- p + geom_point(alpha=0.75, size=0.750) +
scale_color_manual(values=c("FALSE"="gray80","TRUE"="blue"))
p <- p + xlim(c(-8, 14)) + ylim(c(0, 6))
p <- p + xlab("log2 fold change") + ylab("-log10 p-value")
p <- p + theme(panel.grid.major = element_line(size = .25, color = "grey"),panel.grid.minor
= element_line(size = .25, color = "grey"), panel.border = element_rect(size=.5, color =
"black"))
p <- p + geom_text_repel(data=subset(res, res\$GeneSymbol %in% genelist) , aes(x=logFC,
y=Pvalue,label=GeneSymbol), force = 1.5, segment.colour="black", segment.size=0.05,
min.segment.length=0.05, size=2, box.padding= .25, point.padding=.1, segment.alpha=0.5,
alpha=1, color="black")
pdf(file="${ANALYSISDIR}/THP1_Volcano_stat_Pu1_vs_NLS.pdf", height=3, width=4)
print(p)
dev.off()
EOF
chmod 750 "/loctmp/R.volcano.P.${_DATE}.R"
R < /loctmp/R.volcano.P.${_DATE}.R --no-save
rm /loctmp/R.volcano.P.${_DATE}.R
```

```
# Volcano plot of MS data of K-562 cells
cd ${ANALYSISDIR}
cut -f1-3,7 K562_Pu1_vs_NLS_volcplot_ttest_FDRcorr.txt >
${TMPDIR}/tmp.K562_Pu1_vs_NLS_volcplot.txt

# PU1 vs. BirA generate volcano
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.volcano.P.${_DATE}.R" <<EOF
library(ggplot2)
library(ggrepel)
res <- read.table("${TMPDIR}/tmp.K562_Pu1_vs_NLS_volcplot.txt", header=T, sep="\t")
colnames(res) <- c("sig","Pvalue","logFC","GeneSymbol")
genelist <- c("ARID1A", "SPI1", "TAL1", "ARID2", "ARID1B" , "SMARCE1")
head(res)
# Highlight genes that pass FDR
res\$threshold = as.factor(res\$sig == "+")
res\$namethresh = as.factor(res\$sig == "+")
res\$fCategory <- factor(res\$sig)
# Construct the plot object
p <- ggplot(data=res, aes(x=logFC, y=Pvalue, color=threshold))
p <- p + theme_bw(base_size = 8, base_family = "Helvetica") + theme(legend.position = "none")
p <- p + geom_point(alpha=0.75, size=0.750) +
scale_color_manual(values=c("FALSE"="gray80","TRUE"="blue"))
p <- p + xlim(c(-8, 10)) + ylim(c(0, 6))
p <- p + xlab("log2 fold change") + ylab("-log10 p-value")
p <- p + theme(panel.grid.major = element_line(size = .25, color = "grey"),panel.grid.minor
= element_line(size = .25, color = "grey"), panel.border = element_rect(size=.5, color =
"black"))
p <- p + geom_text_repel(data=subset(res, res\$GeneSymbol %in% genelist) ,aes(x=logFC,
y=Pvalue,label=GeneSymbol), force = .75, segment.colour="black", segment.size=0.05,
min.segment.length=0.05, size=2, box.padding= .15, point.padding=.05, segment.alpha=0.5,
alpha=1, color="black")
pdf(file="${ANALYSISDIR}/K562_Volcano_stat_Pu1_vs_NLS.pdf", height=3, width=4)
print(p)
dev.off()
EOF
chmod 750 "/loctmp/R.volcano.P.${_DATE}.R"
R < /loctmp/R.volcano.P.${_DATE}.R --no-save
rm /loctmp/R.volcano.P.${_DATE}.R

# Volcano plot of MS data of CTV-1 cells
cd ${ANALYSISDIR}
cut -f1-3,6 ctv1_Pu1_vs_NLS_volcplot_ttest_FDRcorr.txt > ${TMPDIR}/tmp.Pu1_vs_NLS_volcplot.txt
cut -f1-3,6 ctv1_Pu1_vs_delA_volcplot_ttest_FDRcorr.txt >
${TMPDIR}/tmp.Pu1_vs_delA_volcplot.txt

# PU1 vs. BirA generate volcano (only GO terms for chromatin remodeling and interesting TFs
were selected to be displayed)
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.volcano.P.${_DATE}.R" <<EOF
library(ggplot2)
library(ggrepel)
res <- read.table("${TMPDIR}/tmp.Pu1_vs_NLS_volcplot.txt", header=T, sep="\t")
colnames(res) <- c("sig","Pvalue","logFC","GeneSymbol")
genelist <-
c("SMARCA2","ARID3B","SMARCA4","PBRM1","SMARCD2","FLI1","ARID2","SMARCE1","ACTL6A","PDHX","TAL
1","ZBTB4","LDB1","ARID1A","ARID1B","SPI1","KMT2D","KMT2C")
head(res)
# Highlight genes that pass FDR
res\$threshold = as.factor(res\$sig == "+")
res\$namethresh = as.factor(res\$sig == "+")
res\$fCategory <- factor(res\$sig)
# Construct the plot object
p <- ggplot(data=res, aes(x=logFC, y=Pvalue, color=threshold))
p <- p + theme_bw(base_size = 8, base_family = "Helvetica") + theme(legend.position = "none")
p <- p + geom_point(alpha=0.75, size=0.750) +
scale_color_manual(values=c("FALSE"="gray80","TRUE"="blue"))
p <- p + xlim(c(-8, 12)) + ylim(c(0, 6))
p <- p + xlab("log2 fold change") + ylab("-log10 p-value")
p <- p + theme(panel.grid.major = element_line(size = .25, color = "grey"),panel.grid.minor
= element_line(size = .25, color = "grey"), panel.border = element_rect(size=.5, color =
"black"))
p <- p + geom_text_repel(data=subset(res, res\$GeneSymbol %in% genelist) , aes(x=logFC,
y=Pvalue,label=GeneSymbol), force = .75, segment.colour="black", segment.size=0.05,
min.segment.length=0.05, size=2, box.padding= .15, point.padding=.05, segment.alpha=0.5,
alpha=1, color="black")
```

```
pdf(file="${ANALYSISDIR}/Volcano_stat_Pu1_vs_NLS.pdf", height=3, width=4)
print(p)
dev.off()
EOF
chmod 750 "/loctmp/R.volcano.P.${_DATE}.R"
R < /loctmp/R.volcano.P.${_DATE}.R --no-save
rm /loctmp/R.volcano.P.${_DATE}.R

# PU1 vs. delA generate volcano (same genes as above were selected to be displayed)
cd ${ANALYSISDIR}
_DATE=$(date +%s)
cat >"/loctmp/R.volcano.P.${_DATE}.R" <<EOF
library(ggplot2)
library(ggrepel)
res <- read.table("${TMPDIR}/tmp.Pu1_vs_delA_volcplot.txt", header=T, sep="\t")
colnames(res) <- c("sig","Pvalue","logFC","GeneSymbol")
genelist <-
c("SMARCA2","ARID3B","SMARCA4","PBRM1","SMARCD2","FLI1","ARID2","SMARCE1","ACTL6A","PDHX","TAL
1","ZBTB4","LDB1","ARID1A","ARID1B","SPI1","KMT2D","KMT2C")
head(res)
# Highlight genes that pass FDR
res\$threshold = as.factor(res\$sig == "+")
res\$namethresh = as.factor(res\$sig == "+")
res\$fCategory <- factor(res\$sig)
# Construct the plot object
p <- ggplot(data=res, aes(x=logFC, y=Pvalue, color=threshold))
p <- p + theme_bw(base_size = 8, base_family = "Helvetica") + theme(legend.position = "none")
p <- p + geom_point(alpha=0.75, size=0.750) +
scale_color_manual(values=c("FALSE"="gray80","TRUE"="blue"))
p <- p + xlim(c(-8, 8)) + ylim(c(0, 6))
p <- p + xlab("log2 fold change") + ylab("-log10 p-value")
p <- p + theme(panel.grid.major = element_line(size = .25, color = "grey"),panel.grid.minor
= element_line(size = .25, color = "grey"), panel.border = element_rect(size=.5, color =
"black"))
p <- p + geom_text_repel(data=subset(res, res\$GeneSymbol %in% genelist) , aes(x=logFC,
y=Pvalue,label=GeneSymbol), force = .75, segment.colour="black", segment.size=0.05,
min.segment.length=0.05, size=2, box.padding= .15, point.padding=.05, segment.alpha=0.5,
alpha=1, color="black")
pdf(file="${ANALYSISDIR}/Volcano_stat_Pu1_vs_delA.pdf", height=3, width=4)
print(p)
dev.off()
EOF
chmod 750 "/loctmp/R.volcano.P.${_DATE}.R"
R < /loctmp/R.volcano.P.${_DATE}.R --no-save
rm /loctmp/R.volcano.P.${_DATE}.R
```

## 10.2 Supplementary Tables and Figures

### 10.2.1 Appendix I – Cell Type-Specific PU.1 Binding Site Selection

The following table lists all public available PU.1-ChIPseq data used in this study. The exact sample description and the NCBI GEO SRA (Sequence Read Archive) number is depicted.

**Table 10-1 - Public available PU.1 ChIPseq data used in this study**

| Sample | NCBI GEO SRA number |
|---|---|
| DOHH2-PU.1 | SRR2050990 |
| DOHH2-Input | SRR2050987 |
| GM12878-PU.1 | SRR351880<br>SRR351881<br>SRR578180<br>SRR578181 |
| GM12878-Input | SRR351535 |
| H929-PU.1 | SRR1240634 |
| H929-Input | SRR1240632 |
| HPC-PU.1 | SRR094808<br>SRR094809 |
| HPC-Input | SRR094805 |
| K562-PU.1-R1 | SRR351605 |
| K562-Input-R1 | SRR351541 |
| K562-PU.1-R2.1 | SRR2085865 |
| K562-Input-R2.1 | SRR2085862 |
| K562-PU.1-R2.2 | SRR2085866 |
| K562-Input-R2.2 | SRR2085863 |
| OCILY7-PU.1 | SRR2051012 |
| OCILY7-Input | SRR2051009 |
| RS411-Dex1h-PU.1 | SRR2138409 |
| RS411-Dex1h-Input | SRR2138408 |
| RS411-NotStim-PU.1 | SRR2138399 |
| RS411- NotStim-Input | SRR2138398 |

Ana-Karina Da Silva Mendes, Anna Ratermann, Dr. Claudia Gebhard, Dagmar Glatz, Dr. Christian Schmidl and Lucia Schwarzfischer-Pfeilschifter generated additional preliminary data.

### 10.2.2 Appendix II – Epigenetic Determinants of PU.1 Binding Site Selection

**Table 10-2 - Targeted Bisulfite Amplicon Sequencing Data**

| Sample | pos. 63 | pos.87 | pos.96 | pos.106 | pos.109 | pos.138 | pos.189 | pos.193 | pos.216 |
|---|---|---|---|---|---|---|---|---|---|
| gDNA | 94.2 | 97.8 | 96.5 | 98.8 | 96.2 | 96.8 | 97.7 | 90.1 | 88.1 |
| DMSO | 94.5 | 99.2 | 99.2 | 95.3 | 97.8 | 96 | 99.5 | 84.7 | 95.5 |
| 10 nM | 66.4 | 89.8 | 81.8 | 96 | 86.2 | 86.3 | 94.9 | 73.3 | 91 |
| 100 nM | 25.2 | 41.5 | 41.3 | 58.2 | 45.9 | 42.1 | 50.2 | 35.6 | 33.5 |
| 300 nM | 24.8 | 32.6 | 38.5 | 50.2 | 41.6 | 34.2 | 40.5 | 31.4 | 34.1 |
| 1000 nM | 37.8 | 51.5 | 68.4 | 72.2 | 59.9 | 59.3 | 59.7 | 51.4 | 57.9 |
| 100 nM_PU1_R1 | 28.8 | 47.1 | 53.3 | 65.9 | 55.8 | 48.8 | 57.8 | 41.4 | 41.5 |
| 100 nM_PU1mut_R1 | 21.6 | 40.5 | 42.6 | 58.3 | 49.4 | 42.8 | 51.4 | 34 | 39.3 |
| 100 nM_PU1_R2 | 22.2 | 41 | 47.7 | 68.4 | 57.9 | 48.1 | 61 | 41.8 | 40.3 |
| 100 nM_PU1mut_R2 | 20 | 47.2 | 54.4 | 67 | 54.7 | 46 | 61.6 | 40.2 | 41.5 |
| 100 nM_PU1_R3 | 30.1 | 48.1 | 57.1 | 70.6 | 57.2 | 53.9 | 60.8 | 41.1 | 47.1 |
| 100 nM_PU1mut_R3 | 27.7 | 46.7 | 51.7 | 68.1 | 55.4 | 48.8 | 58.1 | 39 | 45.1 |

| Sample | pos.235 | pos.240 | pos.252 | pos.258 | pos.260 | pos.265 |
|---|---|---|---|---|---|---|
| gDNA | 81.6 | 98.4 | 99.6 | 96.4 | 99.2 | 98.6 |
| DMSO | 82 | 98.3 | 99.8 | 98.1 | 98.8 | 96 |
| 10 nM | 68.9 | 97.8 | 99.7 | 99.6 | 99.7 | 83.9 |
| 100 nM | 35.9 | 47.7 | 69 | 55.8 | 57.7 | 48.2 |
| 300 nM | 35.1 | 47.8 | 71.3 | 57.3 | 56.8 | 45.8 |
| 1000 nM | 58.5 | 63.2 | 79.4 | 79 | 72.1 | 64.3 |
| 100 nM_PU1_R1 | 42.8 | 56.3 | 83.1 | 72.6 | 73.2 | 63.3 |
| 100 nM_PU1mut_R1 | 40.3 | 56.2 | 79.5 | 67.5 | 73.1 | 60.8 |
| 100 nM_PU1_R2 | 46.1 | 58 | 84.2 | 71.4 | 78.6 | 65.5 |
| 100 nM_PU1mut_R2 | 45.6 | 57.7 | 82.3 | 72.5 | 74.7 | 60.2 |
| 100 nM_PU1_R3 | 48.7 | 62.8 | 87.1 | 77.6 | 76.3 | 68.6 |
| 100 nM_PU1mut_R3 | 46.5 | 59.5 | 81.9 | 68.5 | 71.4 | 63.8 |

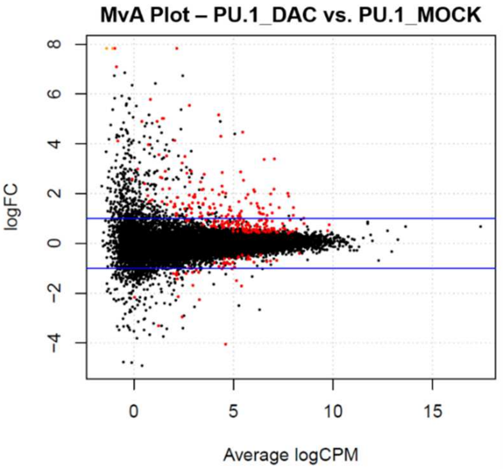


**Figure 10-1 - Multi variance analysis of differential expressed genes between PU.1-transfected CTV-1 cells treated with DAC or left untreated**

The expression ratio (logFC) of the DEGs is plotted for both conditions against their average expression intensity (logCPM). Each point represents one gene. The blue line depicts a logFC of 1. Genes of interest are located above or rather below the blue line, representing the genes, which are at least 2-fold different between PU.1-transfected DAC-treated cells (above) and untreated CTV-1 cells (below).

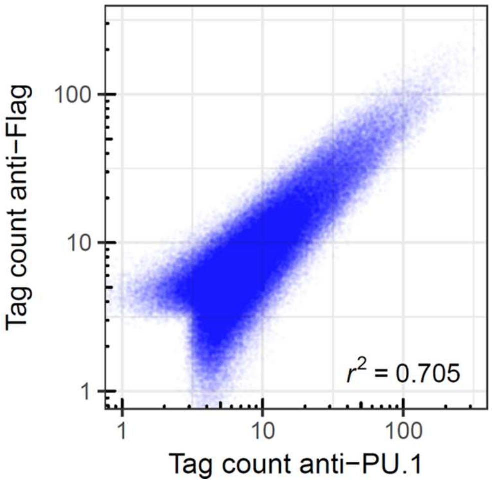


**Figure 10-2 - Scatter plot comparing read counts of CTV-1-specific PU.1 reads**

Read counts obtained in PU.1-transfected CTV-1 cells obtained with an anti-FLAG or anti-PU.1 antibody respectively are plotted against each other. The Pearson correlation (coefficient of determination) is depicted ($r^2$=0.705).

### 10.2.3 Appendix III – PU.1 Binding Site Selection in Lymphoid CTV-1 Cells

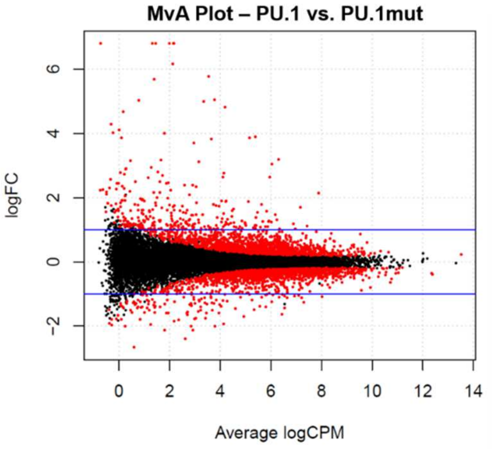


**Figure 10-3 - Multi variance analysis of differential expressed genes between PU.1- and PU.1mut transfected CTV-1 cells**

The expression ratio (logFC) of the DEGs is plotted for both conditions against their average expression intensity (logCPM). Each point represents one gene. The blue line depicts a logFC of 1. Genes of interest are located above or rather below the blue line, representing the genes, which are at least 2-fold different between PU.1-transfected (above) and PU.1mut-transfected (below) CTV-1 cells.

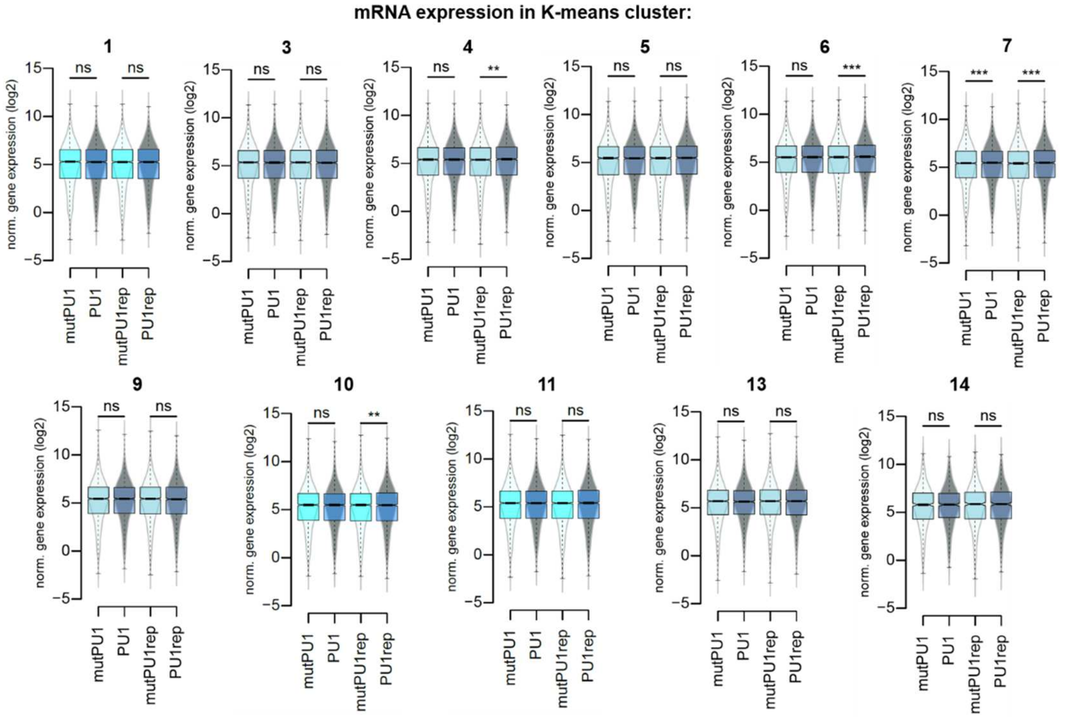


**Figure 10-4 - mRNA expression across remaining K-means ATAC cluster**

The mRNA expression of CTV-1 cells transfected with PU.1 or PU.1mut mRNA is shown in association with the generated K-means cluster of the ATAC signals of the transfected cells. Short-term (mutPU1, PU1) and repetitive long-term transfections (mutPU1rep, PU1rep) used for RNAseq are depicted.

### 10.2.4 Appendix IV – Binding Site Selection of PU.1-deletion mutants

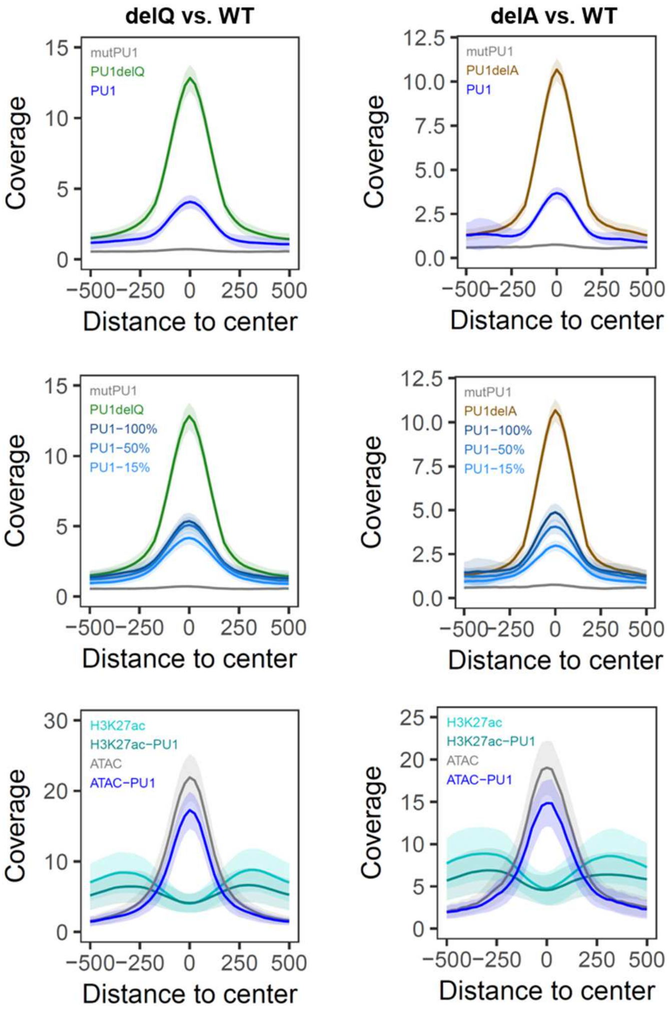


**Figure 10-5 - Distribution of differential peaks of PU.1-mutants in CTV-1 cells**

Histogram plots showing the ChIPseq coverage (y-axis) of PU.1 reads of PU.1 (blue)- and PU.1mut (grey)-transfected CTV-1 cells, as well as of cells transfected with PU.1-delQ (green) and PU.1-delA (brown) across differential peak sets (delQ vs. WT; delA vs. WT). The distance to the PU.1 peak center is indicated on the x-axis and the 95% confidence interval is shown (upper histograms). PU.1 reads of cells transfected with reduced PU.1 levels are depicted in comparison (100%, blue; 50%, medium blue; 15%, light blue; middle histograms). Histogram plots showing the coverage of annotated H3K27ac and ATACseq data of PU.1 (ATAC, blue; H3K27ac, green)- vs. PU.1mut (ATAC, grey; H3K27ac, turquoise)-transfected CTV-1 cells across depicted differential bound regions (lower histograms).

### 10.2.5 Appendix V – PU.1 Interactome in diverse hematopoietic cell lines

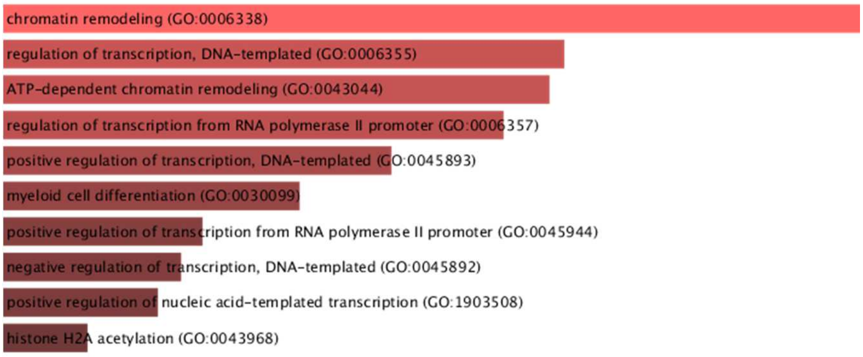


**Figure 10-6 - Gene ontology of enriched proteins in PU.1-BirA-transfected vs. NLS-BirA-transfected THP-1 cells**

Heat map of enriched terms across PU.1-specific input gene list, colored by p-values and generated using the Enrichr suite.

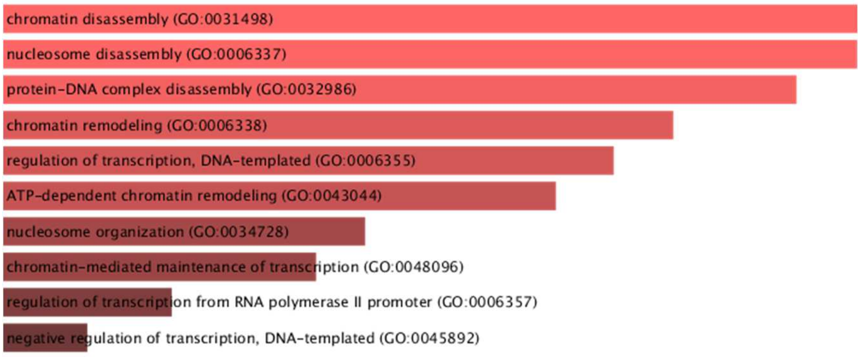
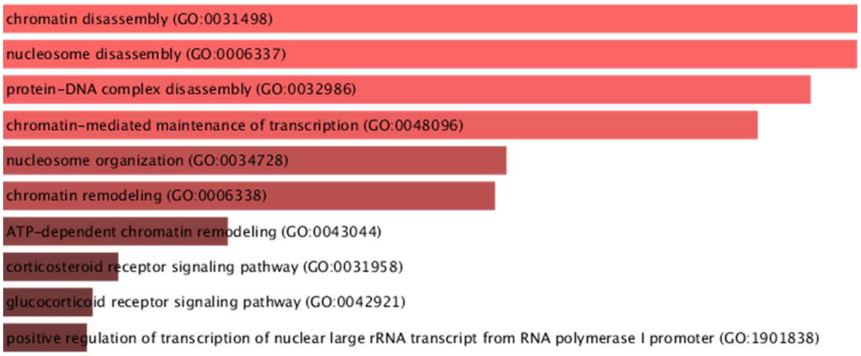


**Figure 10-7 - Gene ontology of enriched proteins in PU.1-BirA-transfected vs. NLS-BirA-transfected CTV-1 cells**

Heat map of enriched terms across PU.1-specific input gene list, colored by p-values and generated using the Enrichr suite.



.

**Figure 10-8 - Gene ontology of enriched proteins in PU.1-BirA-transfected vs. delA-BirA-transfected CTV-1 cells**

Heat map of enriched terms across PU.1-specific input gene list, colored by p-values and generated using the Enrichr suite.

# ACKNOWLEDGEMENT

„Der Langsamste, der sein Ziel nicht aus den Augen verliert, geht immer noch geschwinder als der, der ohne Ziel herumirrt (Gotthold Ephraim Lessing).“

Deshalb möchte ich mich an dieser Stelle bei allen nachstehenden Personen bedanken, ohne deren Hilfe ich mein Ziel wohl so manches Mal aus den Augen verloren hätte:

Mein besonderer Dank hierfür gilt zunächst Herrn Prof. Dr. Michael Rehli, meinem Doktorvater, für die Betreuung dieser spannenden Arbeit, der endlosen Hilfe in bioinformatischen Belangen und der vielfachen Ideengebung, die mir einen unglaublichen Zugang zu dieser Thematik eröffnete. Die zahlreichen Gespräche mit Dir werden mir immer als bereichernder und konstruktiver Austausch in Erinnerung bleiben.

Des Weiteren möchte ich mich herzlich bei Herrn Prof. Dr. Wolfgang Herr, dem Direktor der Klinik und Poliklinik für Innere Medizin III bedanken, ohne dessen Zuspruch die Anfertigung dieser Arbeit nicht möglich gewesen wäre.

Meinem Mentor Prof. Dr. Gernot Längst danke ich herzlich für seine stete Hilfsbereitschaft, den regen Diskussionen zu allen Belangen dieser Arbeit, sowie zur Bereitschaft diese doch sehr „kurz“ geratene Dissertation zu lesen und zu begutachten ;) Auch dem restlichen Prüfungskomitee (Prof. Dr. Herbert Tschochner & Prof. Dr. Christine Ziegler) gilt mein besonderer Dank für die investierte Zeit zur Begutachtung dieser Arbeit und meines Kolloquiums.

Ohne all meine lieben Kollegen aus dem H1 wäre ich wohl so einige Male verzweifelt. Ihr habt mich stets motiviert, es immer wieder geschafft mich aufzubauen, wenn die Dinge gerade nicht so gut liefen, und dieses Labor zu einem unglaublich angenehmen Arbeitsplatz gemacht. Ich danke euch von Herzen…

Lucia, du hast mir mit deiner liebevollen & dennoch direkten Art stets geholfen jede Krise zu überwinden und auf den Boden der Tatsachen zurückzukehren. Die zahlreichen Gespräche mit Dir werde ich nie vergessen und sie werden meinen weiteren Lebensweg immerzu begleiten! Danke!

Johanna, danke, dass du es all die Jahre neben mir ausgehalten hast! Du hattest immer ein offenes Ohr für mich, hast stets all meine Launen ertragen und mich dennoch nie im Stich gelassen! Auch dir danke ich von Herzen!

Claudi, Sandra und Ute tausend Dank für Eure Freundschaft und die zahlreichen schönen Stunden, mit oder ohne Kaffee! Ihr habt meinen Arbeitsalltag sehr bereichert und habt enorm dazu beigetragen, dass ich immer gerne ins Labor gegangen bin!

Auch allen anderen ehemaligen und gegenwärtigen Kollegen der AG Rehli (Claudia, Daggi, Jan, Julia, Karina & Margit), sowie der AG Kreutz, AG Edinger und AG Thomas gebührt mein Dank für die stete Bereicherung meines Laboralltags. Mit Euch gab es immer etwas zu lachen & zu feiern und ich konnte immer auf euch zählen!

Auch all meinen Freunden, die mich stets begleitet, mir Mut zugesprochen haben, oder mich einfach von Allem abgelenkt haben, möchte ich von Herzen danken. Ohne Euch wäre es nicht dasselbe! Inge, Miia & Martin, Mona & Markus, Yvi & Vicky, Ihr seid die Besten!

Meiner Familie und Schwieger-Familie in Spe ;) danke ich von Herzen für ihre Liebe und Unterstützung. Danke Mama & Papa, ohne Euch wäre dieses Studium nie möglich gewesen! Eure nie enden wollende Geduld und Eurer stetes Vertrauen in mich haben mich stets begleitet und mir das Leben umso vieles einfacher gemacht! Lisi, Philipp & Co und Sophia, jeder sollte Geschwister wie euch haben! Besonders Dir, Sophia, danke ich, da du es immer wieder schaffst mich auf andere Gedanken und zum Lachen zu bringen!

Und das Beste kommt natürlich immer zum Schluss:

Malte, du bist der wunderbarste Mensch, den ich kenne. Deine Liebe, deine aufbauenden Worte, deine stete Geduld und immerwährende Hilfe - auch zu allen fachlichen Belangen dieser Arbeit - sind nur einige wenige Dinge, für die ich mich von Herzen bei dir bedanken möchte! Du bist mein bester Freund und ich bin überwältigt, dass du meine Launen stets ertragen hast und mich noch dazu immer unterstützt, aufbaust und auf andere Gedanken bringst! Ohne Dich und unsere Gang hätte ich es nicht so weit gebracht….DANKE, DANKE, DANKE !!!!