Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

presented by

M. Sc. Ivan Berest

born in Luhansk, Ukraine

Oral examination: July 13th, 2020

A computational multiomics method for quantifying activity and regulatory mode of transcription factors and its application in leukemia

Referees:

Prof. Dr. Anne-Claude Gavin

Prof. Dr. Henrik Kaessmann

Acknowledgements

During my road as a PhD student I met a lot of different people. They changed me and now I want to express my gratitude for all their help and support.

When I first entered the office of my PhD supervisor Dr. Judith Zaugg, I almost lost belief in science. I was about finishing my Masters in molecular and cellular biology and trying to switch from wet lab to computational analysis. I think through all my life I will be endlessly grateful for the chance that she gave me at that stage. It is hard to summarize what exactly she gave to me during my PhD. Most importantly she always believes in me and supports me as much as she can. Through the thousands hours of discussions she honed in me ability to look on the data with a scientific approach and always visualise in my head the end result of the analysis. I am especially thankful for engagement in many collaborative projects, through which I had a chance to work with wonderful scientists around the globe.

I want to thank Dr. Christian Arnold. Through his supportive guidance I learned the technical side of bioinformatics. Through all these four years he never didn't answer my question, and believe me, I had plenty of them every day. I could never wish for a better tutor and example. Thank you for all your help, my friend!

I cannot omit with the words of gratitude my TAC members - Prof. Dr. Henrik Kaessmann, Prof. Dr. Anne-Claude Gavin and Dr. Wolfgang Huber. They always supported me during my PhD and guided with their wise advice towards the correct direction. Also, many thanks to Prof. Dr. Benedikt Brors for his willingness to be part of my thesis defence examination committee.

The presented work in this thesis has not been possible without the help of many collaboration partners. First of all, I am very grateful to Dr. Kasper Dindler Rasmussen and Prof. Dr. Kristian Helin for their help and advice in the analysis of the TET2 data. I wanted to say thanks also to Holly Giles, Dr. Peter-Martin Bruch, Dr. Sascha Dietrich and Dr. Wolfgang Huber for their help in the providing external CLL data and statistical support. Especially I want to thank Dr. Bernd Klaus for his statistical advice during different stages of diffTF development.

Warm thanks to the all past and present members of the Zaugg group. This time would never be so much fun without you. Especially I want to thank Armando, Giovanni, Daria, Olga, Manish, Rafail, Irene, Nacho and Mariana for their time, support and guidance. Each of them taught me important lessons and I will never

forget them. I am also thankful to the new generation (Neha, Mathias, Anna, Aryan, Guido, Mikael) of the Zaugg group for the fun times that we spend together. Big thanks to the EMBL International PhD programme and all my PhD fellows from the 2016 batch. I am sure we will keep in touch.

I am especially grateful to my parents and family for their support, love and warm shoulder that gave me strength to go further through all my troubles. They raised me with a specific scientific mindset and never suppressed my curiosity. Last but not least, I want to thank my wife Nadja for her everyday support and understanding. I owe her a lot of nights together, and I am grateful that I have her in my life. Without her, all this would not be possible. Thank you for your love, my sweetie.

Abstract

Recent breakthroughs in sequencing technologies allowed researchers to generate extensive amounts of data characterizing cellular regulation at many levels. Consequently, this boosted our understanding of gene regulatory networks responsible for different biological processes and highlighted the overall importance of transcription factors (TFs). TFs are dynamic mediators that react to both intra- and extracellular changes in order to ultimately transmit signals and execute genetically inherited gene regulatory programs in a time- and location-specific manner. However, it is still challenging to quantify *in vivo* TF specific binding occupancy and dynamics due to the high complexity of the regulatory part of the genome. Modern technologies measuring chromatin changes (e.g., chromatin accessibility, DNA methylation, histone modifications) can now generate testable hypotheses about the effects of TF binding on gene regulation.

In this thesis, I mainly describe the novel computational tool `diffTF`, a multiomics data integration tool for globally assessing differential TF activity and classifying TFs into transcriptional activators and repressors (by integrating chromatin accessibility and gene expression data). We applied it to a recently published ATAC-seq dataset from a cohort of chronic lymphocytic leukemia (CLL) patients and identified dozens of differential active TFs representing two different CLL subtypes that are inherently linked to tumour progression. In addition, we integrated gene expression data from corresponding RNA-seq and were able to globally predict an activating or repressive role for 40% of the expressed TFs. We validated the approach on an independent CLL dataset and showed that the majority of TFs does not change their mode of action upon genetic or environmental perturbations. Finally, we extensively tested and benchmarked `diffTF` to validate its technical robustness.

We also applied `diffTF` to a multiomics dataset from the mouse hematopoietic differentiation system and targeted potential TFs that are disturbed upon epigenetic dysregulation driven by a Tet methylcytosine dioxygenase 2 (TET2) knockout in acute myeloid leukemia (AML). TET2 plays an essential role in the cellular DNA methylation balance and is known to be frequently mutated in leukemia. We used the first high-quality TET2 binding map to identify TF families that can facilitate TET2 binding in the genome.

In summary, we developed a novel hypothesis-generation computational tool that can, in a data-driven way, identify key regulators of cellular biological processes based on chromatin and expression data.

Zusammenfassung

Die jüngsten Durchbrüche bei den Sequenzierungstechnologien ermöglichten es den Forschern, umfangreiche Datenmengen zu hinterlegen, die die zelluläre Regulation auf vielen Ebenen charakterisieren. Dies verbesserte folglich unser Verständnis der Genregulationsnetzwerke, die für verschiedene biologische Prozesse verantwortlich sind, und der allgemeinen Bedeutung von Transkriptionsfaktoren (TFs). TFs sind dynamische Mediatoren, die sowohl auf intra- als auch auf extrazelluläre Veränderungen reagieren, um letztendlich Signale zu übertragen und genetisch vererbte Genregulationsprogramme zeit- und ortsspezifisch auszuführen. Aufgrund der hohen Komplexität des regulatorischen Genoms ist es jedoch immer noch schwierig, die TF-spezifische Bindungsbelegung und -dynamik *in vivo* zu quantifizieren. Moderne Technologien, die Chromatinveränderungen messen (z. B. Zugänglichkeit von Chromatin, DNA-Methylierung, Histonmodifikationen), können nun überprüfbare Hypothesen über die Auswirkungen der TF-Bindung auf die Genregulation generieren.

In dieser Arbeit beschreiben wir hauptsächlich das neue Tool `diffTF`, ein Multiomics-Datenintegrationswerkzeug zur globalen Bewertung der differentiellen TF-Aktivität und zur Klassifizierung von TFs in Transkriptionsaktivatoren und -repressoren (durch Integration von Chromatin-Zugänglichkeits- und Genexpressionsdaten). Wir haben es auf einen kürzlich veröffentlichten ATAC-seq-Datensatz aus einer Kohorte von Patienten mit chronischer lymphatischer Leukämie (CLL) angewendet und Dutzende von differentiell aktiven TFs identifiziert, die verschiedene CLL-Subtypen darstellen, die inhärent mit der Tumorprogression zusammenhängen. Zusätzlich haben wir Genexpressionsdaten aus entsprechenden RNA-Sequenzen integriert und konnten die globale aktivierende oder repressive Rolle für 40% der exprimierten TFs vorhersagen. Wir haben den Ansatz anhand eines unabhängigen CLL-Datensatzes validiert und gezeigt, dass die Mehrheit der TFs ihre Wirkungsweise bei genetischen oder Umweltstörungen nicht ändert. Schließlich haben wir `diffTF` ausgiebig getestet und evaluiert, um seine technische Robustheit zu überprüfen.

Wir haben `diffTF` auch auf einen Multiomics-Datensatz aus dem hämatopoetischen Differenzierungssystem der Maus angewendet und potenzielle TFs identifiziert, die durch eine epigenetische Dysregulation gestört werden, ausgelöst durch

einen TET2-Knockout bei akuter myeloischer Leukämie (AML). TET2 spielt eine wesentliche Rolle im zellulären DNA-Methylierungsgleichgewicht und ist bekanntermaßen bei Leukämie stark mutiert. Wir haben die erste hochwertige TET2-Bindungskarte verwendet, um TF-Familien zu identifizieren, die die TET2-Bindung im Genom erleichtern können.

Zusammenfassend haben wir ein neuartiges Tool zur Erstellung von Hypothesen entwickelt, das auf datengesteuerte Art und Weise Schlüsselregulatoren zellbiologischer Prozesse aus Chromatin- und Expressionsdaten identifiziert.

Table of Contents

List of Figures

List of Abbreviations

CLL	Chronic lymphocytic leukemia
U-CLL	Unmutated chronic lymphocytic leukemia
M-CLL	Mutated chronic lymphocytic leukemia
TF	Transcription factor
TFBS	Transcription factor binding site
TSS	Transcription start site
DBD	DNA binding domain
GO	Gene ontology
AUC	Area under the curve
PWM	Position weight matrix
HSC	Hematopoietic stem cells
GMP	Granulocyte-monocyte progenitors
MPP	Multipotent hematopoietic progenitors
ES	Embryonic stem cells
FDR	False discovery rate
NGS	Next generation sequencing
IDR	Irreproducible discovery rate
DNAse-seq	DNAse I hypersensitive sites sequencing
MNase-seq	Micrococcal nuclease digestion followed by sequencing
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements sequencing
ChIP-seq	Chromatin immunoprecipitation DNA-Sequencing

Chapter 1

Introduction

In the past 50 years our understanding of molecular events happening in the cell and their underlying logic has improved significantly. Science moved from the descriptive state of observing and characterizing biological molecules to the analyzing and predictive state of the molecular systems. With the discovery of the central molecular biology dogma, the role of genes are characterized as a blueprint for the construction of all biological systems. Each eukaryotic cell has a nucleus with DNA inside, compacted in a specific manner. In the tightly controlled process of gene regulation genes are getting expressed to RNA and then transcribed to proteins, therefore transferring information stored in the nucleus to maintain and construct the cell. Each cell, though, is not only a product of genetic information from ancestor cells, but it has an amazing ability to react and adapt to external/internal stimuli. Using molecular receptors and signalling pathways a cell is able to modify and correct the *maintaining plan*, written in DNA, e.g. via the gene regulation mechanisms. The definition of the cell types changed from the description of the cellular phenotypes (size, shape, location of the cells) to the common gene regulation programs and their ability to evolve together (Arendt et al., 2016).

It is obvious, that at this stage of molecular biology development, we are just starting to understand key mechanisms occurring in the cell and taking part in the gene regulation. With the rise of the high-throughput methods (NGS sequencing, imaging, genetic screens) we are gathering more and more information about the characteristics and states of the molecular phenotypes in the cell. However, the major bottleneck for nowadays biology and future generations is to develop analysis strategies to integrate all these massive amounts of data and provide testable hypotheses of cell gene regulation underlying principles in different conditions. Such information is essential in order to provide solutions to the diseases on the multiple levels of biological systems organization.

1.1 Gene regulation in the 3D genome

All the human cell *blueprint* genetic information is encoded within the ~3 billions nucleotides that are stored in the cell nucleus with a diameter of ~10 μm on average. This is an extremely important topological problem solved intracellularly by using additional nucleosome organizational proteins (e.g. histones). At the first level of compactization ~150 bp of DNA are wrapped circularly around histone octamer, thus reducing the needed space 5 times. After this nucleosomes can fold in loops and fibers thus decreasing the occupied 3D space even more. In the past 10 years more and more understanding of the DNA structural folding in the nucleus has been obtained, with the development of the various sequencing and imaging methods/analyses, such as fluorescent *in situ* hybridization (FISH) and chromatin capture (3C, 4C, Hi-C) methods (Kempfer and Pombo, 2019). On top of it, advances in the electron and superresolution microscopy (Boettiger et al., 2016; Ou et al., 2017) provide additional information about chromatin structure in the nucleus. Latest research suggests that chromatin is a highly dynamic, but strongly controlled intracellular phenotype. Genes located in the regions with more compacted chromatin have lower expression levels, in comparison to genes in the open chromatin, so called "A" and "B" compartments (Lieberman-Aiden et al., 2009).

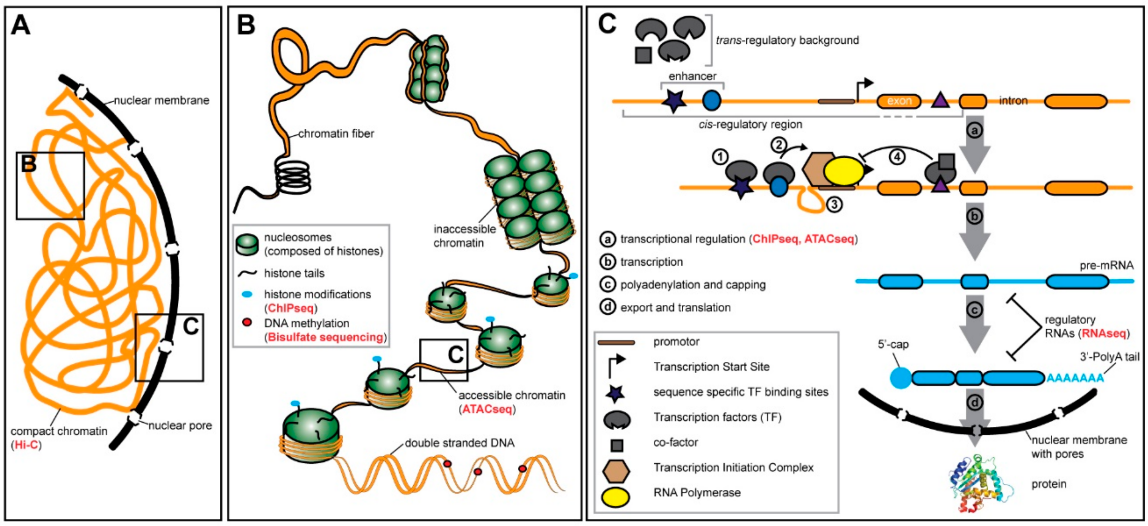

Figure 1.1: **Chromatin affects gene regulation.** (A) Scheme of chromatin organization in the nucleus. (B) Compression of DNA in the nucleus with histones. Scheme of opened and closed chromatin regulated post-translational histone modifications. NGS methods for measuring different information about chromatin dynamics affecting gene expression (ATAC-seq, ChIP-seq, Bisulfite sequencing) are shown. (C) Scheme of the gene expression (a–d). TFs specific DNA binding (1). Activation of the transcriptional machinery (2) through DNA looping (3). Enhancer regulation of transcription (4). Adapted from (Buchberger et al., 2019) under Creative Commons License 4.0.

A remarkable observation of the Human Genome Project was that only 2% of the human genome are occupied by the protein-coding genes. The remaining part is used to regulate the expression of the protein-coding genes in various ways. Compared to the prokaryotic promoter-oriented gene expression scheme, eukaryotic regulation is much more complex, involving a combined regulation of enhancers and promoters, summarized in Figure 1.1. Large international consortia, such as ENCODE and Roadmap Epigenomics (Kundaje et al., 2015), generated deep databases of the specific enhancer regulation of different cells using different epigenetic phenotypes (e.g. chromatin states, chromatin accessibility, DNA methylation, TF binding).

One of the possible explanations of the eukaryotic gene regulation can be characterized with restricted spatial organization of the genome (Li et al., 2018). Chromatin inside the nucleus can be divided by the topological associated domains (TADs), a structure, which explains the enhancer-promoter gene regulation. Thisstructure, on a big scale chromosomal territories, is conserved through evolution (Fudenberg and Pollard, 2019) and highly controlled. All the dynamic changes of the gene regulation, responsible for the fast adaptations and reaction to the intra- and extracellular signals, are occuring with the changing DNA loop structures inside the TADs. The previously proposed loop extrusion (Sanborn et al., 2015) model explains how CTCF together with cohesin, through sliding on DNA, can regulate the 3D structure of the locus specific DNA, and connect enhancer and promoter regions.

Another model attempting to explain chromatin structure is phase separation (Hnisz et al., 2017) based on the biophysical properties of the molecules inside the nucleus. Due to the high amount of different proteins associated with gene transcription (TFs, RNA polymerase, transcriptional machinery) regulatory regions in the genome have gel-like consistency different from the heterochromatin, and therefore forced physically to interact. The same effect is enhanced with increased local concentration of TFs (Cortini and Filion, 2018) and is predicted to occur in super-enhancers. Likely, the truth about principles of chromatin organization lies somewhere in the middle of these two models, taking into account sequence specific features of DNA itself and biophysical properties of molecules around it.

In this thesis I will discuss and use chromatin accessibility data obtained from NGS sequencing. Notably, information about chromatin accessibility and structure can be obtained using imaging techniques. At the moment, these two fields of chromatin biology are developing separately, but future integration of both approaches to define precise 3D genome in the nucleus is essential.

1.2 Chromatin accessibility assays

Recent advances in sequencing technologies together with decreased cost of it, allowed development of high-throughput DNA accessibility methods (Klein and Hainer, 2019). In this section we are going to mainly discuss methods developed in the past ten years to measure bulk chromatin accessibility, such as DNAse-seq, FAIRE-seq, MNase-seq and ATAC-seq. However, more and more techniques are appearing that measure chromatin accessibility as readout on the single-cell level (e.g. scATAC-seq, scDNAse-seq). This is particularly interesting, as we will be able to measure dynamic changes of chromatin for each cell and much more precise integrate this information with the unique cellular external and internal molecular phenotypes to understand the basic gene regulation principles.

Overall, all chromatin accessibility methods are based on the ability of enzymes to access and modify or digest chromatin. A schematic representation of the most popular methods is shown in the Figure 1.2. DNAse-seq uses DNAse I endonuclease, which cuts unprotected DNA, and therefore allows further identification of it after PCR amplification and sequencing (Song and Crawford, 2010). DNA fragments of a certain size (from 50 bp to 160 bp) are used to construct a template for the library construction. This method was the first one to define chromatin accessibility and is widely used till now, especially in the ENCODE consortium. FAIRE-seq, on the contrary, doesn't use enzymatic activity to define opened regions, but rather use formaldehyde crosslinking to identify all DNA-protein connections (Gaulton et al., 2010). DNA that was not crosslinked is sonicated, amplified and then sequenced. As this method crosslinks all the DNA-protein interactions it is much harder to obtain high quality chromatin accessibility signal using it. MNase-seq combines both of the described above techniques, while first crosslinking DNA with proteins and then applying micrococcal nuclease to digest free DNA (Schones et al., 2008). Covered DNA then follows removal of the proteins and further sequencing. Importantly, this method was used to identify precise nucleosome positioning. It is quite sensitive to the MNase digestion kinetics and requires specific titration beforehand.

ATAC-seq is using Tn5 transposase in order to find accessible regions in chromatin (Buenrostro et al., 2013). It cuts accessible DNA and insert specific primers, which can be used for extracting nucleosome-free regions of DNA afterwards and amplification of the signal. In comparison to the other methods, ATAC-seq does not require a lot of time (whole experiment can be done in a few hours) and has a higher signal-to-noise ratio. In recent years this method is used widely and has become the

golden standard to define chromatin accessibility. It is important to note that the majority of these methods, based on the enzymes, have different biases based on the DNA sequence. For example DNase I has a specific bias depending on the width of the minor groove of DNA. On top of it, there are known CG biases of sequencing that should be taken into account.

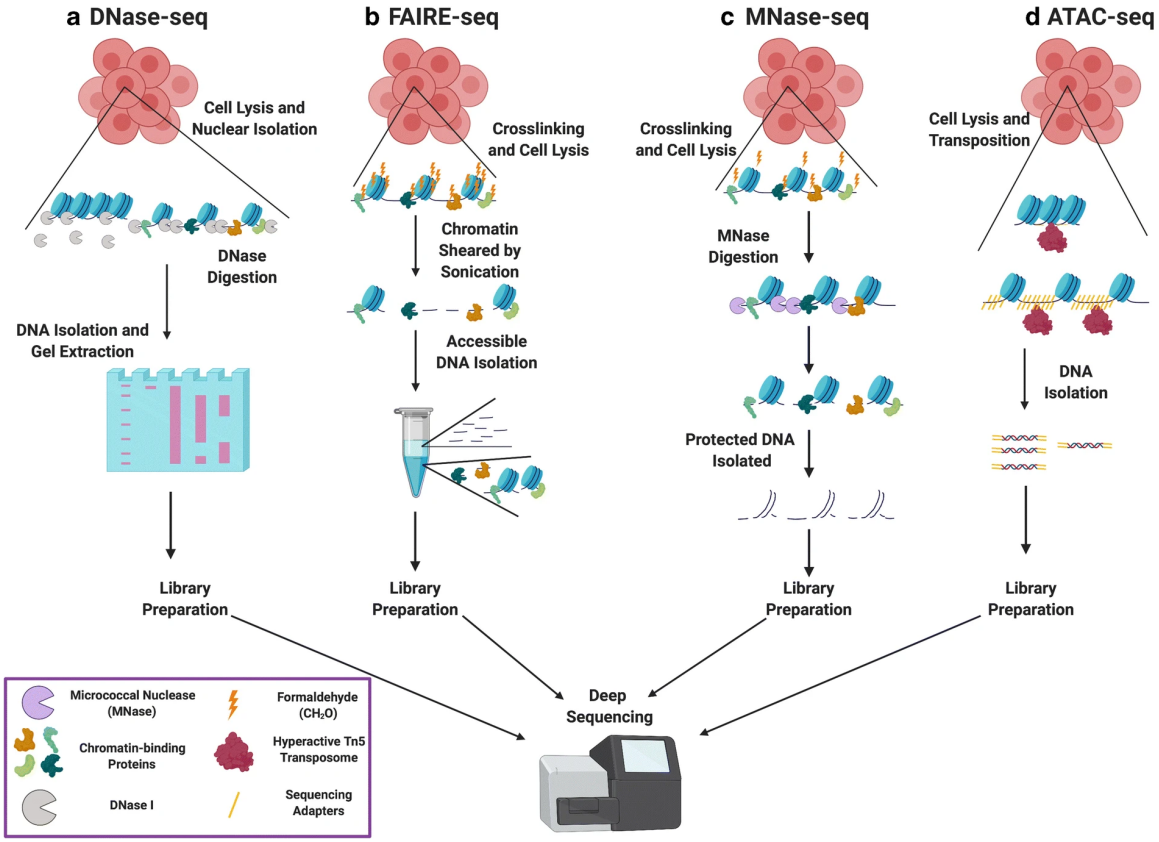

Figure 1.2: **Chromatin accessibility methods.** Schematic representation of DNAse-seq (A), FAIRE-seq (B), MNase-seq (C) and ATAC-seq (D) workflow. Adapted from (Klein and Hainer, 2019).

For each of the mentioned methods specific pre-processing analysis is needed to infer open chromatin regions from the raw sequencing data. However processed data defining chromatin accessibility can be used together in downstream analyses (TF footprinting, differential accessibility analysis, nucleosome occupancy). More importantly, these techniques are not identifying specific binding of different proteins (histones, TFs) and need to be combined with assays that specifically map binding to DNA of different regulatory proteins (e.g. ChIP-seq for TF binding) to understand gene regulation mechanisms.

1.3 Transcription factors

Transcription factors is a unique class of proteins that can directly bind DNA sequence and regulate gene expression. They are an essential part of the cell signalling and often are master regulators of key cellular processes, such as cell differentiation, development, cell cycle and metabolism. Dysregulation of TF functioning often occurs in diseases and cancer. A typical TF consists of a DNA-binding domain (DBD), which facilitate DNA binding, and effector domains, that are responsible for the fine-tuning of the TF activity through the binding of other molecules/cofactors (Figure 1.3A). Places where TFs can bind (TFBS) are essential to know, as the identification of them is the first step of possible linkage of TF activity with the gene expression (Wasserman and Sandelin, 2004). Such binding specificities for each TF are recapitulated in the term TF motif, a predictive model of TF binding based on the DNA sequence. Experimentally, TF motifs were determined, both *in vitro* with protein binding microarrays (Berger et al., 2006), SELEX methods (Jolma et al., 2010) or earlier with DNA footprinting assays (Galas and Schmitz, 1978), and *in vivo* with chromatin immunoprecipitation based methods (Johnson et al., 2007) and DamID-seq (Wu et al., 2016). Despite the fact that there are a lot of methods uncovering locations of TF binding in the genome, it is still a huge limitation towards analyzing complex gene regulation networks. TF binding is highly dynamic, and ChIP based methods, which are known to best characterize TF binding, are not able to measure this dynamics due to crosslinking. Also if it is not possible to get antibodies for a specific TF that makes ChIP based assays for such cases not appropriate. On the other hand it is possible to model *de novo* TF binding sites in a genome, using as initial set TFBS derived from experimental studies. One of the most widespread representations of TF motifs is position weight matrix (PWM), that is basically a base pair oriented relative representation of TF binding specificity in a specific genome (Stormo and Zhao, 2010). There are more ways of TF motif representation, reviewed in (Lai et al., 2019), which take into account similarities between TFs, DNA shape and sequence background. Overall it seems that the combination of the computational predictions of TF binding and experimental validations of such binding is a key for future dissection of the precise TF binding and impact on gene regulation.

In recent years, a lot of progress of TF annotation has been done, with the rise of many TF binding models databases (JASPAR (Mathelier et al., 2016), HOCOMOCO (Kulakovskiy et al., 2015), CisBP (Weirauch et al., 2014)). The whole repertoire

of the human transcription factors consists of 1639 TFs, summarized in (Lambert et al., 2018). Most of the TFs contain one type of DBD and can be grouped based on that into 35 TF families (Figure 1.3B). Importantly, despite DBD domains in the TFs are highly consevative through evolution (Lambert et al., 2019), there are a lot of variation of the other domains of the TF that are regulated by the cofactors, which is much more harder to identify *in vivo*. Most abundant TF families are C2H2 zinc finger factors (~45% of all TFs), homeodomain containing TFs (~11%), bHLH, bZIP, HMG, Forkhead and nuclear hormone receptors. For the zinc fingers it is also possible to further subclassify TFs based on the effector domains (KRAB, BTB, SCAN). Curation of the transcription factors is a very challenging task. A lot of known TFs lack crystal structures, there is high similarity between the motifs and it is hard to dissect the functional effect of individual TFs on the biological regulation.

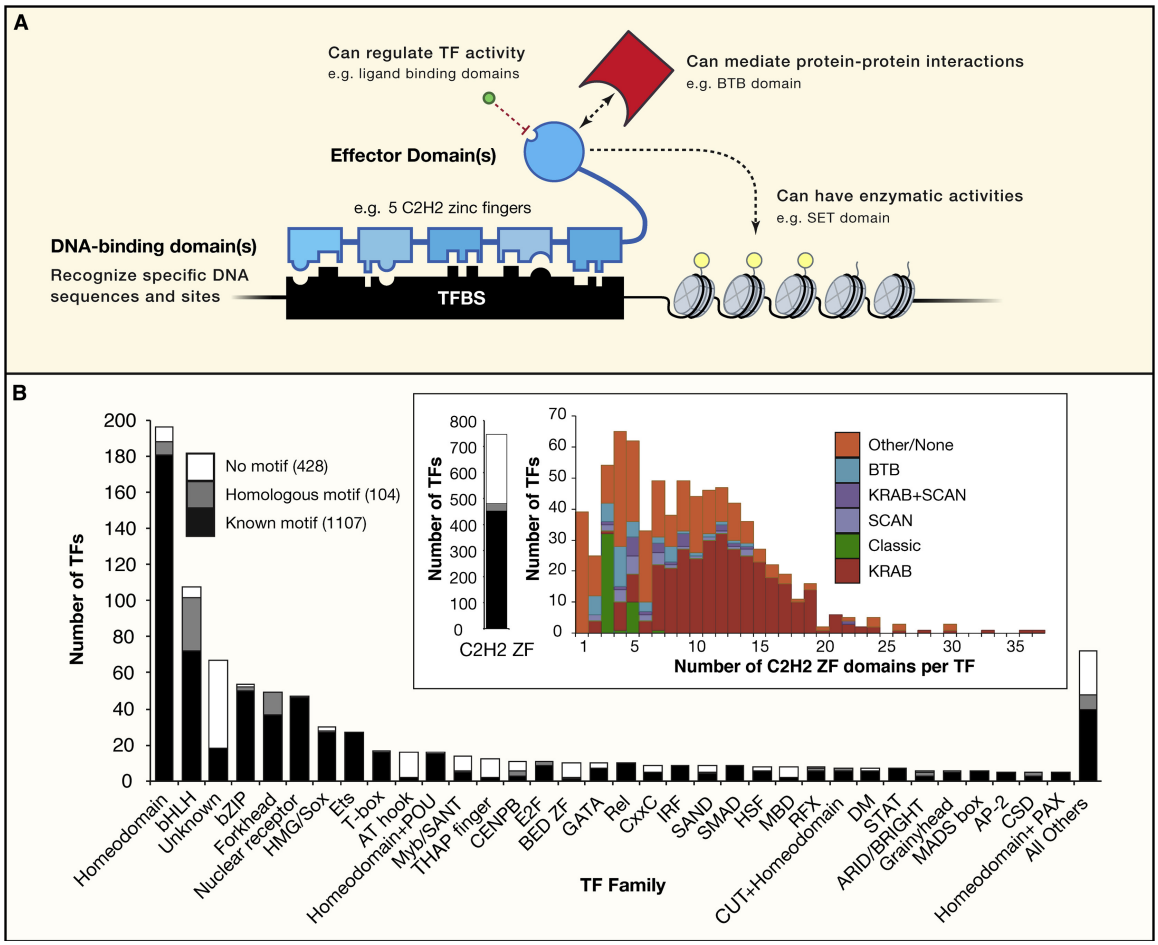

Figure 1.3: **Classification of TFs based on DBDs.** (A) Schematic of a prototypical TF. (B) Distribution of TFs and motif status for each DBD family. Zoomed area describe further division of the C2H2 zinc fingers family into effector domains (KRAB, SCAN, or BTB domains); "Classic" group summarize related and highly conserved SP, KLF, EGR, GLI GLIS, ZIC, and WT proteins. Adapted from (Lambert et al., 2018).

Via binding to DNA, TFs directly or indirectly control gene expression. Depending on the functional context they can activate or repress expression. Notably, they don't necessarily act on their own at any given TFBS. Large amounts of cofactors, that can either enhance or reduce the activity of a TF, as well as local sequence content affect the function of TF. In eukaryotic cells vast amounts of chromatin modifiers are affecting the region-specific TF functionality as well. Taken together, it seems that TFs are maintaining the balance of regulation, which is, despite being tightly controlled, highly variable and gene specific through different regulatory mechanisms. Disruption of TF regulatory networks mainly leads to apoptosis or abnormal cellular functioning. They are an essential part of the chromatin regulation and key mediators of all the changes occuring in the cell.

1.4 Measuring TF activity

In the following chapters we will refer to TF activity as a measure how TFs can regulate chromatin and target genes expression. Interestingly, with a different logic in mind, a massively parallel protein activity assay for TFs, based on the electrophoretic separation of TF-bound DNA sequences from a complex DNA library and followed by mass-spectrometry, was performed recently (Wei et al., 2018) and found a high correspondence with the chromatin accessibility data.

Before the rise of the chromatin accessibility methods, TF activity was often inferred from gene regulatory networks. After defining a set of target genes for each TF, based on the proximity of binding sites, through manual curation or semantic search, TF activity was defined from a linear model taking into account expression of TF and target genes. A number of such approaches are summarized in (Garcia-Alonso et al., 2019). They have been proven to correctly predict the magnitude of TF activity changes in different biological systems and generate testable hypothesis, however not always showing consistent accurate estimates of the regulatory effect of TFs on the specific gene. TFs are regulating gene expression through an effect on chromatin (i.e. binding to the DNA), thus, gene expression is a less direct readout than chromatin accessibility, since it is the result of activities of multiple enhancers, promoters and posttranscriptional regulatory steps. Besides this TFs can also co-operatively regulate expression of the genes, making the inferred predictions from only gene expression data even more noisy. Using chromatin accessibility, as more direct readout of the gene expression and regulation, which allows us to estimate more accurate TF activity from the predicted TFBS. In the following paragraphs I

will briefly describe major computational approaches for estimating TF activity from the bulk and single cell chromatin accessibility samples.

One of the recent tools for the bulk chromatin accessibility samples, BagFoot, calculates TF footprints for the predicted TFBS, and then compares obtained signal between two conditions of interest (Baek et al., 2017). By default, a differential TF footprint is defined as the sum of the genome-wide differential footprint depth and flanking accessibility. Footprint depth serves as a normalisation factor of the TFBS accessibility compared to the genome-wide baseline. Differences in the flanking regions of a footprint are hypothesized to distinguish between more TFs locating in more active or repressive chromatin. BagFoot has an interesting approach purely based on TF footprinting, however it is not clear if this tool can be used for all predicted TFs. First of all, not all TFs have a strong footprint, due to the high dynamics of association/disassociation. Also, it is known that for proper footprint you need to have deeply sequenced samples (Calviello et al., 2019), thus making this approach even more specialized on the input data. Another interesting TF activity defining computational application (Azofeifa et al., 2018), demonstrated also on the bulk samples, uses the assumption that active enhancers should be associated with the presence of eRNAs. Using computed eRNA profiles they calculate the enrichment of TF motif occurrences around eRNA sites and define a differential motif displacement score between conditions. It is possible to apply such an approach to ATAC-seq data (Tripodi et al., 2018), however no proper benchmarking on the effect of eRNA peak calling and comparison to other enrichment tools were shown.

With the technological advancement of chromatin accessibility methods for single cell data, it is possible to estimate TF activity on the single cell level. One of the first tools developed to define such measures is chromVAR (Schep et al., 2017). In this approach TF activities are defined as chromatin accessibility deviations for each TF in each sample compared to other cells. Using predicted TFBS that overlap with accessibility peaks it compares observed and expected fragment counts for each of TFBS. Obtained raw deviation score is then compared to the background distribution of deviations, that takes to account CG bias for each TFBS. The advantage of such a method is that it computes one TF activity measure for each cell/sample, however it doesn't have encoded statistical methods to compare samples from different groups between each other. One can use mean of deviations z-scores to summarize the signal between samples of each group and compare such values with another group. Development of computational strategies defining TF activity, especially for single cell data, are extremely important for the analysis of the cell regulatory landscape.

1.5 Chromatin regulation changes in leukemia

Hematopoiesis is one of the most controlled biological processes. The formation of the spectrum of different blood cells from one source of stem cells is coordinated through specific gene regulation programs. However, dysruption of such regulation and evasion from normal hematopoiesis cause leukemia. Leukemia oncogenesis results from both genetic and epigenetic factors. All leukemias can be divided into four main classes: acute lymphoblastic leukemia (ALL), chronic lymphocytic leukemia (CLL), acute myeloid leukemia (AML) and chronic myeloid leukemia (CML). ALL is the most common type of cancer in children, whereas adults often are diagnosed with the CLL and AML types. As genetic factors are inherited and hardly can be removed, the main focus of possible high-level treatments of leukemia is to understand what is causing the disruption of the gene regulatory mechanisms in the cells, on which stage of the differentiation this is happening, and ultimately prevent or revert this. Such a goal sounds almost impossible taking into account how complex and dynamic the gene regulatory mechanisms are. However, one important cellular molecular phenotype that can be measured nowadays and used for further dissection of gene regulation is chromatin accessibility.

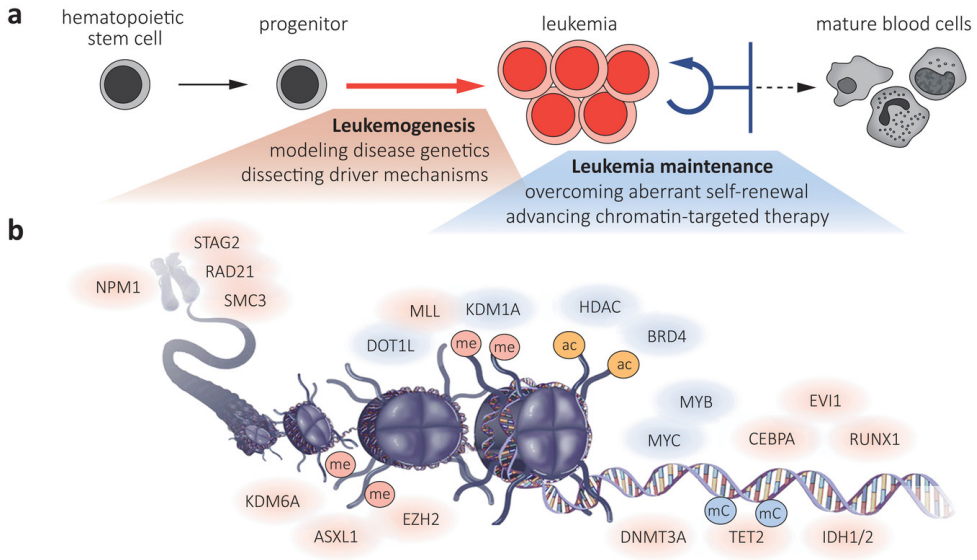

Figure 1.4: **Chromatin regulators in leukemogenesis and leukemia maintenance.** (a) Schematic of leukemia initiation and aberrant self-renewal in the context of hematopoiesis. (b) Chromatin regulators and transcription factors recurrently mutated in leukemia (red) and critical requirements of self-renewal and cell survival (blue). Illustration is taken from [`https://www.imp.ac.at/groups/johannes-zuber/`] and modified from Saygin and Carraway (Journal of Hematology & Oncology, 2017) under Creative Commons License 4.0.

Genomic instability is one of the main characteristics of cancer (Hanahan and Weinberg, 2011), which leads to the abnormal chromatin structure in the nucleus, and it is very common for leukemia. Due to such genomic aberrations normal enhancer-promoter interactions are disrupted. On top of that, leukemia often has specific mutational signatures, especially affecting histones or chromatin modifying enzymes (Boileau et al., 2019). Dysregulation of the DNA methylation is regarded as key event for the development of leukemia development, with increased mutation rate and abnormal functioning of DNMT and TET methyltransferases (Yang et al., 2019). Altogether, it seems that cancer affects gene regulation on multiple levels (chromatin, histone modifications, DNA methylation), and due to high interconnectivity of these molecular phenotypes it is challenging to find the causal dysfunction among them. Moreover, leukemogenesis is a very rapid process, difficult to catch and analyze, with most of the samples coming already on the stage of the leukemia maintenance (Figure 1.4).

Transcription factors play a central role in cancer as well. All the genetic predisposition together with environmental changes are affecting specific TF binding programs. Together with chromatin modifying enzymes TFs are balancing the regulation of normal hematopoiesis, by regulating histone modification codes, DNA methylation and chromatin interactions. Specific TFs are known to be mutated in cancer or have disrupted activity due to genomic rearrangements, such as MYC/MYB (Delgado and Leon, 2010), CEBP family of TFs, RUNX family, NFAT TFs (Beekman et al., 2018). Knowledge of the gene regulation changes occurring in the cancer cells can increase our understanding of basic biological pathways and interactions between different layers of epigenetics. More complex and targeted leukemia treatments directed on the regulators of gene expression can be designed using such knowledge.

We propose that chromatin accessibility is an ultimate measure of chromatin changes occuring in leukemias, which can be used to predict in parallel TF activity of multiple TFs. Recently, a lot of research was done to establish epigenomic landscape of the different leukemic cells (Beekman et al., 2018; Rendeiro et al., 2016), and this data should be used for the further investigation of the TF regulatory changes in leukemia.

Chapter 2

Characterizing differential TF activity with `diffTF`

In this chapter I will describe the methodology to define transcription factor (TF) activity that we developed during my PhD and that was applied to the different data in other chapters. The main ideas about this method were conceived by me under the supervision of Dr. Judith Zaugg and Dr. Christian Arnold. Most of the computational analyses of this chapter were carried by me, with some of the analyses being made by Dr. Christian Arnold, Dr. Armando Reyes-Palomares and Giovanni Palla. Additionally, I received statistical support on different stages of this project from Dr. Bernd Klaus and Dr. Wolfgang Huber. The text in this chapter has been originally written by myself and was taken and adapted from:

***Ivan Berest**†, Christian Arnold†, Armando Reyes-Palomares, Giovanni Palla, Kasper Dindler Rasmussen, Holly Giles, Peter-Martin Bruch, Wolfgang Huber, Sascha Dietrich, Kristian Helin & Judith B. Zaugg (2019) Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF. Cell reports, 29(10), 3147–3159.e12. doi: 10. 1016/ j. celrep. 2019. 10. 106*

2.1 Introduction

As discussed in the first chapter, TFs are playing a crucial role for coordinating and regulating various biological processes and pathways in the cell. It was shown before that disruptions of TF motifs in the TFBS can partly explain variation in H3K27ac histone mark signal across individuals (Grubert et al., 2015). Pioneer transcription factors can bind directly to the nucleosomes and affect the chromatin accessibility, histone modifications and methylation landscape of the nearby region (Liu et al., 2015; Mayran and Drouin, 2018). Here, we revert this statement, and use summarized changes of chromatin accessibility or histone modification marks in the vicinity of TFBSs of a certain TF as a functional TF-specific readout of chromatin changes in the cell, which we called TF activity. Such TF activity is rather cell type specific and determined by the local chromatin microenvironment, potential binding partners and cofactors, as well as local concentration of TF (Kribelbauer et al., 2019; Whyte et al., 2013).

Despite the vast regulatory mechanisms controlled by TFs, they are usually low abundant proteins in the cell, making it difficult to measure in high throughput their abundance and activity with proteomics or biochemical assays (Kim et al., 2007; Komatsu et al., 2010; Liu et al., 2011). On the other hand, chromatin immunoprecipitation DNA sequencing (ChIP-seq), which provides information about condition-specific TFBS for one TF at a time, does not quantify changes in TF activity without additional spike-in normalization method (Bonhoure et al., 2014). On top of this complexity, depending on the epigenetic background, the mode of action of many TFs can vary from activating to repressing transcription of genes (Han et al., 2018). Therefore, understanding the global TF mode of action will help to unbiasedly interpret the effects of a changes in TF expression. Altogether, a unified method for defining differential TF activity between conditions or cell types and assigning global TF mode of action in a high-throughput manner is currently missing in the field.

To overcome these limitations, we developed `diffTF` [`https://git.embl.de/grp-zaugg/diffTF`)], a computational Snakemake based pipeline to define global differential TF activities and to classify TFs into activators or repressors based on the integration of chromatin accessibility (ATAC-seq,DNAse-seq) or histone mark ChIP-seq data with gene expression (RNA-seq) data. In the following chapter we will describe in detail the functionality behind this tool and benchmark it against advanced motif analysis methods.

2.2 Methods

2.2.1 Data sources

To benchmark and test our designed workflow we used previously published multiomics dataset from CLL patients (Rendeiro et al., 2016), downloaded from the European Genome-phenome archive with the accession number EGAD00001002110. From this dataset we used for the further analyses 52 ATAC-seq samples (25 of which were classified as unmutated CLL and 27 as mutated CLL) and 8 RNA-seq samples (4 unmutated CLL and 4 mutated CLL samples). Complete table of the used samples with assigned metadata is shown in the Appendix A.

Apart from the sequencing data, we also used several TF binding profiles databases, such as *JASPAR* and *HOCOMOCO*. We obtained *JASPAR 2016* (Mathelier et al., 2016) core position frequency matrices through the *JASPAR* RESTful API [`http://jaspar.genereg.net/api`]. As main set set of predictive PWMs we used data from *HOCOMOCO v10* (Kulakovskiy et al., 2015) database [`http://hocomoco10.autosome.ru`] containing 640 human and 423 mouse TF binding models. We also downloaded ChIP-seq data for *hg19* genome annotation from *ReMAP 2015* (Griffon et al., 2015) [`http://tagc.univ-mrs.fr/remap/index.php`)].

2.2.2 ATAC-seq processing pipeline

As it was first described in the original ATAC-seq paper (Buenrostro et al., 2013) raw ATAC-seq data is required to undergo several processing and quality filtering steps before it can be used for the downstream analyses. We established an in-house Snakemake ATAC-seq processing pipeline that performs several quality controls, adaptor trimming, alignment, base quality recalibration, removal of mitochondrial reads and indels and removal of CG bias, if needed, see workflow in Figure 2.1.

As the first step we perform sequence quality checks on the raw *fastq* data using `FastQC` (Andrews, 2010), followed by the removal of adaptor sequences from the Nextera Transposase with `Trimmomatic` (Bolger et al., 2014) [internal parameters: ILLUMINACLIP:NexteraPE-PE.fa:1:30:4:1:true TRAILING:3 MINLEN:10]. Further we continue with quality checks with `FastQC` (Andrews, 2010) after adapter trimming and alignment to the reference hg19 human genome using `Bowtie2` (Langmead and Salzberg, 2012) algorithm [internal parameters: -X 2000 –very-sensitive]. We also perform base quality recalibration using `GATK` suite (McKenna et al., 2010) with the known variants for the hg19 genome.

As the next step, the pipeline executes several filtering steps: removing mitochondrial reads and indels, removing reads with a low mapping quality (< 10), marking and removing PCR duplicates using Picard tools [http://broadinstitute.github.io/picard/], adding 4bp on the forward and 5bp on the reverse strand to the reads to deal with the transposase cleavage and also removing reads from non-assembled contigs or alternative haplotypes. We also suggest to perform CG bias detection and removal using deepTools (Ramírez et al., 2014) and Benjamini's method (Benjamini and Speed, 2012), as DNA polymerases used for ATAC-seq library generation have CG bias that emerge in synthetically higher read counts for the regions with high CG frequency.

Finally, the downstream analysis of this pipeline ends with peak calling using MACS2 (Zhang et al., 2008) [internal parameters: -q 0.01 –slocal 10000 –nomodel –nolambda] and removal of reads in the blacklisted region of the hg19 human reference genome (Amemiya et al., 2019). Another function of this pipeline is to generate visual and qualitative summary statistics, for example coverage plots, transcription start site enrichment, fragment length distribution plots, that allow to quickly judge the quality and complexity of the performed ATAC-seq experiment and choose which potential further analysis can be performed.

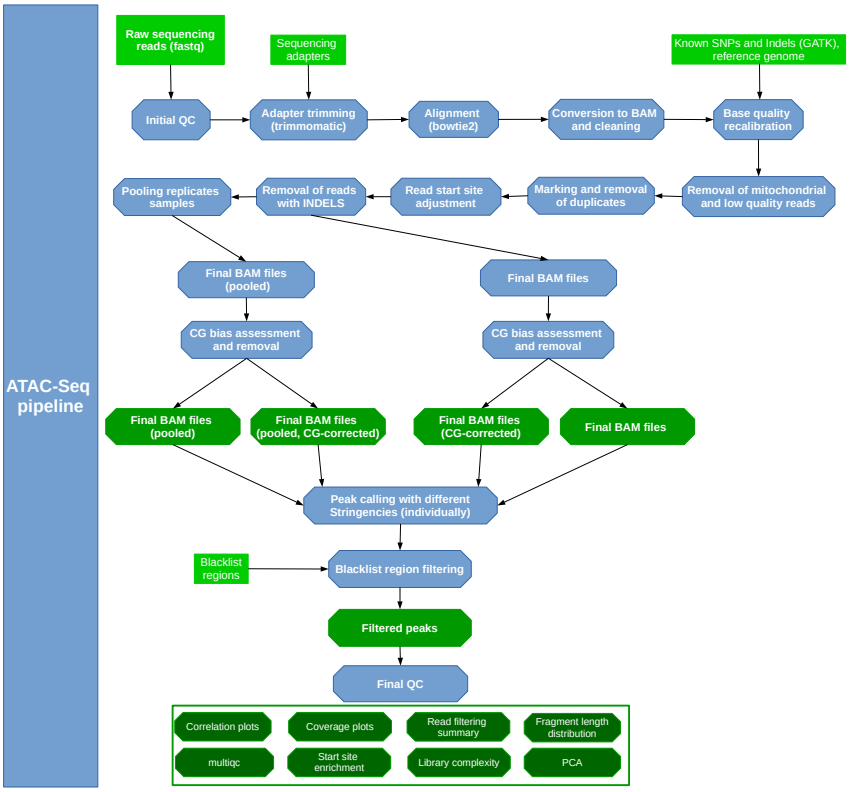

Figure 2.1: **Schematic workflow of the ATAC processing pipeline**

2.2.3 Description of `diffTF` basic mode workflow

`diffTF` is a complex computational pipeline that estimates global differential TF activity for the pairwise comparisons between two or several conditions (basic mode) and predict TF mode of action by integrating gene expression data (classification mode), conceptual scheme shown in Figure 2.2. In the basic mode we summarize genome-wide readout of changes in chromatin accessibility across multiple TF binding sites and perform a statistical test to obtain TFs that are significantly different between conditions. Whereas in the classification mode, by correlating TF expression and chromatin accessibility changes in the TF binding sites, and comparing them to the background, we predict global TF mode of action, for example activator or repressor.

We can split the description of the `diffTF` basic mode to the 7 consecutive steps, which are described below.

Generate consensus peakset

In order to have the same search space for all of the samples used in the analysis we need to define a common consensus peakset. Users can either provide this information or such a peakset will be generated based on the sample-specific peaks. By using `DiffBind` (Ross-Innes et al., 2012) Bioconductor package functionality we define consensus peakset across ATAC-seq samples with a parameter *minOverlap* that state the amount of samples which should contain the peak. Afterwards we only keep peaks from autosomes and sort them by coordinate, which is required for further steps.

Predicting of TF binding sites

One of the essential principles of `diffTF`, is that it is region-oriented, therefore it requires an input set of predicted TF binding sites. We use `PWMscan` tool (Ambrosini et al., 2018) to scan hg19 reference genome with the set of TF binding models/PWMs downloaded from HOCOMOCO v10 database (Kulakovskiy et al., 2015). By default, we are using cutoff p-value equals to 0.00001 and background composition that is similar to the human genome (0.29;0.21;0.21;0.29). Finally we sort the obtained *bed* files of the predicted TFBS by coordinates. By default we are providing this set of TFBS, however, it is important to note that one can use a completely different set of predicted TFBS.

Differential accessibility consensus peaks analysis

First of all, we calculate sample specific counts for the consensus peaks using the *featureCounts* function from the Subread R package (Liao et al., 2019) [internal parameters: -p -B -d 0 -D 2000 -C -Q 10 -O -s 0]. After that, we use obtained counts in the DESeq2 (Love et al., 2014) and determine the fold change for our pairwise comparisons for each peak. By default, at this step we are using cyclic loess normalization defined by *normOffsets* function from csaw (Lun and Smyth, 2016) Bioconductor package. As compared to the default DESeq2 normalization by library size, we found out that loess normalization is more robust to the chromatin accessibility differences between the samples. One of the diffTF features that distinguishes this algorithm from similar ones, is that it can use various design formulas in the differential accessibility analysis, therefore potential batch effects and covariables can be handled to improve statistical output of the analysis. During this step in the pipeline we generate several diagnostic plots (MA plots, density plots of counts before and after normalization, mean standard deviation plots), which can help users to assess whether normalization is working correctly for the input samples.

As diffTF can use all types of design formulas supported in DESeq2, it is also applicable to the time series data. Described above fold changes are then calculated as per unit of change of the variable. For example, negative fold changes would mean that overall it was negative slope per unit of change, whereas positive fold changes would mean opposite behaviour - positive slope.

Read counts for the intersected TFBS

At this step we overlap predicted binding TFBS with the consensus peaks using bedtools (Quinlan and Hall, 2010) [internal parameters: -wa -wb] and extend each overlapped TFBS by some value (default: 100bp) in both ends. Afterwards using same parameters as for consensus peaks we calculate read counts per TFBS using *featureCounts* function from the Subread R package (Liao et al., 2019). From the obtained matrix of peaks and overlapped TFBS we choose one TFBS of certain TF per peak with the biggest average read counts across samples.

Differential accessibility analysis for each TFBS

After counting reads for the intersected TFBS we proceed with differential accessibility analysis for this sites using limma (Ritchie et al., 2015) R package with *lmFit*

and *eBayes* functions. At this step we are also utilising the same design formula and normalization factors that were defined in the differential accessibility analysis of the peaks. As a result of this analysis we obtain log2 fold change values for each of the selected TFBS and various diagnostic plots, similar plots that we generated for the differential accessibility of the consensus peaks and additionally ECDF and density plots summarizing log2 fold changes per TF.

Calculation of the differential TF activity

As different TFBS are positioned in the local environments varying by their CG content and this can strongly affect chromatin accessibility of these sites, we are adding CG binning step in order to summarize TF activity per each TF. First we calculate CG content of the extended TFBS with bedtools (Quinlan and Hall, 2010) *nuc* function. Then we group all extended TFBS in the *n* bins (default: 10) based on their CG content. For each TF for each bin that has at least 20 TFBS, we compare the TFBS distribution of log2 fold changes with all log2 fold changes of all TFBS from all TFs that are assigned to the same CG bin, and calculate the difference in means between these two distributions (TF-specific vs all). As we have *n* amount of bins, the differential TF activity for each TF is defined as weighted mean difference summarized across all CG bins, using as weight percentage of TFBS of this TF in each bin.

Estimation of significance for differential TF activity

To assess statistical significance and the magnitude of the resulting differential TF activity for a large datasets we recommend using a permutation approach. In essence, we permute *n* times (default: 1000; dependent on the dataset size) our metadata table and repeat steps involving computation of differential accessibility (peak- and TF-specific) and then calculate empirical two-sided p-value per TF by comparing the real value of differential TF activity with the permuted ones. Magnitude of the p-values is related to the number of the samples in the input data (minimum p-value equals to the 1/ *n* permutations) and can be defined with the binomial coefficient *(j/k)* with *j* being the sum of samples across all dataset and *k* amount of the samples for the smallest condition. Finally, we apply multiple testing correction using Benjamini-Hochberg correction (Benjamini et al., 1995) and plot the distribution of the permuted differential TF activity values as compared to the real one. Such an approach is heavily computationally expensive. Finally, we observed no correlation

between the number of TFBS per TF and the p-value obtained.

On the other hand, if the number of samples in a dataset is small, and it is impossible to calculate meaningful p-value, we offer a different approach to assess the significance of the differential TF activity that we called analytical approach. Following this approach to calculate log2 fold changes for the consensus peaks and TFBS we use DESeq2 instead of limma workflow, which is appropriate for the small number of samples. Instead of starting permutations, we run extra steps to define statistical significance of the differential TF activity. From the CG binning step we perform for each bin per TF Two Sample Welch t-test and transform the resulting t-score of the difference to z-score, which allows us to summarize them across multiple bins per TF to the one p-value. As we calculate differential TF activity as weighted mean difference between the CG bins, we also need to calculate the weighted mean of the T-scores with estimation of the expected variance.

For the variance estimation for each TF we use propagation of uncertainty formula in which ω_i and x_i represent the weight, as ratio of TFBS for each TF, and T-score from the t-test for each bin *i*. $Cov(x_i, x_j)$ is the covariance of two independent bins *i* and *j*, and $var(x_i)$ is the variance of the T-score for x_i.

$$var(\underline{x}) = \sum_{i=1}^{n} var(\frac{\omega_i}{\sum_{k=1}^{n} \omega_k x_i}) + \underbrace{2 \sum_{i \geq 1}^{i<j} \sum_{j>i}^{j \leq n} \omega_i \omega_j Cov(x_i, x_j)}$$

(weights sum up to 1)

$$var(\underline{x}) = \sum_{i=1}^{n} var(\omega_i * x_i) + k$$

If all values are scaled by constant, the variance is scaled by the square of the constant.

$$var(\underline{x}) = \sum_{i=1}^{n} \omega_i^2 var(x_i) + k$$

This scaling is done to prevent a biased estimation of the variance with the assumption that sum of it should be 1 for all bins. To not be biased for this, we add to the variance $var(x)$ calculation bootstrap using the boot library in R with a number (default: 10000) of bootstrap replicates. During this we resample the bin-specific data and use t-test against the full sample. After this we calculate the variance $var(x_i)$ of the bootstrapped T-scores per each bin per each TF. Since we can guarantee that the T-scores across all bins for each TF are independent, we correct this variance using the pairwise covariance, that is estimated by $Cov(x_i, x_j)$ for each

bin using bootstrap data. From the 9,099 covariances for all the combinations of TFs and bins, only 19 cases exceeded an absolute value of 0.05 and only one 0.2, while over 91% had absolute values of smaller than 0.02, which make us believe that such covariances have very small effect and may be neglected in the analysis.

Afterwards we calculate weighted T-score for all TFs and centralize the distribution by subtracting the mean, which we predict using maximum likelihood estimate of the distribution mean with *locfdr* function from the `locfdr` (Efron, 2007) R package. Then we calculate p-values out of the obtained z-scores using the variance as estimated above, followed by multiple testing correction of this p-values using Bejamini-Hochberg method (Benjamini et al., 1995).

2.2.4 `diffTF` classification mode

There are already published databases, like TRRUST (Han et al., 2018), that classify TFs on the mode of action based on text-mining in the published research. Nonetheless, when we tried to classify most differentially active TFs from the `diffTF` basic mode using this database, we observed that most TFs are classified equally often both as activator and repressor. This classifying strategy is of course biased to the amount of published papers for a specific TFs and does not take into account different cell types and conditions, which makes it difficult to judge objectively about data specific TF mode of action.

To overcome this limitation, we decided to employ a novel data-driven approach for classification of TFs having global activating or repressing functions in the genome, based on the integration of TF specific gene expression and chromatin accessibility at the predicted TFBS, conceptual scheme shown in Figure 2.2. Major assumption of this method is that if TF is predominantly activator, then increasing expression of it would lead to the increased chromatin accessibility of its putative binding sites, whereas for repressive TFs, increased expression will result in the decreasing chromatin accessibility.

We start with the normalization of the RNA-Seq count data for TFs (default: quantile normalization) to reduce effects from gene expression outliers on the further correlation analysis. After that we calculate the Pearson correlation coefficients between the chromatin accessibility read counts of each TFBS per individual TF and respective TF expression from each sample. Median of the resulting distribution serves as a classifier for the mode of TF action. TFs with more positive correlations are predicted to be activators and TFs with larger ratio of negative correlations are

repressors. However, on this step we also calculate the correlation median for the background set of regions for each TF, that consists of correlations from all TFBS, excluding TF-specific TFBS, and TF expression. Such measure allows us to estimate the noise level for each specific TFs. To distinguish real dependency of accessibility from the TF expression from the noise, we use percentiles of the background distribution across all TFs, as a cutoff for defining activators and repressors from the undetermined TFs. In the output classification we provide several variants of such classification based on the different stringency thresholds (default: 0.1/0.9, 0.05/0.95, 0.01/0.99, 0.001/0.999, 0.00001/0.9999) that allow user to choose needed level of stringency and flexibility for the downstream analyses.

At last, we measure if the foreground distribution of correlations (unique TFBS per TF) and background distribution (all TFBS without unique TFBS per TF) are significantly different between each other using one-sided Wilcoxon rank sum test. P-value from this test serves as the additional threshold for defining activators and repressors. TFs that were classified in previous steps as activators or repressors, but are not significantly different from the background distribution, are moved to the undetermined class. We store resulting values in the output files and provide a visual representation of such correlations per each TF and summarized across all TFs used in the analysis. This analysis was originally conceived by Dr. Armando Reyes-Palomares.

2.2.5 Benchmarking robustness of diffTF

Importance of the TFBS scanning parameters

The essential part of diffTF analysis is defining the regions in the genome that are potentially bound by the TFs. We are limiting predicted TFBS to the space of the open chromatin peaks, but as we used external PWMscan (Ambrosini et al., 2018) algorithm, we need to control if diffTF is unbiased from the cutoffs used in the scanning procedure. For this we tested several p-value cutoffs in PWMscan (0.00005, 0.00001, 0.000001) and generated for these cutoffs different sets of predicted TFBS. Also, we generated different sets of TFBS specific for the consensus peakset background composition (0.27;0.23;0.23;0.27), as used by default nucleotide background composition ratios of the human genome (0.29;0.21;0.21;0.29). We transformed *bed* file of consensus peaks to the *fasta* file with bedtools *getfasta* function, and then calculated respective background nucleotide composition using *fasta-get-markov* function from the MEME suite (Bailey et al., 2009).

To compare different databases storing TF binding models information (in this case HOCOMOCO v10 vs JASPAR 2016), we downloaded the respective PWMs from HOCOMOCO and PFMs from JASPAR and converted them into integer log likelihoods models with *pwm_convert* function from `PWMscan` standalone version [internal parameters: -f "real" for HOCOMOCO; -f "jaspar" for JASPAR]. Finally, for all the scenarios we performed prediction of possible TFBS using *pwm_bowtie_wrapper* script from `PWMscan`. Part of these analyses were done by Giovanni Palla.

Impact of the differentially accessible peaks

A lot of approaches of ATAC-seq data analysis are centered around differentially accessible peaks. Therefore, we tested if the signal from these peaks is prevalent in comparison to all other peaks from consensus peakset. We used `DiffBind` (Ross-Innes et al., 2012) R Bioconductor package [design formula: *batch_number* + *Condition*] and determined 389 differentially accessible peaks for mutated CLL and 3569 peaks for unmutated CLL. We excluded these peaks from the consensus peakset and reran `diffTF` on the whole CLL dataset using the default parameters. Obtained differential TF activity we correlated using Pearson correlation with TF activities gained from `diffTF` run with all peaks.

Validation using ReMap ChIP-seq data

For the validation of our predicted TFBS we used data from human ChIP-seq experiments collected by ReMap (Griffon et al., 2015) and intersected them using bedtools *intersect* function with respective predicted TFBS. Out of 640 human TFs available from HOCOMOCO v10 we found overlap only for 157 common TFs from ReMap. We reran `diffTF` pipeline using an intersected set of TFBS and excluding the intersected set from the predicted TFBS (potentially noise). After this we correlated these groups with each other and with the all predicted TFBS from HOCOMOCO v10.

Robustness to the internal `diffTF` parameters

As we checked the robustness of `diffTF` to the external scanning parameters, we also need to verify if the signal that we obtain is robust to the internal `diffTF` parameters. One of the key parameters is the motif extension parameter [default: 100bp]. We consistently changed only the motif extension width starting from 0bp to 600 bp (0,50,100,200,400,600) and reran `diffTF` pipeline. Also, as discussed at the previous section regarding the estimation of p-value, we checked the effect of

amount of permutations for permutation approach and amount of bootstraps for the analytical approach and observed that output of diffTF is not biased on these parameters. This analysis was done by Dr. Christian Arnold.

Dependence on the dataset size and depth of sequencing

In order to check the dependence of the diffTF results on the number of samples per dataset or per condition and sequencing depth of coverage of ATAC-seq data, we implemented a subsampling scheme, repeating the diffTF workflow changing read depth and sample size. Importantly, for this analysis we used the whole CLL dataset consisting of 84 ATAC-seq samples. For the subsampling based on the read depth, we used non-GC corrected CLL dataset with 84 samples from 52 individuals and randomly downsampled *bam* files using *DownsampleSam* function from Picard tools with random seeds and *probability* = k, ranging from 0.01 to 1 (0.01;0.02;0.06;0.125;0.25;0.5;0.75;1). For the downsampled *bam* files we calculated the median number of reads across all samples to determine the estimate of the sequencing read depth of the sample required to gain particular accuracy. After this we permute also the number of samples per condition, however, keeping the original ratio of the samples from the initial dataset (60% of unmutated CLL vs 40% of mutated CLL samples). The calculated ratios of conditions in the diffTF runs were varying from 34+50 to 3+5 (decreasing to the 9+14 step by 5 and 7 respectively; and from 9+14 to 3+5 in steps of 1), where the first and second number correspond to the absolute number of distinct unmutated and mutated samples, respectively. For each of these subsampling schemes, we performed 50 repetitions to minimize the effect of the sampling noise. After running diffTF for each subsampling we matched the results with the differential TF activity observed in the full dataset and calculated the fraction of significantly differentially active TFs that show the same sign of change with the full dataset. To differentiate changes between TFs with different effects of signal, we separated all TFs into 3 bins using 0.33 and 0.66 percentile threshold of the differential TF activity from the full data. This analysis was performed by Dr. Christian Arnold.

2.2.6 Comparison with similar tools

Comparison with HOMER

To run HOMER (Heinz et al., 2010), as one of the standard motif enrichment analysis tools, first, we extract *fasta* files from *bed* differentially accessible peaks for each condition. Using obtained *fasta* files we generated a set of two-fold background sequences using `BiasAway` (Hunt et al., 2014) software, taking to the account the length and GC bias as main covariables. Lastly, we used `HOMER` (Heinz et al., 2010), setting as foreground *fasta* files from the differentially accessible peaks and as background corresponding unbiased *fasta* files for each of the conditions. To compare obtained enrichments with `diffTF` results we transformed them to the percentage of differentially accessible peaks enriched for the given motif and correlated them with TF activities from `diffTF` using Pearson correlation.

Comparison with `chromVAR`

To compare `diffTF` results with the output of `chromVAR` (Schep et al., 2017) we adjusted the `chromVAR` framework to work with the same input files. We started by importing to R all TFBS that were predicted with PMWscan from HOCOMOCO v10 database using `chromVAR` *getAnnotations* function. To have similar regions of interest in the genome we also loaded our set of consensus peaks with *getPeaks* function without resizing to keep them as similar as possible to the `diffTF` settings. Using this data and metadata related to the CLL dataset we counted fragments in paired-end mode from the *bam* files with the function *getCounts* [internal parameter: *by_rg = FALSE*]. Resulted counts matrix was corrected for the CG bias using *addCGBias* function and filtered for non-overlapping peaks with default parameters. To normalize the resulting counts for each condition we computed expectations, which summarized the average fraction per peak in each sample for the two different conditions. Fragment counts matrix, background set of peaks, expectations and peak-TFBS matrix were used to compute deviations using *computeDeviations* function. AS chromVAR was originally designed for the single-cell data, and does not provide log2 fold changes for the pairwise comparisons, we needed to summarize sample specific deviations and deviations scores. As for the `diffTF` we used weighted mean difference between the CG bins as main measure of effect size, we decided to also use mean of deviations within each condition in the chromVAR to define one number of TF activity per TF per condition. It is important to note that as `diffTF` utilize more complicated analysis to define log2 fold changes it is not prone to outliers as much

as deviations from chromVAR summarized by simple mean. Finally, we calculated Pearson correlation coefficient between summarized deviations/deviations scores and diffTF TF activity values. Majority of this analysis was done by Dr. Christian Arnold.

2.3 Results

We developed `diffTF` as a complex computational workflow to assess genome-wide differential TF activity based on the chromatin accessibility or chromatin marks distribution between two or multiple conditions in a pairwise manner (basic mode) and to provide data-driven classification of TFs based on their mode of action (classification mode), see Figure 2.2. In the basic mode `diffTF` calculates log2 fold change for each of the predicted *in silico* TFBS and summarize them globally across genome to define differential TF activity while normalizing for GC content. Depending on the dataset size, the significance is estimated using a permutation approach or an analytical approach, for a detailed explanation see section 2.2.3.

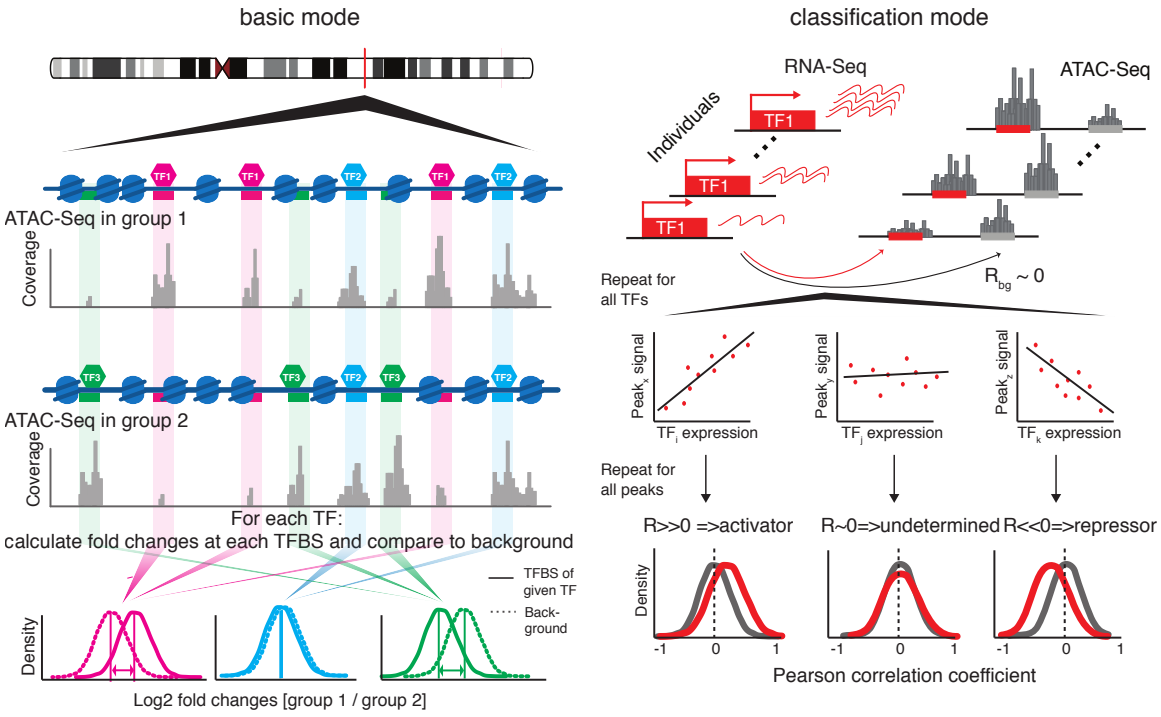

Figure 2.2: **Schematic representation of the `diffTF` workflow.** On the left scheme of the `diffTF` basic mode: For each TFBS of a given TF, the fold change between the two conditions is computed, followed by comparing their distribution to a background set of fold changes obtained from GC content-matched loci that do not contain the putative TFBS. On the right description of the `diffTF` classification mode: TF expression levels are correlated with the accessibility of their target sites. If correlations with its target sites are more positive than with the background distribution (non-target sites), the TF is classified as putative activator; if they are more negative than with the background, it is classified as putative repressor; and if they are indistinguishable from the background, it is classified as undetermined. This figure was contributed by Dr. Judith Zaugg and is published in (Berest et al., 2019).

In the classification mode, `diffTF` utilizes additional RNA-seq data and calculates correlation of TF expression and every TFBS for each sample. We classify each TF into putative activators (having mostly positive correlations), repressors (mostly

negative correlations), or undetermined (similar to background) class, by comparing the distribution of correlations between TF abundance and peaks with predicted binding sites against all other peaks, see described analysis in the section 2.2.4.

All the following benchmarking analyses were completed on the CLL dataset, see section 2.2.1 for the source of the data. In the next methods section 3.2.2 we will discuss in detail all preprocessing parameters and results of diffTF pipeline applied on this dataset. However, as such dataset is still one of the biggest up to date bulk multiomics datasets (including ATAC-seq and RNA-seq), it is particularly useful to perform various benchmarking analyses of our pipeline on it in order to define limits and potential biases of diffTF.

2.3.1 diffTF is robust to the internal and external parameters

Before investigating biological interpretation of the diffTF results on the CLL dataset, described in the Chapter 3, we used this dataset to check the effects of different internal diffTF and external programs parameters on the final output, see methods from section 2.2.5.

As one of the main input data for diffTF are predicted *in silico* TFBS (Ambrosini et al., 2018), first, we concentrated on assessing potential influences of the different parameters of TFBS predictions on the diffTF results. Varying the p-value threshold in the PWM scanning, we observed that overall the correlation for the differential TF activities derived using more lenient p-value as compared to default (5e-5 versus 1e-5) is quite high (R=0.87; Figure 2.3B), suggesting that we capture most of the signal. At the same time by using more stringent parameters (1e-6 versus 1e-5) we have lower correlation (R=0.62; Figure 2.3A), however the amount of predicted TFBS with such parameters drops significantly. The proposed by authors p-value cutoff of 1e-5 seems to be optimal for PWM scanning by capturing most of the signal and still having sufficient amounts of TFBS per each TF. Similarly, we found that by changing nucleotide background composition for the PWM scanning from genome-wide to the consensus peaks specific we are not altering the signal (R=0.93; data not shown). diffTF seems to be also robust to the input TF binding models that were used from the different databases. For 412 common TFs from HOCOMOCO v10 and JASPAR 2016 databases we observed high correlation (R=0.69; Figure 2.3C), considering that normalization steps and binding models formats are different for each database. We also varied internal diffTF extension size from 0bp to 600bp, but haven't observed significant differences in the diffTF results.

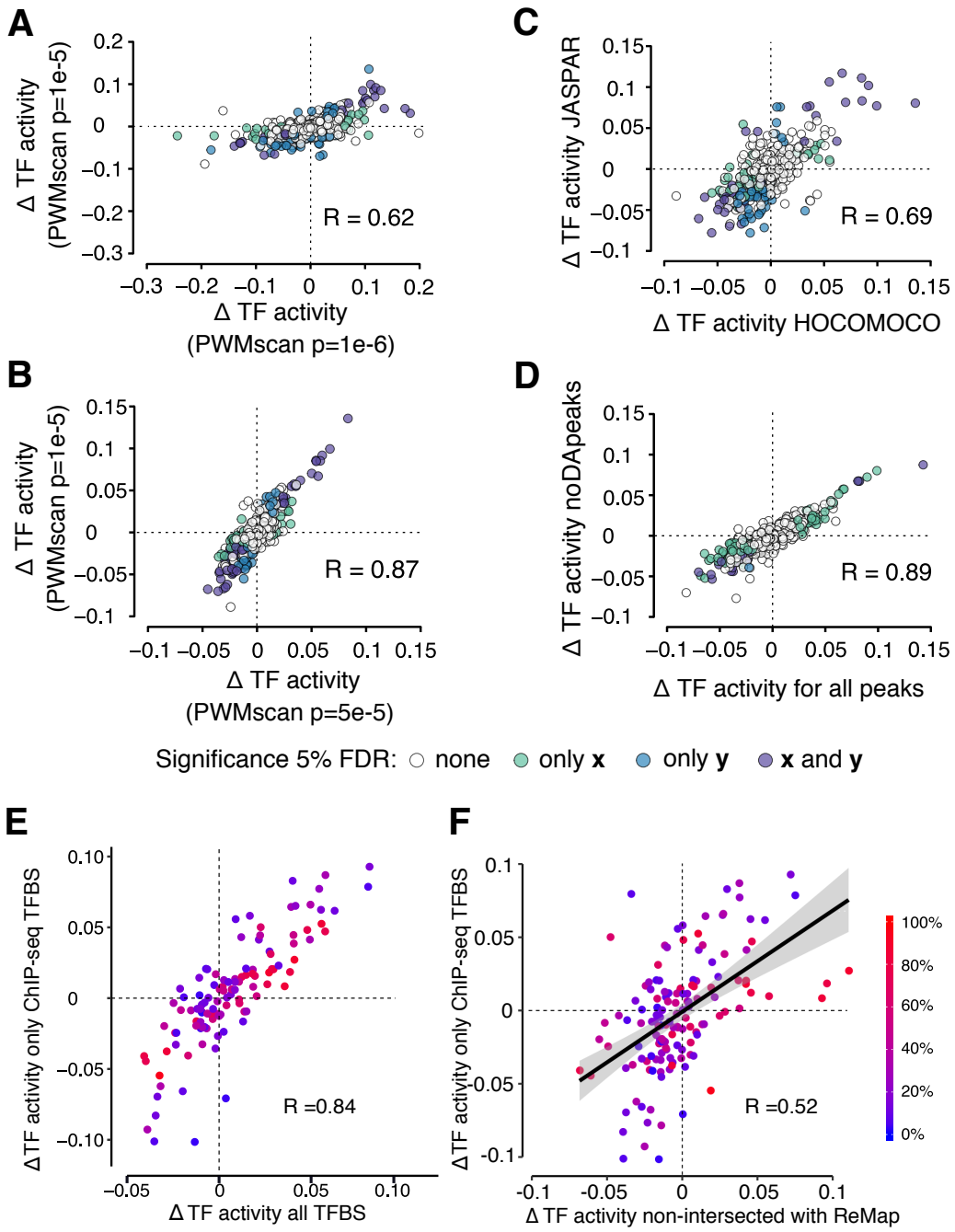

Figure 2.3: **Technical robustness of `diffTF` results in basic mode.** (A and B) Comparisons between p value thresholds in PWMScan to predict TFBS (n = 628 TFs): (A) standard (1e-5) versus stringent (1e-6) and (B) standard (1e-5) versus relaxed (5e-5). (C) Comparisons between different motif databases (HOCOMOCO v10 versus JASPAR 2016; n = 412 TFs). (D) Comparisons between different peak sets (full consensus peakset [all Peaks] versus non-differentially accessible peaks [noDApeaks]; n = 640 TFs). For (A)-(D), R indicates Pearson correlation and TFs are colored by `diffTF` significance (5% FDR) in the compared analyses (white, not significant; light green or blue, significant for the x axis or y axis only; purple, significant in both). (E-F) Scatterplots of the differential "TF activity" related to ReMap. (E) Comparison of all predicted TFBSs and TFBSs experimentally validated by ChIP-seq data from ReMap. (F) `diffTF` results only using TFBS not intersecting with ReMap against results from only TFBS intersecting with ReMap. For each TF (n = 157), the percentage of TFBSs that overlap ChIP-seq data is indicated from blue (0%) to red (100%). This figure was produced by myself and was published in (Berest et al., 2019).

All performed above benchmarks showed the robustness of our pipeline to the parameters for TF binding sites prediction. One of the possible reasons for this is the fact that diffTF summarizes signal across many TFBS genome-wide and therefore obtained signal is much stronger than the false-positive prediction noise.

By aggregating differential signals across multiple TFBS per TF diffTF can show differences between samples in experiments with low biological signal. We tested this by excluding differentially accessible regions (FDR=5%) from the consensus peakset and rerunning diffTF analysis. Resulting TF activities showed high correlation (R=0.89; Figure 2.3D) with the default full consensus peakset, therefore supporting the hypothesis that diffTF can be applied to low-signal experiments.

As previously reported (Jayaram et al., 2016; Landt et al., 2012), TF binding sites *in silico* predictions are intrinsically noisy and have a high ratio of false-positives when compared to the experimental ChIP-seq experiments, we decided to compare diffTF robustness via changing predicted TFBS to the ChIP-seq peaks. Overall, while computing TF activities for the 157 common TFs between used predicted TFBS and ReMap (Griffon et al., 2015) data we observed strong correlation (R=0.84; Figure 2.3E) with TF activities derived from predicted TFBS. When excluding from all TFBS the ones that overlap with ChIP-seq peaks we still observed sufficient correlation (R=0.52; Figure 2.3F), suggesting that potential low affinity binding sites not recognized by ChIP-seq but predicted *in silico* show the signal that has the same direction high affinity binding sites. The analysis described above suggests that diffTF utilize the power of the quantity of predicted *in silico* TFBS, however, does not lose in the quality of signal.

Considering the fact that we are working with very large ATAC-seq datasets, we decided to perform subsampling experiments which provide guidelines to the sample sizes needed for the diffTF analysis. We found that diffTF analysis results in the highly consistent results across permutation on the sample size and sequencing depth for both significant (Figure 2.4A) and not significant differential active TFs (Figure 2.4B). Taking into account subsampling analysis we state that for diffTF the number of samples are more important than the sequencing depth of the samples. We observed that even with 1 million reads per sample while using all samples from dataset, we had a very high correlation (R=0.92) with the full dataset that has 21.4 million reads on average for the significant TFs. Such observations are in line with the recent statements about the robustness of single-cell ATAC-seq data analyses to low coverage for genome-wide summary statistics (Mezger et al., 2018).

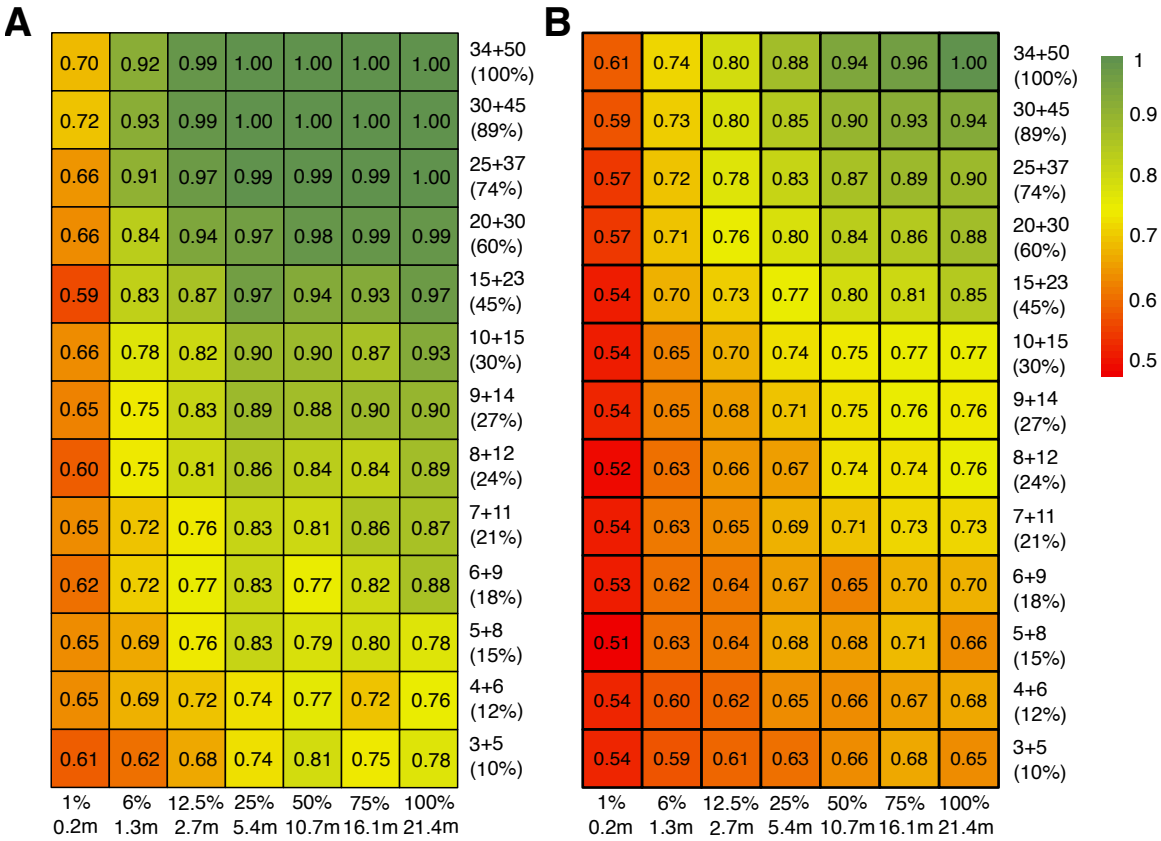

Figure 2.4: **Robustness analysis for the sequencing depth and sample size.** Each cell in the heatmap shows the fraction of TFs that have the same direction of change as in the full dataset for varying degrees of down-sampling sequencing depth and number of samples, averaged over 50 independent repetitions to minimize sampling noise. Sequencing depth is shown as a fraction of the original data and median number of reads across samples, while the number of samples is given as unmutated + mutated. (A) Only TFs that were deemed significant in the full dataset are considered (5% FDR). (B) The fraction of TFs that were not assigned as significant with the full data are shown. This figure was contributed by Dr. Christian Arnold and is published in the (Berest et al., 2019).

In summary, all results described above showed that `diffTF` analysis is robust either to external and internal parameters in detecting differential TF activities, and highlight the importance of summarizing signals from TFBS genome-wide, as efficient and sensible way to deal with technical limitations, such as little biological variation between conditions and low coverage.

2.3.2 Comparing `diffTF` results with similar tools

Lastly, we quantitatively compared `diffTF` pipeline results with modern computational tools widely used in the field that define motif enrichment and activity (Baek et al., 2017; Heinz et al., 2010; Schep et al., 2017). More technical description of the following analysis is described in the section 2.2.6.

We started by comparing `diffTF` with a typical TF motif analysis from HOMER

(Heinz et al., 2010), that by matching two sets of sequences identify TF motifs that enriched in one set relative to another. As described by HOMER tutorial, we used a set of differentially accessible condition specific peaks and compared them to the set of two-fold background peaks defined by BiasAway (Hunt et al., 2014). Notably, we haven't found any enriched motifs for M-CLL, while 32 of them pass the significance threshold of 10% FDR for U-CLL. This can be due to the fact that for M-CLL samples only 389 peaks were differentially accessible, as compared to 3569 differential peaks defined in M-CLL. We correlated results of HOMER motif enrichment analysis and `diffTF` differential TF activity, and observed significant correlation (R=0.63; Figure 2.5B) between the significant TFs from both analyses. Such correlation for U-CLL samples supports the results of `diffTF`, however also highlight the impact of the extra step of defining differential accessible peaks for the motif enrichment analysis. Since `diffTF` uses the union of peaks from the open chromatin in all samples, it tends to capture more signal than currently available differential peaks-oriented motif enrichment approaches.

We also compared `diffTF` to the computational tools that have similar input files and estimate, as final output, measures similar to `diffTF` differential TF activity, chromVAR (Heinz et al., 2010) and BaGFoot (Baek et al., 2017). Unfortunately, we were incapable of using our CLL dataset with BaGFoot workflow due to the incomplete documentation and absence of the example files.

By comparing differential TF activity from `diffTF` to chromVAR results, which was designed for analysis of single cell ATAC-seq data, we detected high correlation overall (see Figure 2.5C-D). As chromVAR outputs two final measures (deviations and deviations scores) we used both of them for correlation analysis, and observed Pearson correlation between 0.75 and 0.93 dividing also by TF significance from `diffTF` (splitting into 2 groups with a threshold of 10% FDR). Highly significant differential active TFs from `diffTF` have the strongest correlation with deviations from chromVAR analysis.

Even though we quantified very strong significant correlation between chromVAR and `diffTF` results, there are distinct methodological discrepancies between them worth noticing. First, chromVAR counts ATAC-seq fragments for peaks and not exactly TFBS, with subsequent assignment of them to TFBS overlapping certain peaks. On the contrary, `diffTF` always work in the TFBS specific space and count the exact number of reads overlapping with TFBS. While we observed that overall log2 fold changes for peaks and respective TFBS are highly correlated (R=0.91; data not shown), the range of TF-specific differences can vary greatly depending on the TF.

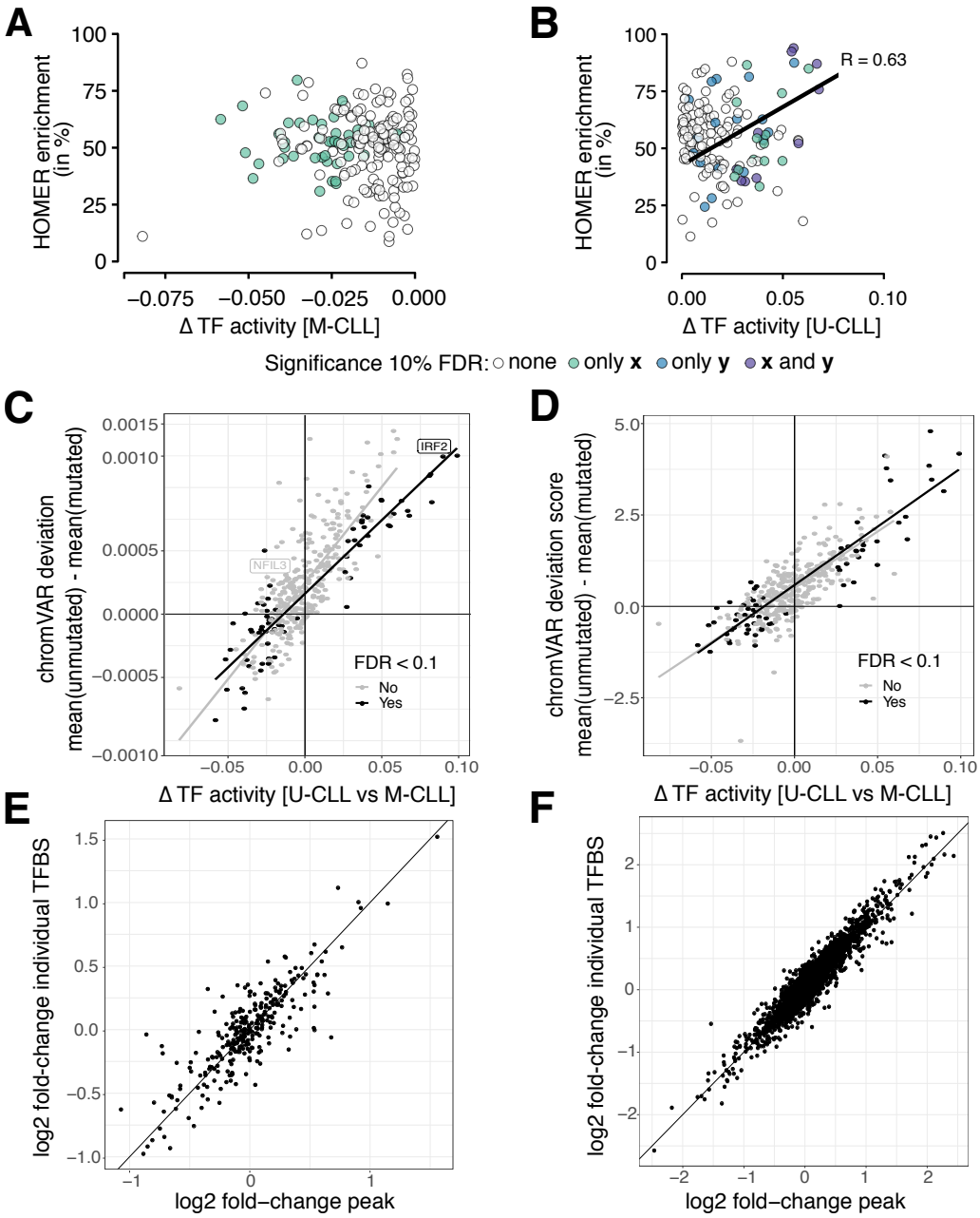

Figure 2.5: **Comparison of `diffTF` with similar tools.** (A-B) Comparison with HOMER for M-CLL (A) and U-CLL (B). Each point represents one TF. The x-axis shows the differential "TF activity" from `diffTF`, while the HOMER enrichment (in %) is shown on the y-axis. Colors represent significance (FDR < 10%): white - not significant in either analysis; light green and light blue - significant only for the analysis from the x-axis or y-axis, respectively; purple - significant for both analyses. Spearman correlation was computed only for TFs with FDR < 10% for the adjusted p-values from the HOMER enrichment using Benjamini-Hochberg. 0 and 32 motifs were enriched in HOMER in (A) and (B), respectively. In (B), a linear model is shown as a black line. (C-F) Comparison plots for `diffTF` vs. chromVAR. (C)-(D): The `diffTF` "TF activity" is shown on the x-axis and the chromVAR deviation (C) and deviation score (D) on the y-axis, as measured by the difference of the means between the two conditions U-CLL and M-CLL. Only the 370 expressed TFs from the CLL analysis are included. The two TFs that are mentioned in (E) and (F) are labeled in (C). (E)-(F): Correlation of log2 fold-changes between U-CLL and M-CLL from peaks versus individual TFBS for two selected TFs. Note that for each peak, the TFBS per TF with the highest average read count across all samples was selected (see section @ref(basicdiffTF). (E) TF with the lowest correlation among all expressed TFs from the second quadrant (NFIL3, 290 TFBS). (F) TF with the highest correlation among the set of significant TFs according to `diffTF` (IRF2, 4362 TFBS). This figure was partly (chromVAR analysis) contributed by Dr. Christian Arnold and is published in the (Berest et al., 2019).

TFs with a low number of binding sites, such as NFIL3 (Figure 2.5E), can be affected by this more, as peak-TFBS variation is stronger there due to the sample size, whereas TFs with a high number of binding sites (Figure 2.5E-F) seems to not have such strong variation. Also, diffTF calculates differential TF activity always compared to the mean effect of differences in chromatin accessibility using common space of TFs, therefore the resulting measure is relative. Whereas, chromVAR is defining deviations summarizing only chromatin accessibility for only one TF across all samples and assigning rather absolute value for each TF and each sample. Such phenomenon can explain the observed shift to the positive side for the chromVAR deviations values (424 TFs > 0; 66.3%; Figure 2.5E-F).

As chromVAR is not designed for bulk datasets, it never precisely compares the differences in TF activity between two conditions (f.e. by computing log2 fold-changes), however it can take into account sample metadata to calculate condition-specific expectations. It provides one deviation value/score for each sample and each TF, which was recommended by the authors to summarize by calculating the mean deviation within each condition. The resulting measure is similar to the diffTF differential TF activity, as we can see from the high correlation between these two methods, though it is prone to outliers. diffTF does not have this problem as it uses log2 fold-changes between pairwise comparisons taking into consideration different covariates (f.e. batch, gender) that can be stated in the design formula.

In conclusion, the comparisons described above reported comparability between diffTF and state-of-the-art tools in motif analysis, although featuring strength of diffTF approach. It shows more sensible results, as compared to HOMER, and provides a more dynamic and detailed workflow structure than chromVAR. As compared to the chromVAR it also saves output from different steps of pipeline, which can be used for various downstream analyses, and has higher flexibility of the workflow. In addition to this diffTF goes one step further by directly integrating chromatin accessibility data and gene expression with further TF classification into activators or repressors, thus providing additional insights into their molecular function.

2.4 Discussion

Here, we presented a novel method for calculating genome-wide differential TF activity from chromatin accessibility for a large set of TF motifs at the same time, called `diffTF` (comprehensive documentation and starting vignette available at `https://diffTF.readthedocs.io/`). Instead of chromatin accessibility data, as input to the `diffTF`, without changing global assumptions sequencing data from the active histone marks (e.g. H3K27ac) can be used (Reyes-Palomares et al., 2020). By integration with gene expression classification mode of this method is able to classify TFs by their general mode of action to the transcriptional activators or repressors in the pairwise comparisons.

The idea of summarizing genome-wide correlations between TF expression and putative target genes followed by defining differential TF activity based on such correlations was already described before (Boorsma et al., 2008; Bussemaker et al., 2001). Recent advances in the chromatin accessibility sequencing techniques, such as ATAC-seq development, entitle us to use chromatin readout instead of gene expression to calculate TF activity. First, chromatin is an upstream cellular phenotype compared to the gene expression. Chromatin takes into account TF expression, posttranscriptional and posttranslational modifications, and ideally acts as direct sensitive readout of the gene regulation. However, defining correct TF binding patterning based on the DNA sequence (Movva et al., 2019) and chromatin data is still a limitation and further research in this area is required. Also, usually there are much more opened peaks than genes, therefore enabling us to use stronger statistical methods and increasing signal-to-noise ratio. Third, compared to the enhancers, which regulate gene expression and often not well-defined in each cell, effects on chromatin are usually locally defined.

As we described in this chapter we designed different statistical methodologies for the estimation of significance of differential TF activity for the datasets with different sample sizes. We recommend to use, so-called permutation approach, with the heterogeneous datasets containing at least 15 samples. For the small-scale datasets, which are often generated nowadays, we recommend using an analytical approach, that is summarizing differences in each CG bin. We extensively tested `diffTF` by changing different internal (motif size extension; peakset with and without differentially accessible regions) and external parameters (TF motif scanning parameters; using ChIP-seq ReMap data as a source of TFBS) and demonstrated the technical robustness of `diffTF` basic mode. For the tested dataset `diffTF` tackled

expected technical noisiness of TFBS prediction by summarizing signals across many binding sites genome-wide. Using the power of the big dataset we identified that diffTF is more sensitive to the amount of samples in dataset compared to the sequencing depth of the samples.

Compared to the analogous methods that use chromatin accessibility as readout of TF activity (Baek et al., 2017; Schep et al., 2017), diffTF is specifically designed to analyze bulk ATAC-seq data and can accept as input ChIP-seq data for the histone marks. Apart from this, diffTF integrates gene expression and chromatin accessibility data and classifies TFs into activators or repressors. As diffTF operates with log2 fold changes of the accessibility for each TFBS, such information is saved and can be used for the further downstream analyzes specific for single TFBS. Such fold changes are calculated taking into account local read depth biases between conditions, thus diffTF is insensitive to region-dependent and sequence biases. As we used the same design formula as in DESeq2, diffTF also allows easy analysis of the time course data by calculating TF-specific slope change. However, as a consequence of diffTF extensive output and flexibility in parameter and approach choices, it is quite a computationally exhaustive pipeline. It is written with the Snakemake (Köster and Rahmann, 2012) functionality and optimized for running in a cluster environment.

In summary, in this chapter we presented a brand-new computational method, that utilizes chromatin and gene expression data, calculates differential TF activity and assigns in a data-driven way the molecular mode of action to the TFs. The main goal of diffTF is to help users by integration of multiple cellular molecular phenotypes to generate provable hypotheses regarding TF regulation. Ultimately this method can improve our understanding of the gene regulation mechanisms through the TF regulatory networks.

Chapter 3

Identifying TF regulatory changes between two subtypes of CLL

In the following chapter I will describe in detail the application of `diffTF` to the CLL multiomics dataset. We will highlight relevant TFs that distinguish two subtypes of CLL and validate `diffTF` classification mode into activators and repressors. All the analyses applied to the CLL dataset were conceived by me under the supervision of Dr. Judith Zaugg. An independent CLL dataset used for validation purposes was provided by Holly Giles under the supervision of Dr. Sascha Dietrich and Dr. Wolfgang Huber (Giles et al., in preparation). I received help in the computational analysis on the different stages of this project from Dr. Christian Arnold. The text in this chapter has been originally written by myself and was taken and adapted from:

***Ivan Berest†**, Christian Arnold†, Armando Reyes-Palomares, Giovanni Palla, Kasper Dindler Rasmussen, Holly Giles, Peter-Martin Bruch, Wolfgang Huber, Sascha Dietrich, Kristian Helin & Judith B. Zaugg (2019) Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF. Cell reports, 29(10), 3147–3159.e12. doi: 10. 1016/ j. celrep. 2019. 10. 106*

3.1 Introduction

In the previous chapter we described a novel computational framework `diffTF`, that defines TF activity and provides TF activator/repressor classification. In this chapter we will mostly focus on the application of this method to the largest at a time multiomics (ATAC-seq, RNA-seq, histone marks ChIP-seq) dataset (Rendeiro et al., 2016), generated from the B-cells of the chronic lymphocytic leukemia patients, and show that `diffTF` recover known and novel transcription factors, which are different between two divergent subtypes of the B-cell CLL.

CLL is a highly heterogeneous disease, originating mostly in the lymph nodes and affecting lymphocytes, as well as lymph node microenvironment. It is one of the most common cancers in the world, especially for adults. Based on the progression of the disease and particularly on the molecular phenotypes of the cancer cells, there are two major subtypes of CLL (mutated: M-CLL and unmutated: U-CLL), which are described in detail in the research for the past 20 years (Cordoba et al., 2015; Guièze and Wu, 2015; Strati and Shanafelt, 2015). This classification is purely based on the mutation status of the IGHV locus in the B-cells, however, distinguish different epigenetic programs in the CLL cells. In general, patients with M-CLL have B-cells that proceed through normal affinity maturation and have high activity of the BCR and NF-κB signaling pathways, that results in a longer survival rates (Neu and Wilson, 2016). Patients with U-CLL do not reach the affinity maturation, potentially due to the multiple genomic aberrations (Döhner et al., 2000), and overall have shorter survival time and higher frequency of relapse after treatment (Furman et al., 2014).

Here, while using chromatin accessibility and gene expression data as input to diffTF, we discovered novel and already known in the literature TFs, associated with CLL, that are differentially active between U-CLL and M-CLL. We also predicted their mode of action with the `diffTF` classification mode, and validate these results both experimentally, using external CLL dataset treated with Ibrutinib, and *in silico*. By performing TF footprint analysis based on the ATAC-seq data we also highlight the importance of the flank regions accessibility of the TFBS in the distinguishing between activators and repressors.

3.2 Methods

3.2.1 Data sources

We used CLL multiomics data (ATAC-seq and RNA-seq) previously published in (Rendeiro et al., 2016), below referred as original CLL dataset, which was described in detail in the previous chapter in the section 2.2.1. For the `diffTF` analysis below we used HOCOMOCO v10 database with 640 human TF binding models (Kulakovskiy et al., 2015). To compare our CLL TF classification for activators and repressors we used available TF-target interactions database *TRRUST v2* (Han et al., 2018) [`http://www.grnpedia.org/trrust/`)]. For the chromatin state enrichment below we downloaded 18-state model from the *chromHMM* (Ernst and Kellis, 2012) primary B cells data (Kundaje et al., 2015).

For the independent validation of the AR classification we used paired ATAC-seq and RNA-seq dataset, later called Ibrutinib CLL dataset, from different 4 CLL patients (2 U-CLL versus 2 M-CLL), where all samples were treated with DMSO and Ibrutinib (Giles et al., in preparation). The following 3 paragraphs about generation of ATAC-seq and RNA-seq data for the Ibrutinib CLL dataset is a direct copy from (Berest et al., 2019) and were provided by Holly Giles.

Cell source of the Ibrutinib CLL dataset

"Peripheral blood was taken from 4 CLL patients (2 male, 2 female, aged between 61 and 74) and separated by Ficoll gradient (GE Healthcare), mononuclear cells were cryopreserved on liquid nitrogen. Samples were later thawed from frozen as previously described (Dietrich et al., 2018) and MACS sorted for CD19 positive cells (Milteny autoMACS®). The cells were resuspended in RPMI (GIBCO, Cat.No. 21875-034), with the addition of 2mM glutamine (GIBCO, Cat.No. 25030-24), 1% Pen/Strep (GIBCO, Cat.No. 15140-122) and 10% pooled, heat-inactivated and sterile filtered human type AB male off the clot serum (PAN Biotech, Cat.No. P40-2701, Lot.No:P-020317). 5ml of cell suspension was cultured in 6-well plates (Greiner Bio-One Cat.No. 657160). To prepare the treatments, Ibrutinib (Selleckchem, Cat.No. S2680) was dissolved in Dimethyl sulfoxide (DMSO; SERVA, Cat.No. 20385) and stored at -20°C. After thawing, Ibrutinib was prediluted in DMSO and was added to the plates. Control wells were treated with DMSO in the same concentration as with Ibrutinib treatment. In both treatment and control, the final DMSO concentration was 0.2%. Cells were incubated at 37°C and 5% CO2 for 6 hours with or without

500nM ibrutinib. The final cell concentration was 2x10 ^ 6 cells/ml. After treatment, cell viability and purity was assessed using FACS. All samples had a viability over 90% and over 95% of CD19+/CD5+/CD3- cells."

Generation of ATAC-seq libraries for the Ibrutinib CLL dataset

"ATAC-seq libraries were generated as described previously (Buenrostro et al., 2013). Cell preparation and transposition was performed according to the protocol, starting with 5x10 ^ 4 cells per sample. Purified DNA was stored at -20°C until library preparation was performed. To generate multiplexed libraries, the transposed DNA was initially amplified for 5x PCR cycles using 2.5 μL each of 25 μM PCR Primer 1 and 2.5 μL of 25 μM Barcoded PCR Primer 2 (included in the Nextera index kit, Illumina, San Diego, CA, USA), 25 μL of NEBNext High-Fidelity 2x PCR Master Mix (New England Biolabs, Boston, Massachusetts) in a total volume of 50 μL. 5 μL of the amplified DNA was used to determine the appropriate number of additional PCR cycles using qPCR. Additional number of cycles was calculated through the plotting of the linear Rn versus cycle, and corresponds to one-third of the maximum fluorescent intensity. Finally, amplification was performed on the remaining 45 μL of the PCR reaction using the optimal number of cycles determined for each library by qPCR (max. 13 cycles in total). The amplified fragments were purified with two rounds of SPRI bead clean-up (1.4x). The size distribution of the libraries was assessed on Bioanalyzer with a DNA High Sensitivity kit (Agilent Technologies, Santa Clara, CA), concentration was measured with Qubit® DNA High Sensitivity kit in Qubit® 2.0 Flurometer (Life Technologies, Carlsbad, CA). Sequencing was performed on NextSeq 500 (Illumina, San Diego, CA, USA) using 75bp paired-end sequencing, generating ~450 million paired-reads per run, with an average of 55 million reads per sample."

RNA-seq library generation for the the Ibrutinib CLL dataset

"RNA was isolated using the miRNeasy Mini Kit (QIAGEN, Cat.No. 217004), starting with 1x10 ^ 7 cells per sample. Cells were lysed in QIAzol Lysis reagent and homogenized using QIAshredder (QIAGEN, Cat.No. 79654), homogenized cell lysates were stored at -80°C until RNA extraction. RNA extraction was performed according to miRNeasy protocol and purified RNA was stored at -80°C until further processing. RNA integrity was checked using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, CA), and concentration was mea-

sured with Qubit® RNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, Carlsbad, CA). Stranded mRNA-Seq libraries were prepared from 250ng of total RNA using the Illumina TruSeq RNA Sample Preparation v2 Kit (Illumina, San Diego, CA, USA) implemented on the liquid handling robot Beckman FXP2. Obtained libraries that passed the QC step, which was assessed on the Agilent Bioanalyzer system, were pooled in equimolar amounts. 1.8 pM solution of each pool of libraries was loaded on the Illumina sequencer NextSeq 500 High output and sequenced uni-directionally, generating ~450 million reads per run, each 85 bases long."

3.2.2 Data processing

ATAC-Seq data processing

For both CLL datasets used in this chapter (original and Ibrutinib treated) we used previously described in the section 2.2.2 *in-house* ATAC-seq processing pipeline. We ran it using default parameters for the *hg19* reference genome. Output diagnostic plots showed a typical pattern for ATAC-seq data of retaining a number of reads across processing steps, with major filtering at the removing duplicates and mitochondrial reads procedures. The fragment length distributions for the CLL data showed peaks with nucleosomal length periodicity and were similar to the quality check plots of the ATAC-seq data described in the [buenrostro_2013]. As we observed for this data that the amount of reads were biased by the GC content of the region, we used GC bias removal procedure explained in detail in the section 2.2.2 to remove it from the ATAC-seq samples. Lastly, for the original CLL dataset we generated PCA plots for variables from the metadata and didn't observe separation of the samples for majority of them (batch, IGVH homology, gender, the patient age at data collection or the patient age when diagnosed), except the *IGVH mutational status* separates data clearly into U-CLL and M-CLL subgroups.

Running `diffTF` analysis

For the `diffTF` analysis of the original CLL dataset we used 52 samples (25 U-CLL and 27 M-CLL) ATAC-seq samples from 88 downloaded, see Appendix A for full dataset metadata. We filtered out samples with undefined IGVH mutational samples and took only first (ending with "_1" extension) replicate per sample in order to remove potential over representation individuals bias. First using `DiffBind` (Ross-Innes et al., 2012) we made a consensus peakset, consisting of 48065 peaks, with

the minimum overlap parameter of the unique samples defined as 5. We added 100 bp on each side of the predicted TFBS, as motif extension parameter. To calculate differential chromatin accessibility signal at the TFBS and peaks we used following design formula ("~ *batch* + *IGVH mutational status*") and fitType "*local*" parameter. *Batch* variable refers to the sequencing batch of the data and *IGVH mutational status* to the U-CLL or M-CLL metadata. Finally, we used 10 GC bins parameter in our GC binning approach, in order to remove additional GC related bias for the distribution of TFBSs.

We performed `diffTF` analysis with the same parameters as described above for the Ibrutinib CLL dataset, except changing of the design formula for the `DESeq2` analysis: "~ *treatment* + *IGVH mutational status*. *Treatment* parameter refers to the treatment of the cells with DMSO or Ibrutinib.

RNA-seq processing for CLL data and AR classification

We downloaded ten RNA-seq raw *fastq* files for the original CLL dataset with respective metadata. Similarly to the ATAC-seq processing pipeline we started with initial quality control using `FastQC` and adaptor trimming with `Trimmomatic` [internal parameters: ILLUMINACLIP:Truseq-2.fa: 1:30:4:5:true TRAILING:3 MINLEN:20]. After that we aligned cleaned RNA-seq *fastq* files to *hg19* reference genome using `STAR` algorithm (Dobin et al., 2013) [internal parameters: -outFilterMultimapNmax 2 -quantMode GeneCounts] with GENCODE (Harrow et al., 2012) v29 gene annotation. Based on the resulted amount of reads for the *bam* files and quality controls we removed 2 RNA-seq samples with low read counts (< 10 millions of reads). For the remaining RNA-seq samples we used `DESeq2` package with "~ *IGVH mutational status*" design formula to define log2 fold changes for each gene. Afterwards we filter out genes that have less than 5 reads on average in each condition, and have a median expression of 0. Based on this filtering we removed 270 not-expressed TFs from the used HOCOMOCO v10 640 human TFs. The resulting set of expressed TF gene expression values were used to define TF mode of action with `diffTF` classification mode. As default stringency percentile threshold we used 0.05/0.95 cutoff.

RNA-seq data for the Ibrutinib CLL dataset was analysed with the same tools and parameters as described above, with only exception that we used "~ *treatment* + *IGVH mutational status*" design formula for the DESeq2 analysis.

3.2.3 Biological validation of the `diffTF` analysis

GO terms related with CLL differentially active TFs

Using *org.Hs.eg.db* Bioconductor annotation package v3.8.2 we assigned ENSEMBL gene IDs for the TFs of interest from `diffTF` CLL analysis with GENCODE v29 (Frankish et al., 2019) annotation and associate them with the respective GO terms for all biological processes. To define receiver operating characteristic curves for all GO terms we used `precrec` (Saito and Rehmsmeier, 2017) v0.10.1 R package with all significant TFs (FDR < 10%) from the `diffTF` analysis on the original CLL dataset. After that we kept only GO terms with AUC > 0.6 and having not more than 90 TFs and not less than 9 TFs, data is shown in the Appendix C.

PWM similarity clusters

We grouped HOCOMOCO v10 TF motifs based on their PWM similarity using `matrix-clustering` (Castro-Mondragon et al., 2017) algorithm from the `RSAT` suite (Medina-Rivera et al., 2015). For that we downloaded PWMs for all human TFs from HOCOMOCO v10 database and input them to the web interface if the matrix-clustering, which performed the clustering with the following parameters: *ncor* = 0.4, *cor* = 0.6 and average linkage rule. Described clustering resulted in 127 PWM clusters [`https://bit.ly/2J9TaaK`], which we used to summarize the signal per cluster for CLL `diffTF` analysis.

Differential TF activity correlation with target gene expression

To predict potential target genes for TFs of interest we used *in silico* predicted TFBS obtained from `diffTF` and annotated each of them to the closest gene with *annotatePeak* function from `ChIPseeker` (Yu et al., 2015) Bioconductor (Huber et al., 2015) package using the annotation to the hg19 reference genome. Only TFBSs located within a distance of -2kb to +500 bp from the transcription start sites of the genes were used for further analysis. Using expression data for the target genes from RNA-seq of the original CLL dataset, we calculated the median log2 fold changes of the target genes per TF, using only unique target genes. For the final analysis we filter out TFs that have less than 200 and more than 1500 unique target genes.

Enrichment of chromatin states for CLL differentially active TFs

For every significant TF (adjusted p-value < 0.1) from `diffTF` analysis for the original CLL dataset we extracted unique TFBSs per TF. After that we overlapped these TFBSs with downloaded expanded 18-state model from the `chromHMM` (Ernst and Kellis, 2012) analysis of the primary B cells (Kundaje et al., 2015) with bedtools *intersect* function. Then we calculated the fraction of the binding sites overlapping each chromatin state per TF. Obtained data we subsequently grouped using predicted TF mode of action (activator or repressor) and visualized the distributions as boxplots for each chromatin state. Activator and repressor distributions of TFBS fractions were compared using Wilcoxon rank sum test.

3.2.4 TF footprinting analysis

To perform *in-house* developed TF footprinting analysis for the original CLL dataset, we first downsampled all 52 used for `diffTF` analysis samples to the sample with lowest amount of reads (∼14.5 millions of reads) and merged equal amount of samples (25 samples) into the U-CLL and M-CLL *bam* file. For all 370 expressed TFs that were used for `diffTF` classification mode, we took unique set of TFBSs, excluding TFBSs that were overlapping between all activators and all repressors, and used *dnase_to_javatreeview.py* script from the `pyDNase` software (Piper et al., 2015). By doing so, we obtained a set of base-specific Tn5 insertions matrices for all unique TFBS, importantly splitting activators and repressors, per each TF. We removed regions overlapping with hg19 blacklisted regions and huge outliers having more than 1000 counts per position. To normalize Tn5 insertions per base pair, we calculated averaged amount of reads for all consensus peaks with *featureCounts* function from the `Subread` (Liao et al., 2019) R package [internal parameters: -p -B -d 0 -D 2000 -C -Q 10 -O -s 0].

To generate comparable genomic chromatin accessibility background, we binned the consensus peaks in a 200 bp bins with a *makewindows* bedtools function and randomly selected 10000 regions for which use the same *dnase_to_javatreeview.py* script with the same parameters and normalization as described above. Using scaled by each TFBS average Tn5 insertions per bp for all expressed TFs we performed a PCA analysis and visualized comparison of the first two PC components covering 36% of the variance. To generate summarized footprint plots for each quadrant of the obtained PCA plot we used the same scaled data.

3.3 Results

3.3.1 Differentially active TFs between U-CLL and M-CLL

To gain more insights into the differences in the regulatory mechanisms between U-CLL and M-CLL, we applied `diffTF` to the large multiomics dataset containing almost equally these two subtypes of CLL, dataset described in detail in the section 3.2.1. After running `diffTF` basic mode we identified 68 differentially active TFs [FDR < 10%] between respective cancer types, see Figure 3.1A. Taking into account previously published research about TF regulatory networks of the CLL we annotated 44% (30 TFs) of them as being already known with CLL, with 80% (24 TFs) completely agreeing with the direction of the comparison (e.g. active in M-CLL or U-CLL), see Appendix B for more details of this assignment and potential functions of each TF. Independently, we also performed GO enrichment analysis of the differentially active TFs, and found that they mostly reflected the known differences in molecular regulation between M-CLL and U-CLL, such as cell-surface signaling, immune response and leukocyte differentiation GO terms. Selected enrichments are shown in the Figure 3.1B, as well as full list of enrichments in the Appendix C.

As our `diffTF` results completely rely on the TF motifs of interest and they show a lot of similarities between each other (Lambert et al., 2019), we decided to group used human TFs into the TF motif clusters based on the PWM similarities using *matrix-clustering* RSAT tool (Castro-Mondragon et al., 2017) [clustering is available here:`https://bit.ly/2J9TaaK`]. Most of the differentially active TFs from the CLL analysis aggregated into ten clusters, see top plot in the Figure 3.2A.

For the U-CLL we identified cluster 40, which contains various IRF TFs and STAT2, having the strongest TF activity compared to the M-CLL. All of these factors were previously associated with U-CLL and are part of the Toll-like receptor signalling pathway, that is known to have disrupted activity between U-CLL and M-CLL and influences the rate of cell proliferation, affect cell cycle regulation and apoptosis (Arvaniti et al., 2011; Havelange et al., 2011; Slager et al., 2013). U-CLL is known to be very fast progressive cancer, therefore proliferation rates of it are much more pronounced compared to M-CLL. Cluster 18, which is mostly active in the U-CLL, includes MYC factors, that are known to increase proliferation in the cells (Landau et al., 2015; Yeomans et al., 2016). Another known U-CLL identified TF is PAX5 which affects B-cell to plasma cell differentiation, that leads to overall decreased cell survival and poor patient prognosis (Ghamlouch et al., 2015).

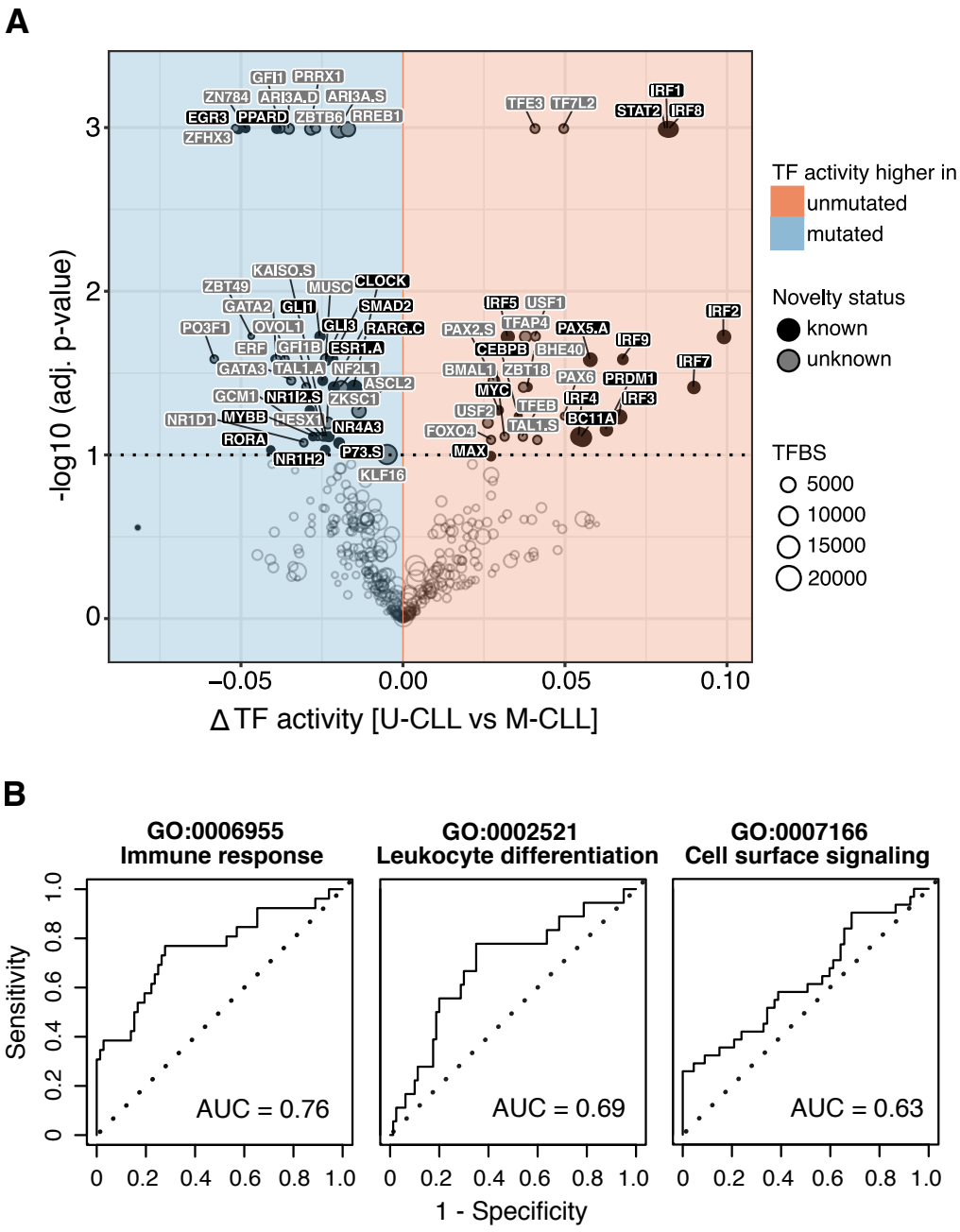

Figure 3.1: **diffTF results for the CLL dataset.** (A) Volcano plot of differential TF activity between U-CLL (n = 27 biological replicates) and M-CLL (n = 25 biological replicates). Significance threshold (10% FDR) is indicated with a dotted line. TFBS, number of predicted TFBSs. p values are obtained through `diffTF` using the empirical approach and adjusted by the Benjamini-Hochberg procedure (y axis). (G) Receive-operator characteristic (ROC) curves for three selected Gene Ontology (GO) terms with high area under the curve (AUC) based on all differentially active TFs (FDR < 10%) between U-CLL and M-CLL. See Appendix C for the full list of significant GO terms. This figure was produced by myself and was published in the (Berest et al., 2019).

In comparison to U-CLL, for M-CLL we found TFs that usually correspond to the normal functionality of B cells and usually are parts of the B-cell receptor (BCR), nuclear factor κB (NF-κB) and Wnt signaling pathways. TF clusters with the most

activity for M-CLL resemble ROR factors, which are known to activate NF-κB and Wnt pathways in CLL (Minami et al., 2010), and the GATA family of TFs, that are essential factors of self-renewal mechanism in the hematopoietic stem cells and priming towards lymphoid differentiation (Kikushige et al., 2011). Among known TFs for M-CLL we also found EGR TF cluster, which regulates BCR signaling (Damm et al., 2014) and is important for DNA methylation regulation in CLL (Oakes et al., 2016). Interestingly, we also found an active GLI1 factor, which is downstream regulator of the Hedgehog signaling pathway, supporting the known effect of this factor on the apoptosis and survival of the B cells in M-CLL (Kern et al., 2015). Recently, also identified for M-CLL with diffTF, PPARδ factor was linked to M-CLL regulation through metabolic pathways (Li et al., 2017).

Apart from 30 TFs that were already associated with CLL regulation, we also identified 38 novel TFs, that are significantly differentially active between U-CLL and M-CLL, see Figure 3.1A. It seems that there is a disrupted regulation of the circadian clock, as we have found several new TFs (BMAL1 and NR1D1) being differentially active in both conditions, as well as it was known association of CLOCK with CLL (Rana et al., 2013). In light of the recent research, suggesting that escaping of the circadian clock is an emerging hallmark of cancer (El-Athman and Relógio, 2018), such findings can lead to potential new discoveries in the CLL development. As novel differentially active TFs more active in U-CLL we can highlight BHE40 (basic helix-loop-helix family), which is a regulator of B-1a cells differentiation (Kreslavsky et al., 2017); TFAP4 that affects mitotic division (D'Annibale et al., 2014); TFE3 and TFEB factors that are known to be overexpressed in renal cancers, mainly because of the chromosomal fusions (Kauffman et al., 2014); TF7L2 is regulating MYC activity in the cancer cells (Hou et al., 2016). TFs that are active in M-CLL compared to U-CLL are mostly associated with terms relevant for the differentiation of B-cells and cancer cells maintenance, such as regulation of apoptosis by ZN784 (Kasim et al., 2017), cell-cycle progression by ZBTB6 (Chevrier et al., 2014) and B-cell lymphopoiesis regulated by ARI3A factor (Zhou et al., 2015). We also observed that GFI family members(GFI1 and GFIB) are less active in U-CLL, therefore their activation in M-CLL can affect regulation of apoptosis, that is decreased in the M-CLL B cells (Coscia et al., 2011).

In summary, the described `diffTF` results from the original CLL dataset cover much of the known biology of the differences between U-CLL and M-CLL and identifies approximately the same amount of significantly differentially active novel TFs, that have a functional relevance for the development and maintenance of CLL.

3.3.2 Molecular function of the CLL differentially active TFs

In the previous section we showed that `diffTF` is able to identify differentially active TFs between different types of CLL. However, it is important to know the molecular mode of action of the TFs, if they activate or repress target genes expression. Using known TF-target gene TRRUST database (Han et al., 2018) we were not able to classify TFs by their mode of action, as the majority of TFs were annotated in the published literature having both repressing and activating activity, see Figure 3.2C.

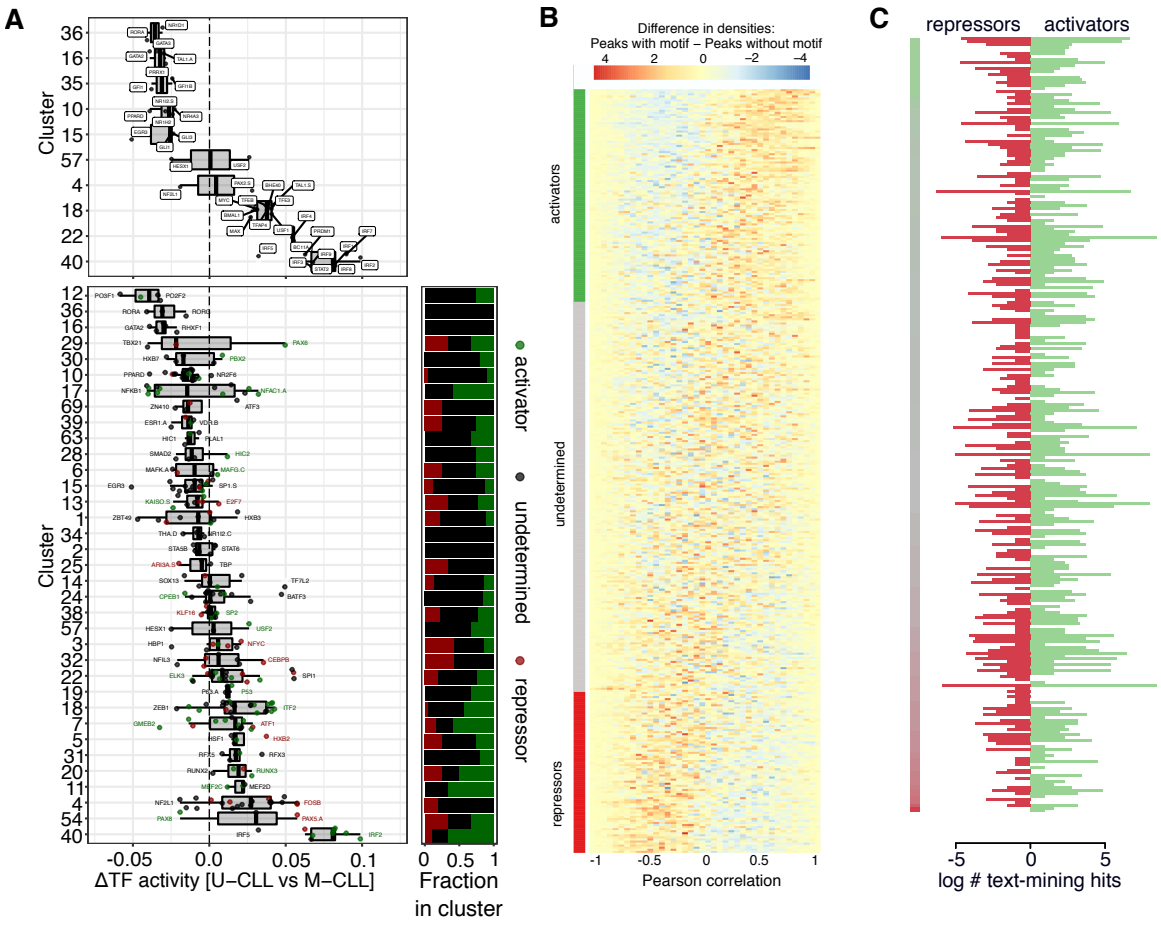

Figure 3.2: **Activator/repressor classification for the CLL dataset** (A) Boxplots for the clustering of TFs based on the similarity of their PWMs as defined in the section 3.2.3 for the differential "TF activity" between U-CLL and M-CLL. The top cluster plot is based on the significant TFs only (FDR < 10%), with all TFs being labeled. The bottom plot shows all remaining PWM clusters with at least 2 members (TFs), with the distribution of activators, repressors, and undetermined TFs displayed for each cluster at the right side. Only the most negative and most positive TFs are labeled in each cluster,and all TFs are colored by their classification. (B) Difference between foreground and background distributions for all TFs are displayed as a heatmap. Each horizontal line represents the subtraction of the binned foreground minus the binned background correlation distributions (40 bins) for one TF. TFs are sorted from strongest predicted activator to repressor. (C) Barplot of the number of studies detected by text-mining for expressed TFs in CLL whose transcriptional activity on studied target gene was classified as activator (green bars) or repressor (red). Panel A was produced by myself and panels B-C were devised by Dr. Armando Reyes-Palomares. Shown graphs are published in the (Berest et al., 2019).

We decided to use the available CLL RNA-seq data and developed a data-driven classification mode of diffTF, that integrates gene expression and chromatin accessibility data and classify TFs into activators or repressors, described in the section 2.2.4 and visualised in the Figure 2.2. The main assumption of the method is that TFs classified as activators will have increased chromatin accessibility of the target sites when highly expressed, and repressors will have decreased chromatin profiles of the target sites with increased abundance.

Upon employing `diffTF` classification mode to the original CLL dataset we classified almost 40% of the expressed TFs (146 out of 370) as supposed activator or repressor. Interestingly, when calculating distributions of devised functional classes per TF cluster based on PWM similarity, we didn't observe a single cluster with the same role (Figure 3.2A). Such observation highlights that TFs from the same TF cluster, even though might have redundant TFBS, are controlled differently on the gene expression level. The heatmap in the Figure 3.2B summarizes the differences between foreground (peaks with TFBS of specific TF) distributions of correlations between TF expression and peaks accessibility as compared to the background distributions (peaks without TFBS of specific TF) of correlations for all TFs. As this heatmap is sorted by median correlation, we observe typical enrichment of signal for activators in the top right part of the heatmap, and respective enrichment in the bottom left part for the repressors. However, more than 200 TFs were not showing significant differences between foreground and background correlation distributions and classified as undetermined. As our classification is based on only gene expression and chromatin accessibility it will be not able to assign activator or repressor role to the TFs that are regulated on the posttranslational level and have low variation of the expression level. Also if TF is acting relatively equally genome-wide as activator or repressor, it would be classified as undetermined.

3.3.3 Validations of `diffTF` classification mode

As described above we tried to validate `diffTF` classification mode using text mining on the external TF-target genes TRRUST database, but were not able to define significantly prevalent mode of action for TFs of interest. Therefore, we started with possible *in silico* validations of the method. First, based on the global assumption of the classification mode we expected to observe significant correlation between differential TF expression and differential TF activity from `diffTF` basic mode for activators, and negative correlation for the repressors respectively.

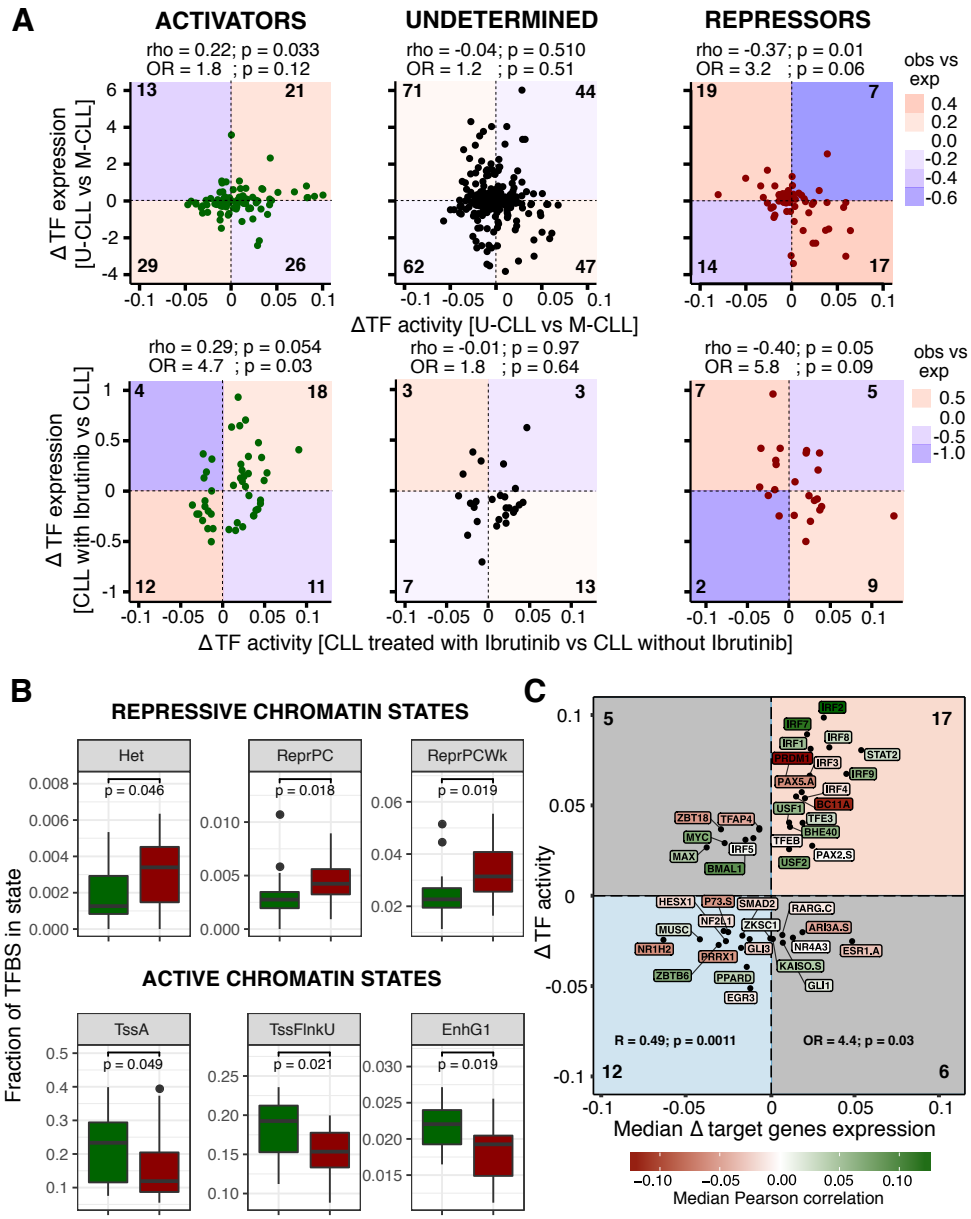

Figure 3.3: **Experimental validation for the activator/repressor classification** (A) Correlation of differential TF activity and differential gene expression for predicted activators, undetermined TFs, and predicted repressors for the comparison of U-CLL and M-CLL (top) and for CLL samples treated with ibrutinib versus control treatment with DMSO (bottom). TF classifications were obtained from the original CLL dataset. Spearman's rho and p value, as well as the odds ratio (OR) and p value of Fisher's exact test, are reported in the figure. The number (n) of TFs is indicated for each quadrant. For the bottom row, only TFs that were significant in `diffTF` in the CLL dataset (FDR < 10%) are shown. Color shadings indicate the observed versus the expected ratio for each quadrant (blue, less than expected; red, more than expected). (B) Fraction of TFBSs overlapping specific chromatin states are shown for putative activators (green) and repressors (red). Only chromatin states with significant differences between activators and repressors are displayed (Wilcoxon test, $p < 0.05$). (C) Correlation of differential TF activity (diffTF U-CLL versus M-CLL) and the differential expression of target genes (median log2 fold change U-CLL versus M-CLL) are shown. Color of TF labels represents mode-of-action class (activator, green; repressor, red) on a continuous scale based on the correlation strength (odds ratio [OR] and p value are given for Fisher's Exact test; R and corresponding p value are given for Pearson's correlation).This figure was produced by myself and is published in the (Berest et al., 2019).

Using RNA-seq data from the original CLL dataset we calculated differential TF expression for each of the expressed TF, and then compared these values with differential TF activity obtained from the `diffTF` basic mode, with stratification of the TFs based on the activator/repressor classification into 3 groups (activators, undetermined and repressors). As expected we observed positive correlation for activators (ρ=0.22; p-value=0.033), no significant correlation for undetermined class(ρ=-0.04; p-value=0.51) and negative correlation for repressors(ρ=-0.37; p-value=0.01), as shown in the Figure 3.3A top. As such validation is based on the same data that `diffTF` classification mode used for the assignment of TF function, we corroborated described CLL specific TF functional classification on the independent data. We used ATAC-seq and RNA-seq dataset generated from four patients from a separate CLL cohort treated with DMSO or Ibrutinib (data is provided by Holly Giles and described in the section 3.2.1). Afterwards, we similarly compared differential TF activity and differential TF expression between CLL samples treated with and without Ibrutinib, however using activator/repressor classification from the original CLL dataset described above. Out of 68 activators and repressors, which were expressed both in the original CLL dataset and Ibrutinib dataset, 46 TFs followed the expected relation (Figure 3.3A bottom). For remaining activators we observed even stronger positive correlation (ρ=0.29; p-value=0.054), whereas for repressors similar negative correlation (ρ=-0.4; p-value=0.05). Such results indicate that even when using independent CLL data, e.g. perturbed by Ibrutinib, differentially active TFs showed consistent functional mode of action predicted on the bigger CLL dataset.

On top of this we used known chromatin states data for primary B cells (Kundaje et al., 2015) and overlapped with it unique combined activator or repressor TFBSs. By comparing fraction of TFBSs in the state we observed that presumable repressors have more regions in the chromatin repressive states, while activators are located mostly to the active states (see Figure 3.3B).

Lastly, we decided to investigate the relationship of the differential TF activity and target genes expression. For this we overlapped TFBS with the promoter regions of the protein coding genes (-2kb/+500bp from TSS) and defined a list of target genes for each CLL differentially active TF. By comparing median expression of the target genes and TF activity we observed the expected positive correlation (R=0.49; p-value=0.0011), shown in the Figure 3.3C. Interestingly, such correlation was independent of the predicted TF mode of action, thus once more validating `diffTF` classification mode based on the chromatin accessibility and corresponding TF regulation of the target genes expression.

In summary, using *in silico* and experimental validations described above, show that `diffTF` classification mode is able to identify TFs mode of action by integrating RNA-seq and ATAC-seq data. As shown, such classification is robust to cellular perturbation by chemical drug and corroborates through the enrichment of the chromatin states and effect on the target genes expression. However, we also observed TFs that are not following the expected relation, which are potentially regulated on posttranscriptional or posttranslational levels.

3.3.4 TF footprinting analysis reveals activator/repressor patterns

After the observation from the validation with chromatin states that repressors are located in the repressive chromatin, we decided to perform TF footprinting analysis based on ATAC-seq data. We selected the well-known activator and repressor from the expressed in CLL TFs, REST and STAT2 respectively. Interestingly, we observed clear differences in their footprints (normalized amount of Tn5 insertions per bp) between them, see Figure 3.4B. For the repressor TF REST, we observed increased accessibility directly in the motif with overall low accessibility around the binding site (lower than genome-wide average accessibility). STAT2 footprint showed the opposite picture, with maximum accessibility outside the binding site (+/-25bp) and slow decrease of accessibility to the genome-wide average over +/-100bp from the center of motif. These findings provide additional validation to the `diffTF` classification mode, showing that despite the local increase of accessibility in the repressor motif center the surrounding chromatin for repressors are highly compact.

When combining footprints from all expressed TFs for the original CLL dataset with PCA analysis we observe that PC1 resolve accessibility differences in the motif center and PC2 explain accessibility differences on the surroundings of the binding sites (see Figure 3.4A). Based on these 2 components we define four TF footprint classes: class I and II contain a lot of predicted activators (53 out of 89), and class III and IV mostly have repressors (37 out of 57). Typical I and II class activator footprint for IRF2 showed increase of accessibility on the surroundings and decrease in the center of the motif, both in M-CLL and U-CLL, see Figure 3.4C top. PAX5 footprint on the other hand summarizes class III and IV, with increased accessibility in the center and decreased flank accessibility (Figure 3.4C bottom). We observed similar shapes and trends of the chromatin accessibility when plotting footprints for all predicted with `diffTF` classification mode activators and repressors (Figure 3.4D).

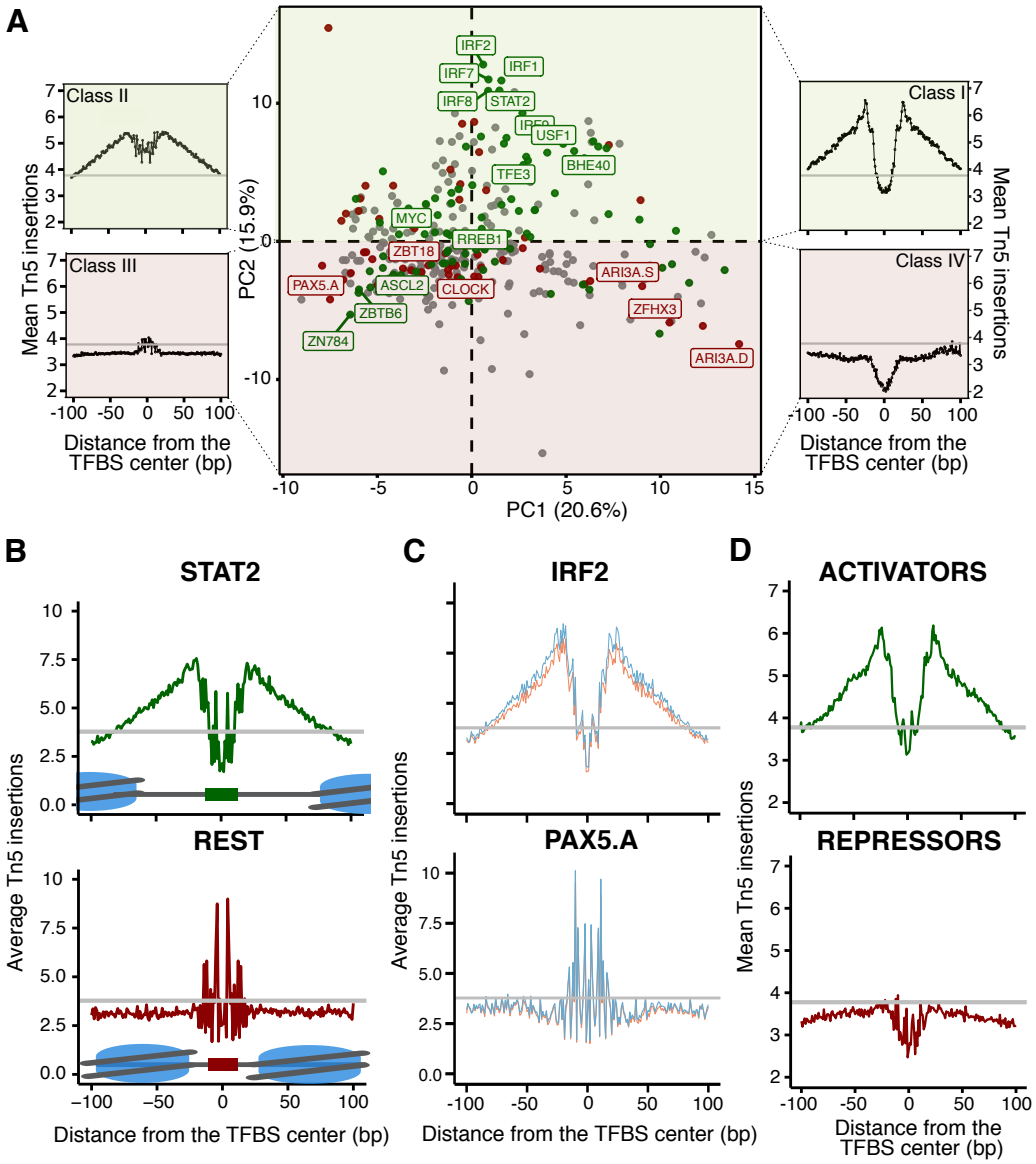

Figure 3.4: **TF footprinting analysis for the CLL data** (A) Scatterplot comparing PC1 and PC2 from a PCA of the footprints of all expressed TFs (n = 370). The insets (classes I–IV) display the average footprint across TFs in that quadrant. (B) Exemplary footprints (mean Tn5 insertions centered at TFBS) for a well-known activator (STAT2, top) and a well-known repressor (REST, bottom). Tn5 insertions were normalized to the library size and numbers of samples between U-CLL and M-CLL. The genome-wide average of insertions within accessible chromatin is shown as a solid gray line. (C) Footprint analysis for one of the strongest putative activators (IRF2, right) and one of the strongest putative repressors (PAX5.A, left). Footprints are shown separately for M-CLL (blue) and U-CLL (orange) based on the normalized number of Tn5 insertions. (D) Average footprints for all significant (FDR < 10%) activators (top) and repressors (bottom) are shown. This figure was produced by myself and is published in (Berest et al., 2019).

3.4 Discussion

In the previous chapter we discussed technical robustness of `diffTF` and advantages in the comparison to the current state-of-the-art methods of defining sample/cell TF activity. Applying our method on the large multiomics CLL dataset we were able to find 68 differentially active TFs between U-CLL and M-CLL. As discussed before, these subtypes of CLL have substantial epigenetic differences, that we were able to recapitulate by summarizing our TF activity signal. Overall, by checking in the published research, we observed that U-CLL had more active TFs that are responsible for the repetitive genomic aberrations, high cell proliferation, disrupted cell cycle and apoptosis. As for the M-CLL TFs that are downstream targets of BCR and Wnt signaling were more active. Such activity results in the less severe damage to the B-cells, decreased apoptosis and longer survival of the cells.

As our method is purely based on the TFBSs, which were predicted using PWM binding models, there could be a lot of redundancy between the TFs that have very similar PWM. Due to the constant variation of TF binding it is very difficult to generate an evaluation golden set of TFBS (Jayaram et al., 2016), that are experimentally validated for each sample for each TF specifically. Knowing this limitation of the upstream data used in `diffTF` we checked the consistency of signal between the TFs that have very similar PWM. As expected most of the differentially active TFs (Figure 3.2A top) clustered together and showed similar TF activity, as well as high level of common TFBS for each PWM cluster. However, by clustering TF activities from all TFs we observed that some clusters (cluster 29,17,7) had very different signals even in one cluster (Figure 3.2A bottom). Such graphs are important to have to judge realistically TF activity changes between two conditions. As `diffTF` provides log2 fold change value for each TFBS, this information can be used to leverage out the effect of other TFs by removing regions that were predicted to be TFBS of the different TFs from the same PWM cluster.

Using TF expression data from RNA-seq we were able to classify TFs into activators and repressors. This classification is based on the correlations of the peaks chromatin accessibility signal containing TFBS of interest and TF expression. Summarizing such correlations and comparing to the background correlations without TFBS of interest we assign global genome-wide mode of action for the particular TF. Importantly, each TF has both negative and positive correlations, therefore, might have activating and repressing impact on the gene expression depending on the region and chromatin environment, but we define prevalent mode based on majority of

the signals. Using independent CLL data we demonstrated that `diffTF` classification mode can correctly predict the directionality of the TF abundance impact on the TF activity, even upon stimulation with Ibrutinib. Almost 70% of differentially active TFs from the original CLL dataset showed a similar relation between TF expression and TF activity, as shown in the Figure 3.3A bottom. Such results suggest that even while the TF regulatory network changed in the B-cells due to the treatment, the global patterns of TF functional modes did not change significantly. We also observed that TFBS of the classified activators are mostly located in the active chromatin state, whereas repressor TFBS are significantly enriched in the repressive chromatin (Figure 3.3B). Also TF activity was highly correlated with changes in expression of the target TF genes, predicted by having at least one TFBS in the promoter region, showing significant effect of the chromating changes on the gene expression of the target genes.

Nevertheless, as a consequence of the methodological assumptions that we made for classification mode algorithm we cannot classify TFs mode of action, if they are regulated posttranslationally or RNA-seq data has low variation between samples (correlations of gene expression and TF activity are not relevant). Also, because of summarizing, `diffTF` is also not able to classify TFs separately that can act equally as activators or repressors, however such observation can be made by looking at the correlation plot of individual TFs. As discussed above quantified TF activities are biased by the PWM similarity of the TFs, however such bias is not occurring for the TFs in the gene expression data. TFs with similar TF activity from the same PWM cluster can be expressed differently, e.g. PRDM1 and IRF PWM cluster 40, and classified with different functional roles (PRDM1 - repressor; IRFs - activators). Such cases should be taken into account and validated separately using biochemical assays. Another way to handle this situation is to calculate TF activity for each TF from the same PWM cluster using unique TFBS.

Finally, we performed TF footprinting analysis on top of the `diffTF` classification mode and found out that TFs that were classified as activators are having opened chromatin not directly at the binding site, but rather in the flanking regions. Summarizing footprints from all expressed TFs for CLL dataset we observe that the biggest sources of variation for the footprints are coming from the differences in accessibility directly in the binding site and flanking regions, see PC1 and PC2 from Figure 3.4A. TFs classified as repressors, showed low chromatin accessibility in the flanking regions with more pronounced changes directly in the motif center. We hypothesize that activators, which are located in the active chromatin, have

increase in the flanking regions due to the binding of other TFs or DNA transcriptional machinery, whereas repressors have more pinpointed binding to the closed chromatin. Repressors can be part of the silencers, which function for fine-tuning gene regulation together with enhancers was described before using histone marks data (Huang et al., 2019; Wang et al., 2012). Linking chromating TF footprinting analysis and gene expression can provide more details about TF regulation of gene expression and need further investigations in the future.

While we only uncover one layer of epigenetic differences with TFs between U-CLL and M-CLL, we were able to recapitulate known CLL biology, and generate new exciting hypothesis of the TF regulation in each CLL subtype. Apart from that we found out that TF functional role is not changing significantly upon chemical perturbation and uncover new observations about chromatin microenvironment of the activators and repressors. Overall, identified new TFs and their predicted functional role can lead to development novel precise treatments of the CLL.

Chapter 4

TF-mediated regulatory effects on TET2 across hematopoiesis

In the current chapter I will present recent findings about the changes in TF regulatory dynamics in the hematopoietic differentiation upon deletion of the methylcytosine dioxygenase TET2. Majority of the bioinformatical analysis was made by me under the supervision of Dr. Judith Zaugg and Dr. Kasper Dindler Rasmussen. Data was generated from the experiments performed by Dr. Kasper Dindler Rasmussen in the lab of Prof.Dr. Kristian Helin. The text in this chapter has been originally written by myself or was taken and adapted from:

*Kasper Dindler Rasmussen†, **Ivan Berest†**, Sandra Kessler, Koutarou Nishimura, Lucia Simon-Carrasco, George S. Vassilou, Marianne T. Pedersen, Jesper Christensen, Judith B. Zaugg & Kristian Helin (2019) TET2 binding to enhancers facilitates transcription factor recruitment in hematopoietic cells. Genome research, 29(4), 564-575. doi: 10.1101/gr.239277.118*

4.1 Introduction

Over the last twenty years, a lot of research in the hematology field revealed association of the mutational status of the ten-eleven-translocation (TET) genes with the development of various hematological malignancies (Scourzic et al., 2015). TET enzymes start the process of demethylation of 5-methylcytosines and, overall, regulate the DNA methylation turnover in the cell together with DNMT enzymes (Lio et al., 2019; Rasmussen and Helin, 2016; Ross and Bogdanovic, 2019). TET triple-knockout cells showed an increase in the global DNA methylation level, especially in the enhancer regions (Dawlaty et al., 2014; Lu et al., 2014). Despite the impact of TET1 function on the chromatin were shown (Gu et al., 2018), studies are only starting to investigate the effect of TET2 on the epigenetic regulation through the knockout systems. This is particularly interesting, because TET2 mutations are the most frequent genomic aberrations associated with myeloid cancers, such as AML (Delhommeau et al., 2009; Weissmann et al., 2011).

Surprisingly, in comparison to the TET1 and TET3 enzymes, TET2 protein does not have direct DNA binding domain. It was demonstrated that TET2 DNA binding can be facilitated through the zinc finger CxxC proteins, which are recepting non-methylated DNA and attracting chromatin modifiers to the CpG islands (Long et al., 2013), or with the help of specific TFs, such as PU.1, CEBPα and KLF4 (Lio et al., 2016; Sardina et al., 2018). Eventhough, such results suggested the potential binding of TET2 through described TFs, due to the lack of the high quality specific TET2 ChIP-seq data, the accurate prediction of the impacted by TET2 TFs was not possible till now.

In this chapter we discuss obtained via crosslinking of C-terminus of TET2 and epitope tag (V5 or FLAG) novel set of the TET2 binding regions, show enrichment of TET2 binding in the enhancers, and importantly, define multiple TFs and TF families using `diffTF` that are regulated directly and indirectly by TET2. We obtain such results not only from the samples from the different stages of steady-state hematopoiesis (ES, MPP, GMP), but also from the mouse AML model with and without TET2 knockout. Apart of that, we were able to benchmark `diffTF` on the dataset with small number of sample (n=8), recapitulate validation of the `diffTF` classification mode under the genetic perturbation and show similar principles of the chromatin microenvironment dynamics for activators and repressors in mice with footprinting analysis.

4.2 Methods

4.2.1 Data sources

For this chapter we used paired ATAC-seq and RNA-seq data (4 samples each) from mouse ES cells and three hematopoietic cell types (MPP, GMP and AML) from wild-type and upon TET2 functional knockout. Also, for the ES and AML conditions TET2 ChIP-seq was performed in the wild type and TET2 knockout. For the `diffTF` specific analyses we used HOCOMOCO v10 database with 421 mouse TF binding models combined with HT-SELEX and Methyl-SELEX data from the previously published dataset (Yin et al., 2017). These data and analyses are in detail described in the following publication (Rasmussen et al., 2019). Following methods sections about generation of ES cell lines, mouse model, FACS cell sorting and ATAC-seq,RNA-seq,TET2 ChIP-seq libraries generation are direct copies from (Rasmussen et al., 2019) and (Berest et al., 2019) methods and was provided by Dr. Kasper Dindler Rasmussen.

ES cell lines

"Mouse ES cell lines were derived from blastocysts harvested from $Tet2^{fl/fl}$ animals (Moran-Crusio et al., 2011). To generate a clean *Tet2* knockout ES cell line (to use as ChIP control), cells were transiently transfected with a *Cre* recombinase-expressing plasmid and subcloned to identify a constitutive *Tet2* knockout ES cell line. Endogenous tagging of TET2 was performed using CRISPR homology-directed repair with a single-stranded DNA oligonucleotide repair template. Briefly, a sgRNA targeting the insertion site was cloned into pSpCas9(BB)-2A-GFP (PX458) (Addgene #48138). This plasmid was cotransfected in ES cells with oligonucleotides encoding either two copies of Flag (DYKDDDDKDYKDDDDK) or a single copy of V5 tag (GKPIPNPL LGLDST) as well as homology arms (60 bp each). Transfected cells were single-cell sorted, and the resulting clones were screened for stable expression of epitope-tagged TET2 by Western blot. All mouse ES cell lines were cultured in feeder-free, gelatinized plates in "Serum/2i/LIF" conditions: Dulbecco's Modified Eagle Medium (DMEM) supplemented with 15% fetal bovine serum, 2mM Glutamax (Gibco), 0.1mM 2-Mercaptoethanol (Gibco), 1× nonessential amino acids (Gibco), 1×Pen/Strep (Gibco), 3 μM GSK3 inhibitor (CHIR99021), 1 μM MEK1 inhibitor (PD0325901), and leukemia inhibitory factor (LIF)."

Mice

"For analysis of aging-related *Tet2*-deficient hematopoiesis, cohorts of age-matched litter mates (8-weeks old) of $Tet2^{fl/fl}$ or $Tet2^{fl/fl}; Mx1-Cre^{+}$ mice (Quivoron et al., 2011) were injected three times intraperitoneally with 250 μg polyinosinic-polycytidylic acid (PolyI:C LMW, InvivoGen) at experimental days 0, 2, and 4. The mice were subsequently allowed to age and sacrificed at 10 month of age. To generate a *Tet2* knockout AML model, the following genetically modified mouse lines $Npm1^{cA-Flox}$ (Vassiliou et al., 2011), *Flt3-ITD* (Lee et al., 2007), and $Tet2^{fl/fl}; Mx1-Cre^{+}$ were intercrossed and $Npm1^{+/cA-Flox}; Fl3^{+/ITD}; Mx1-Cre^{+}$ or $Npm1^{+/cA}; Fl3^{+/ITD}; Tet2^{fl/fl}; Mx1-Cre^{+}$ mice were monitored for disease development. AML cells were harvested and $2.5Ö10^{4}$ $c-Kit^{+}Gr1^{-}Mac1^{-}$ AML splenocytes transplanted into sublethally irradiated (650Rad) Ly5.1 recipient animals by tail vein injection. Ly5.1 mice were maintained on medicated water (Ciprofloxacin 100 μg/mL) for 3 week following the irradiation procedure. To establish in vitro culture of AML cells, $c-Kit^{+}Gr1^{-}Mac1^{-}$ splenocytes har- vested from moribund mice were purified by FACS and cultured in suspension in nontissue culture treated plasticware in StemPro-34 SFM media (Thermo Fisher Scientific) supplemented with 2 mM GlutaMAX (Gibco), 1× Pen/Strep (Gibco), 0.1 mM 2-mercaptoethanol (Sigma-Aldrich), as well as the cytokines SCF (50 ng/mL), IL3 (10 ng/mL), and IL6 (10 ng/mL) (Peprotech). All animal work was carried out in compliance with ethical regulation under license by the Danish regulatory authority."

FACS sorting step for HSC cell types

"Single-cell suspensions of mouse bone marrow were erythrolysed, enriched for Kit expression (CD117 microbeads, Miltenyi Biotech) and stained with antibodies against surface markers: Lineage (B220-PECy5 (RA3-6B2, eBioscience), CD11b-PECy5 (M1/70, eBioscience), Ter119, PECy5 (TER-119, eBioscience), CD3e-PECy5 (145-2C11, eBioscience), Gr1-PECy5 (RB6-8C5, eBioscience)), Sca1-BV421 (D7, BD biosciences), cKit-AlexaFlour 780 (2B8, eBioscience), CD150-APC (TC15-12F12.2, Biolegend), CD48-PE (HM48-1, eBiocience), CD16/32-PECy7 (93, eBioscience). The following combination of surface markers was used to define hematopoietic progenitor populations: MPP cells - $Lin^{-}cKit^{+}Sca1^{+}CD150^{-}CD48^{+}$; GMP cells - $Lin^{-}cKit^{+}Sca1^{-}CD16/32^{+}$. AML cells cultured *in vitro* were harvested and stained with antibodies against the surface markers: CD11bPE (M1/70, eBioscience), Gr1-AlexaFlour 700 (RB6-8C5, eBioscience), CD16/32-PECy7 (93, eBioscience), CD34-

FITC (RAM34, eBioscience). The following combination of surface markers was used to define the leukemic precursor population purified for ATAC-seq analysis: $CD11b^{-}Gr1^{-}CD16/32^{int}CD34^{+}$. Total live bone marrow cells were stained with CD317-FITC (PDCA-1, eBioscience) and B220-APC (RA3-6B2, eBioscience) to enumerated plasmacytoid dendritic cells (pDCs). Cells were sorted on FACSAria III (BD Biosciences) and analyzed using the FlowJo software (Tree Star inc.)."

ATAC-seq libraries generation

"ATAC-Seq libraries were generated as described previously (Buenrostro et al., 2013; Lara-Astiaso et al., 2014), with the following modifications. Briefly, 10.000 freshly isolated cells (MPP, GMP, AML and ES) from individual wild-type mice were sorted into ice-cold FACS buffer (PBS + 2%FBS). The cells were pelleted using a swinging bucket centrifuge (500 x g, 10min, 4°C) with settings for low acceleration/deceleration and washed once in ice-cold PBS. The cell pellets were resuspended in 50 muL lysis buffer (10mM $Tris-HCl$ pH 7.4, 10mM $NaCl$, 3mM $MgCl_2$, 0.1% Igepal CA-630) by gentle pipetting and immediately centrifuged one additional time (500 x g, 10min, 4°C). The supernatant was discarded and the pellet containing released nuclei were resuspended gently in 25 muL 1xTD buffer containing 1.25 muL Tn5 transposase (Nextera sample preparation kit, Illumina). The transposition reaction was allowed to proceed for 45min at 37°C whereafter DNA fragments were isolated using MinElute PCR purification columns (QIAGEN) according to manufacturer's instructions.

To generate multiplex libraries, the transposed DNA were initially amplified for 5x PCR cycles using 2.5 muL each of dual-index primers (Nextera index kit, Illumina) and 2.5 muL PCR primer cocktail (PPC, Illumina) in a 25 muL reaction volume of 1x KAPA HiFi hot-start ready-mix (Kapa BioSystems). The hot-start polymerase was activated prior to adding to the reaction mix by performing a brief pre-incubation step of 3min at 95°C. The amplified fragments were size-selected with AMPure XP beads (0.5X) to remove fragments larger than 600bp and an aliquot was quantified to determine the optimal PCR cycle number to obtain 1/3 of maximum fluorescence intensity (Library quantification kit, Kapa Biosystems). Finally, PCR amplification was performed using the optimal number of cycles determined for each library (max. 18 cycles in total), size-selected with AMPure XP beads (0.5X) and eluted in resuspension buffer (Illumina). The size distribution of the libraries was evaluated on Bioanalyzer (Agilent) and sequenced on NextSeq 500 (Illumina) using 75bp paired-end sequencing with an average of 25 million reads per sample."

RNA-seq libraries generation

"A proportion of total RNA (2 ng) isolated using RNeasy Micro Kit (Qiagen) from four biological replicates (individual mice with wildtype or Tet2-deficient hematopoiesis) was amplified and size-selected with the Ovation RNA amplification system v2 (NuGen, Cat: 7102) and sequencing adaptors were added to the resulting cDNA using the Ovation Ultra Low v2 system (NuGen) according to manufacturer´s instructions. RNA-seq libraries were quality checked using a DNA 1000 Kit (Agilent) and sequenced using an Illumina NextSeq 550 instrument (75bp single-end)."

ChIP-seq libraries generation

"ES cells or in vitro cultured hematopoietic cells were washed twice in ice-cold PBS and pelleted by centrifugation. The cells were resuspended in 10 ml ice-cold PBS and freshly prepared 0.25M disuccinimidyl glutarate (DSG) stock solution (dissolved in DMSO) to obtain a final concentration of 2mM DSG in PBS (ThermoFisher Scientific) and incubated at a rotating wheel for 30 min to allow equilibration to room temperature. Then, formaldehyde (Sigma) was added to obtain a final concentration of 1% and rotated for another 10 min at room temperature. Finally, the crosslinking reaction was stopped by addition of glycine to a final concentration of 125mM. The cells were spun down for 5 min at 350 x g at room temperature and washed twice with ice-cold PBS. The cells were then resuspended in 5 mL SDS Buffer (50mM $Tris - HCl$ pH 8.1, 100mM $NaCl$, 5mM EDTA, 0.5% SDS) containing 1 mM Phenylmethylsulfonyl fluoride (PMSF) and allowed to rotate for another 5 min. Finally, chromatin was pelleted and resuspended in IP buffer (100mM $Tris - HCl$ pH 8.6, 100mM $NaCl$, 5mM EDTA, 0.3% SDS, 1.7% TritonX-100) with proteinase inhibitors according to pellet size. Chromatin dissolved in IP buffer was sheared to an average size of 200-500 bp DNA fragments in a Bioruptor (Diagenode). The sonicated chromatin was diluted in SDS-free IP buffer to achieve a concentration of 0.1% SDS, spun down at 20,000 x g for 20 min to remove insoluble chromatin fraction and precleared with protein G Sepharose beads (GE healthcare) prior to immunoprecipitation.

For immunoprecipitation of endogenous TET2, 1μg affinity-purified rabbit polyclonal antibody raised against N-terminal TET2 protein (TET2-N) (as described above) was incubated with 300μg chromatin (measured by Bradford assay) overnight. The chromatin-antibody complexes were captured in a 3h incubation with protein-G Sepharose beads (GE healthcare). For immunoprecipitation of 2xFL-TET2, 20μl of

anti-FLAG M2 affinity gel (Sigma, A2220) was incubated with 300ug chromatin for 3h. Washes of chromatin-antibody-bead complexes were performed as follows: Three washes with ice-cold 150mM wash buffer (20mM $Tris-HCl$ pH 8.0, 150mM $NaCl$, 2mM EDTA, 0.1% SDS, 1% Triton X-100), two washes with ice-cold 500mM wash buffer (20mM Tris-HCl pH 8.0, 500mM NaCl, 2mM EDTA, 0.1% SDS, 1% Triton X-100), and one wash with ice-cold IP buffer with a final concentration of 0.1% SDS. After the last wash, DNA from TET2-N immunoprecipitations was de-crosslinked by overnight incubation at 65^oC in decrosslinking solution (1% SDS, 0.1M $NaHCO_3$). In FLAG M2 immunoprecipitations, IP'ed chromatin was initially eluted by three consecutive incubations (each 20 min on ice) in elution buffer (20mM $Tris-HCl$ pH 8.0, 150mM $NaCl$, 2mM EDTA) with 0.5mg/ml FLAG peptide (DYKDDDDK, Peptide 2.0). The eluted fractions were pooled and de-crosslinked by overnight incubation at 65^oC in decrosslinking buffer. IP'ed DNA was purified using the QIAquick PCR purification kit (Qiagen) according to manufacturer's instructions.

ChIP-seq libraries for Illumina sequencing was prepared using the NEBNext Ultra II DNA library preparation kit (New England Biolabs) using an input of 1-3ng of IP'ed DNA (quantified using DNA HS assay kit (Qubit)) following the manufacturer's instructions. Adaptor-ligated fragments were size-selected using AMPure XP beads (Beckman Coulter) to retain inserts of approximately 200bp prior to PCR amplification. Equimolar amounts of sample, with compatible indexes, were pooled and sequenced on Illumina NextSeq 550 (75bp single-end)."

4.2.2 Data processing

ATAC-seq processing

For the GMP, MPP, AML and ES both for wild-type and TET2 knockout ATAC-seq samples (each group n=4) we used described in the section 2.2.2 our custom ATAC-seq processing pipeline with default parameters using *mm10* reference genome. Overall, about 60-70% of the reads were removed through different steps of this pipeline. Majority of the removed reads were classified as optical PCR duplicates or were aligned to the mitochondrial genome. We observed typical for ATAC-seq data fragment length distributions with nucleosome size periodical peaks (∼140 bp). We also performed GC bias correction already on the *bam* files level to remove technical sequencing GC bias.

RNA-seq processing

After sequencing, we trimmed adaptors from the obtained RNA-seq *fastq* files and checked the quality of them using `FastQC`. Passing quality checks samples were aligned to the *mm10* reference mouse genome using `STAR` (Dobin et al., 2013) aligner with default parameters. We assigned transcripts to genes using GENCODE v7 gene annotation and calculated differential gene expression with DESeq2. For each `DESeq2` analysis we used four biological replicates for each condition. Finally, for some visualization purposes gene counts were normalized by variance stabilizing transformation (VST).

TET2 ChIP-seq processing

ChIP-seq single-end raw sequencing data for ES and AML cells were trimmed for adapters with `Trimmomatic` and mapped to the *mm10* mouse genome with `Bowtie2` (Langmead and Salzberg, 2012) using –very-sensitive preset parameters. Then we marked and removed duplicates using `PicardTools` *markDuplicates* function. All these steps are included in the Snakemake *in-house* ChIP-seq processing pipeline.

For the ES cells we defined a confident set of TET2 binding sites from ChIP-seq data by performing differential binding analysis between wild-type and TET2 knockout samples. First, we used `MACS2` (Zhang et al., 2008) peal caller to define significant (p-value < 0.05) peaks for each ChIP-seq sample with either TET2-N or FLAG M2 antibodies compared to the samples treated only with IgG antibody. Afterwards we used `DiffBind` Bioconductor package to assemble consensus peakset

merging all peaks from rather TET2-N or FLAG M2 biological replicates with minoverlap = 1 and removing peaks that has less than 5 reads on average. Resulted counts for consensus peaksets were used as input to the `DESeq2` workflow. Finally, we identified for ES cells differentially bound TET2 binding sites (p-value < 0.1 and minimum log2 fold change > 0.5) resulting in the 19466 peaks for samples with TET2-N antibody and 15308 peaks for FLAG M2 ChIP-seq samples.

To identify differentially bound ChIP-seq regions corresponding for TET2 binding in the AML cells, we used slightly lenient cutoffs in the MACS2 peak calling. We called peaks with standard MACS2 parameters, adjusting only q-value threshold as 0.05 comparing wild type cells versus cells lacking TET2 upon functional knockout. Afterwards, we overlapped all peaks across biological duplicates experiments and merged the union of peaks overlapped at least by 1 bp. Using such procedure we identified 7002 potential TET2 binding regions in the bulk AML cells. Potential reason why we detect fewer peaks compared to the ES cells, is that we had a lower signal-to-noise ratio for the AML ChIP-seq samples.

`diffTF` analysis

For the listed comparisons (TET2 KO versus WT; GMP versus MPP) we used `diffTF` analysis with default parameters, always comparing 4 against 4 samples. We required minimum 2 samples to have a common peak to be included in the consensus peakset. Overall the amount of opened chromatin peaks for different conditions varied from 55098 peaks to 120765 peaks (GMP versus MPP comparison = 77678 peaks), later labelled with "all" extension. For the TET2 knockout versus wild type comparisons we splitted consensus peakset on several groups and for each of the resulted peaks ran `diffTF` analysis: peaks overlapping with -1.5kb/+0.5kb from the TSS of the genes - labeled "Pro"; peaks overlapping +100kb/-100kb from the TSS excluding promoter regions - labeled "ProDist"; peaks overlapped with TET2 ChIP-seq binding sites for ES and AML cells - labelled "CHIP". To obtain p-values for the predicted TF activities we used an analytical approach in diffTF, as discussed in the subsection "Estimation of significance for differential TF activity" in the section 2.2.3. To integrate ATAC-seq and RNA-seq data for the GMP versus MPP comparison we used only 268 expressed TFs (minimum 5 reads in average for each condition and mean expression counts bigger than 0).

Defining TF clusters based on PWM similarity

Overall we proceed with a similar procedure, as for the CLL dataset, described in detail in the section 3.2.3. For this project we used modified mouse HOCOMOCO v10 TF binding models. We replaced retracted from this database CDX4 and EVX1 PWM motifs with 6 PWMs for CDX4 and EVX1 (labelled later as CDX4HT, CDX4MET1, CDX4MET2, EVX1HT, EVX1MET1, EVX1MET2) from the methyl-SELEX study (Yin et al., 2017). After that we clustered modified PWM dataset with `matrix-clustering` (Castro-Mondragon et al., 2017) tool from `RSAT` suite [internal parameters: -lth Ncor 0.4 -lth cor 0.6] and obtained final 88 PWM clusters for the downstream analyses.

TF footprinting analysis

We performed TF footprinting analysis, similar as in the previous chapter for the CLL data, for the 8 ATAC-seq samples from GMP and MPP cell types without TET2 knockout. First, we downsampled each of these *bam* files to the lowest one (∼8.3 millions of reads) with the following merging by the condition. For all expressed differentially active TFs (adjusted p-value < 0.05) from the GMP versus MPP `diffTF` analysis we used *dnase_to_javatreeview.py* script from the `pyDNase` package (Piper et al., 2015) to get for each base pair in 200 bp range Tn5 insertions for all TFBS. For normalization we calculated amount of the reads per each cell type in the consensus peakset with *featureCounts* function from Subread R package with the same parameters as defined for CLL TF footprinting. Afterwards we scaled the obtained data for each TF by dividing by the mean average Tn5 insertions in the whole matrix. After PCA analysis on the scaled data we plotted first 2 PC components (comprising 40% of the global variance). To define "gray" line on the footprint plots we binned consensus peakset in 200 bp bins with *makeWindows* function from samtools and randomly selected 10000 regions, which we used in the same footprinting scheme as real TFBS. Afterwards we summarized the values of scaled averaged Tn5 insertions across 200 bp into one value, which we plotted on the shown plots, and referred later as background chromatin openness. For each quadrant of the PCA plot, we generated separate footprint plots summarising data across different TFs. For that we divided final Tn5 insertions for each base pair for each TF by the mean of the Tn5 insertions in the whole matrix of the TFBS in the one quadrant. This is done just for the visual representation and never was used for further statistical analysis. In such a way we highlight stronger differences between footprint signal in the center of the motif and its surroundings.

4.3 Results

4.3.1 TET2 binding overlaps with enhancer regions

One of the key limitations of the TET2 biology is to obtain a set of genomic regions where it can bind. Such TET2 binding is hypothetically facilitated by chromatin modifiers or transcription factors, due to the lack of the DNA binding domain. In this study we generated a high-quality genome-wide map of TET2 occupancy in myeloid, AML and ES cells. Such ChIP-seq dataset in ES cells consists of data from experiments with antibodies (1) to the endogenous TET2 (TET2-N) and (2) FLAG-tagged TET2 (FLAG M2), which was expressed endogenously. Obtained data showed a high level of correlation between biological replicates (Figure D.1 A-B). To define a confident set of TET2 binding sites we took peaks that were differentially bound either over TET2 KO cells or cells without FLAG-tagged TET2. Overall, we were able to report 26512 differentially bound regions in both conditions, of which 8262 TET2 binding sites were present in experiments with both antibodies (Figure 4.1 A; Figure D.1 E-F). Out of the high-confident set of TET2 peaks, 90% of them are associated with DNase hypersensitive sites (open chromatin accessibility peaks) and H3K27ac histone mark signal, and almost half of them overlap with EP300 binding regions (Figure 4.1 B). Using *repitools* R package we also found that ES TET2 binding sites have a low percentage of CpG dinucleotide compared to the CpG islands in mice (Figure D.1 G).

We downloaded previously published for ES cells dataset of 5-methylcytosine and 5-hydroxymethylcytosine changes under the TET2 knockout (Hon et al., 2014), and observed increase of DNA methylation signal for ∼80% and loss of hydroxymethylation for ∼65% of the high confident TET2 binding sites upon TET2 loss. Interestingly, we didn't observe significant enrichment of our TET2 binding regions in the differentially methylated regions from the same methylation/hydroxymethylation data (more than 180000 regions). These results suggest that, even though TET2 binding leads to active DNA demethylation through the transition to hydroxymethylation, such binding cannot predict differential signal alone.

Apart from the ChIP-seq experiments in the ES cells, we also defined a set of TET2 binding sites in the myeloid cells from mice immortalized with AML1-ETO and with inducible promoter of TET2 (Rasmussen et al., 2015). For the ChIP-seq biological duplicates samples with TET2-N antibody we identified 19706 significantly differentially bound TET2 regions compared to the TET2 knockout control (Figure D.1 C). Despite distribution of TET2 binding sites showed enrichment in the promoter

distal regions, only 7.4% of these regions overlapped with TET2 binding sites from ES cells. By comparing GO enrichments of ES and myeloid specific TF binding using GREAT enrichment tool (McLean et al., 2010) we observed a lot of immune system GO terms for myeloid TET2 regions, whereas for ES cells we found mostly GO terms associated with stem cell maintenance. In summary, described above data show that TET2 binds is facilitated to the chromatin active regions with enhancer potential and regulate DNA methylation/demethylation turnover in these regions.

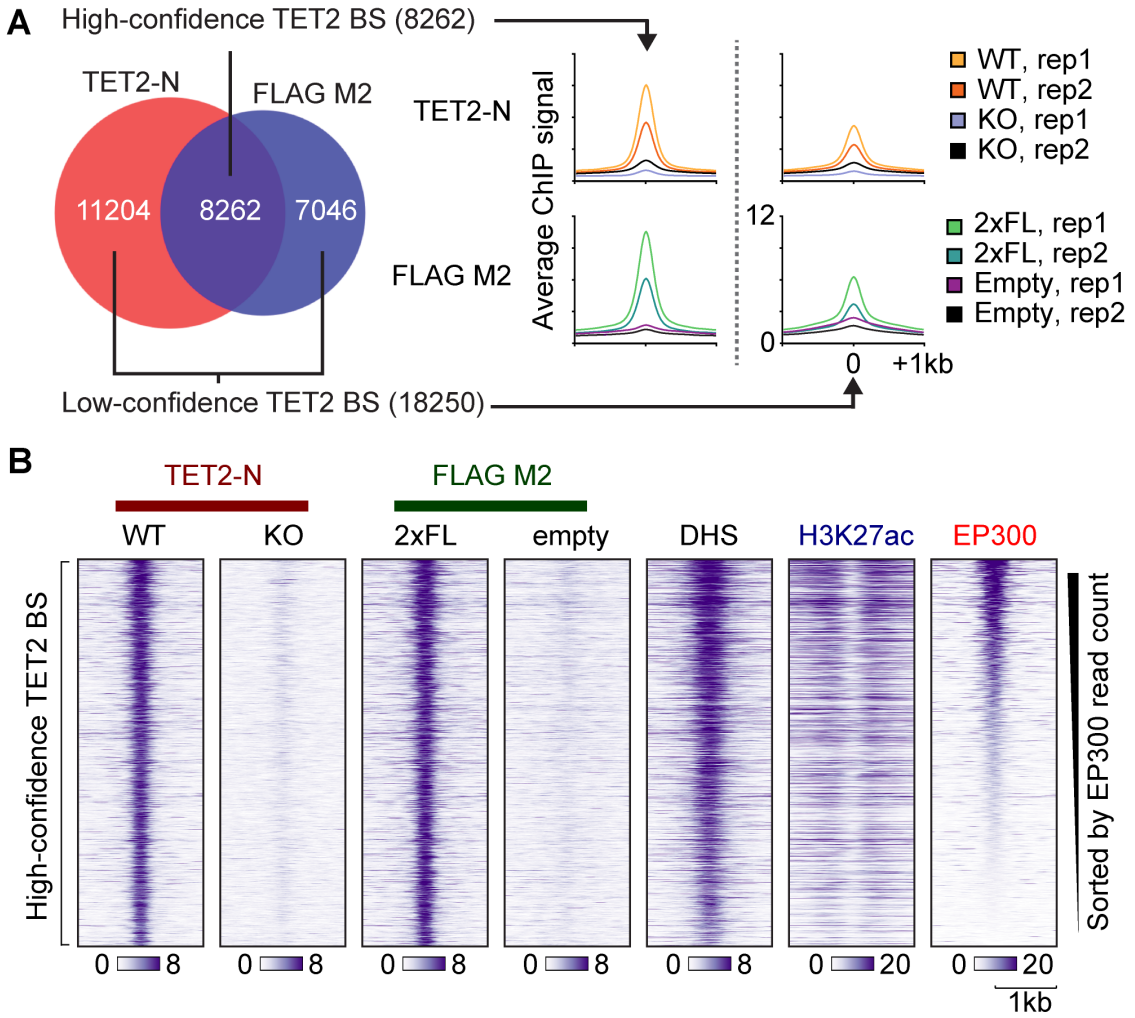

Figure 4.1: **Summary of the TET2 binding regions in the ES cells.** (A) Venn diagram showing overlap of called peaks in TET2-N or FLAG M2 ChIP-seq experiments (left) as well as average ChIP signal from replicate samples (right). High- and low-confidence TET2 binding sites are defined, respectively, as regions showing evidence of TET2 binding in both peak sets (high) or only supported by a called positive region in one peak set (low). (B) Heatmaps of ChIP-seq signal from wild-type TET2 or TET2 with two copies of a FLAG tag (2xFL). Tracks of H3K27ac and EP300 enrichment as well as regions of DHS in ES cells are also shown. The vertical axis contains all high-confidence TET2 binding sites defined in A, sorted by decreasing EP300 read counts. The horizontal axis is centered on TET2 peaks. This figure was provided by Dr. Kasper Dindler Rasmussen and is published in the (Rasmussen et al., 2019).

4.3.2 Changes in the TF activity upon TET2 loss

For the further analyses we generated paired ATAC-seq and RNA-seq datasets from the ES, GMP and MPP cells from the HSC differentiation path, and AML cells, that recapitulate malignant hematopoiesis in the acute myeloid leukemia. For all of these four conditions we obtained samples with and without TET2 knockout (n=4). We preprocessed ATAC-seq data accordingly and defined open chromatin peaks for each sample. Using obtained peaks we generated for each condition consensus peakset, and found that the majority of differential accessible regions were specific for the TET2 knockout samples and partially overlapped with the enhancer annotation (Lara-Astiaso et al., 2014). After this we used described in the previous chapters tool `diffTF` to determine changes in TF activity upon TET2 knockout. To test `diffTF` on the comparatively low size dataset (8 samples in total), using analytical approach to compute p-value (more in the section 2.2.3), we first compared GMP and MPP wild-type conditions (Figure 4.2 A). We found, as expected, specific increased activity of CEBP TFs for the GMP cells (Cirovic et al., 2017) and high activity of the IRF family of TFs for MPP wild-type cells (Seré et al., 2012). Afterwards we ran `diffTF` separately for MPP and GMP comparing samples with TET2 KO versus wild-type, and identify cell-type agnostic set of TFs (ITF2,ZEB1) that were similarly downregulated in TET2 KO samples upon loss of TET2 (Figure 4.2 B).

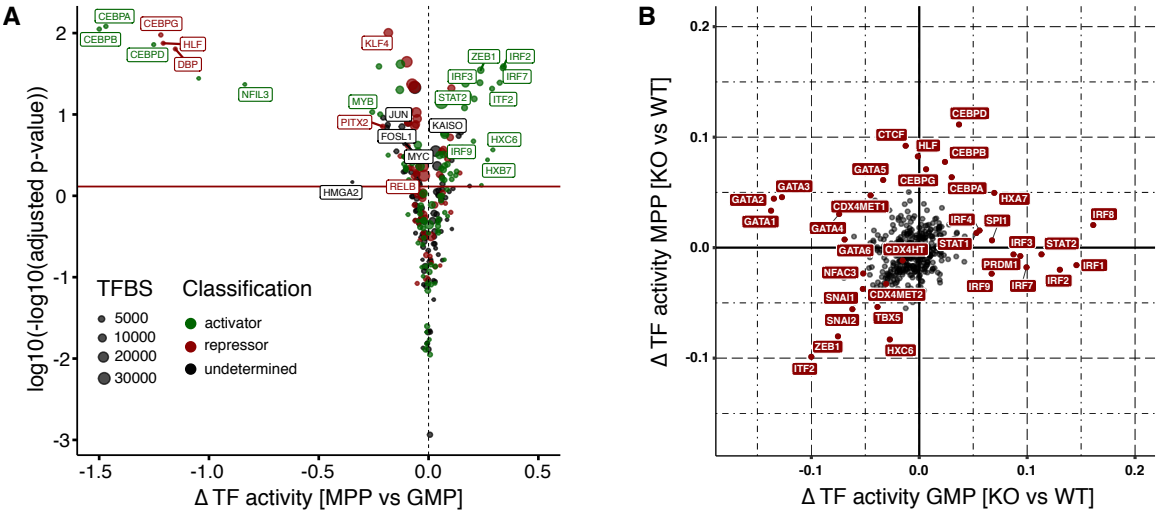

Figure 4.2: **diffTF results for the MPP and GMP.** (A) Volcano plot of differential TF activity between MPPs and GMPs. TFs are colored according to their predicted classification (red, repressors; gray/black, undetermined; green, activators). P values are obtained through `diffTF` using the analytical approach and adjusted by the Benjamini-Hochberg procedure (y-axis). (B) Scatterplot of differential TF activities between MPP and GMP when comparing TET2 KO versus TET2 WT. Only expressed TFs are shown (n = 268). Only the most significant TFs that are relevant for hematopoietic stem cell differentiation are labeled. This figure was produced by myself and part of it (panel A) is published in the (Berest et al., 2019).

Although we were able to highlight potential individual TFs with altered activity upon TET2 knockout, as we discussed in the previous chapter, high PWM similarity between the TFs can lead to the biased interpretation of the effect of the single TF from the shared PWM cluster. To summarize TF activity signal based on PWM similarity we ran *matrix-clustering* tool (Castro-Mondragon et al., 2017) on the combined HOCOMOCO v10 database (Kulakovskiy et al., 2015) and individual motifs from CDX4 and EVX1 motifs from the recent methyl-SELEX paper (Yin et al., 2017). We added these motifs, because original motifs of CDX4 and EVX1 were retracted from the HOCOMOCO database, but they showed interesting changes of TF activity in our dataset. In total, we obtained 88 TF clusters for which we compared overall differential TF activity between TET2 wild-type and knockout states for each condition (Figure 4.3).

We observed widespread significant changes in TF activity across TF clusters in all analyzed cell types. For GMP cells (Figure 4.3 A), the most active clusters upon TET2 knockout are IRF (cluster 18) and STAT (cluster 27) TFs, with significant decrease of activity for the GATA TF family (cluster 3). Whereas for MPP cells (Figure 4.3 B), in the TET2 deficient state we observed an increase of TF activity for CEBP family of TFs (cluster 23) and GATA TF cluster. In comparison to the normal epigenetic differences between MPP and GMP wild-types, as shown in the Figure 4.2 A, such changes suggest anomalous HSC lineage differentiation upon TET2 deficiency. For the AML (Figure 4.3 C) and ES (Figure 4.3 D) cells we haven't observed the same ranges of TF activity in TET2 knockout and the majority of the TFs were more active in the wild-type condition. Similar to GMP cells AML TET2 wild-type cells showed increased activity of the GATA and RFX (cluster 36) TF family. ES cells with TET2 knockout showed decreased activity of the PWM cluster 6, containing nuclear receptors of steroid hormones (e.g. NR4A1). Finally, for all the cell types we characterized decreased activity of the basic helix-loop-helix (bHLH; cluster 12) family of TFs and HOX protein family (cluster 8), especially CDX4 and EVX1 SELEX motifs upon TET2 knockout.

As we already showed previously that TET2 binding predominantly occurs in the enhancer regions, we decided to split consensus peaks in each condition comparison in four groups: *all* - correspond to the all peaks; *Pro* - peaks overlapped with the promoter regions of the protein-coding genes, *ProDist* - all peaks without promoter associated peaks (enhancer regions) and for ES and AML cell types we had also *CHIP* categorie, that is basically overlapped set of regions between consensus peaks and TET2 ChIP-seq data for these conditions.

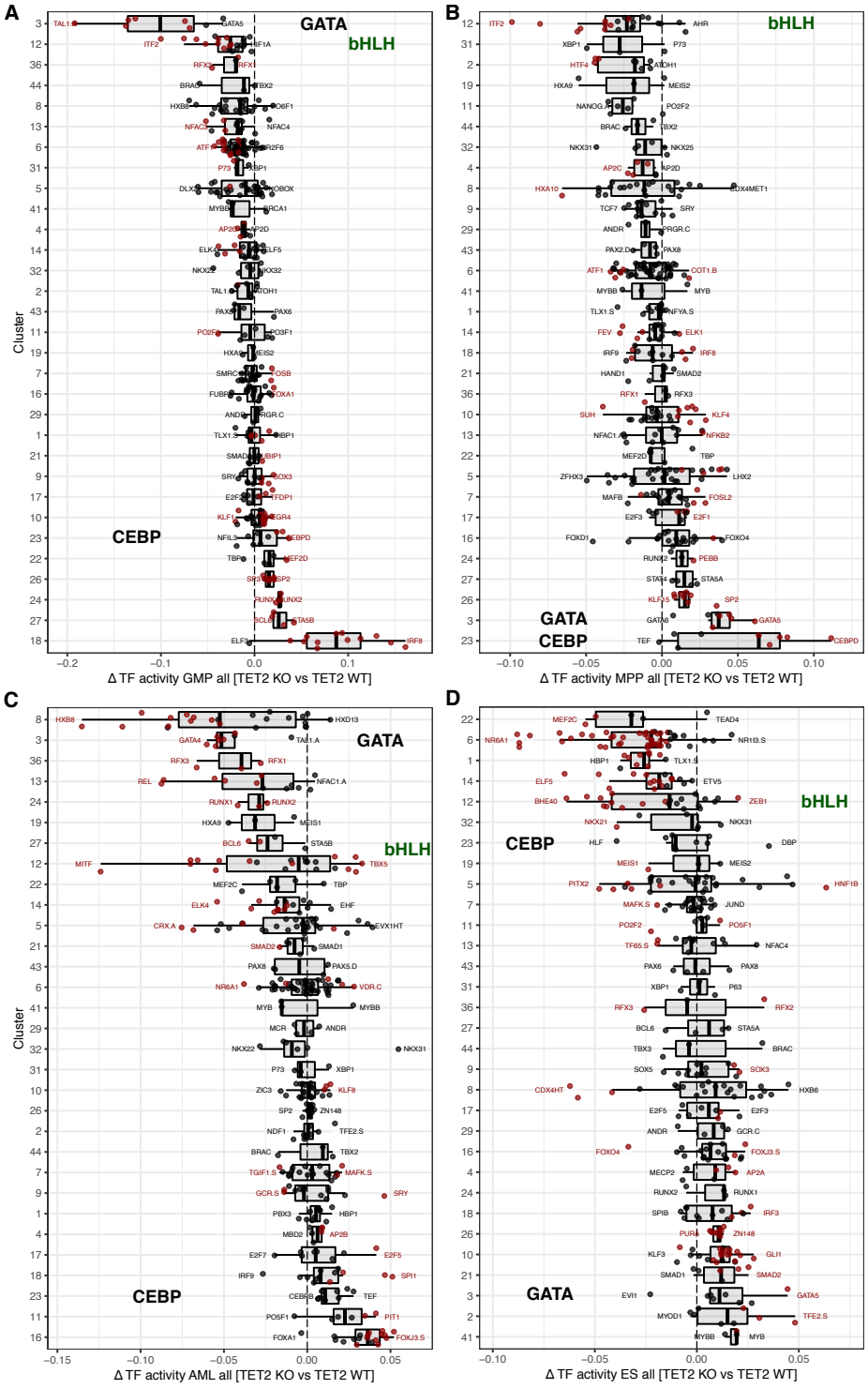

Figure 4.3: **TF activity summarized by PWM similarity upon loss of TET2 .** Box plot showing weighted mean difference values obtained from `diffTF` analyses using all consensus peaks for each TF cluster in GMP (A), MPP (B), AML (C) and ES (D) cells comparing Tet2 knockout versus wild type. Individual TFs within a cluster are shown (black dots), and TFs passing a significance threshold (p-value < 0.1) are highlighted (red dots). The predominant TF identity of selected clusters (12, 3, and 23) are marked. This figure was produced by myself and part of it (panel B) is published in the (Rasmussen et al., 2019).

After that, for each condition and peaks group we ran `diffTF` to compare TF activity changes between TET2 knockout samples and wild-type, grouped resulted TF activity values by PWM clusters and visualized obtained results as heatmap (Figure 4.4 A). Intriguingly, we observed consistent changes of TF activity for several clusters, that can explain impaired regulation of the HSC differentiation program under the TET2 knockout.

For the cluster 12, containing mostly bHLH TFs, we observed the lowest median TF activity across all conditions and other clusters, with specific decrease of signal in *AML.ProDist* and *AML.CHIP* comparisons (Figure 4.4 C). This cluster consists of 22 TFs that facilitate the binding of E-box elements with a specific 5'-CANNTG-3' binding sequence and initiate gene transcription (Massari and Murre, 2000). All motifs of this cluster contain central CpG site, and it was shown before that altered methylation/demethylation regulation of this site can impair *in vitro* DNA binding (Yin et al., 2017). Also, we observed an unexpected increase of TF activity for half of the TFs in this cluster for the ES and AML cells, suggesting that apart from effects mediated by TET2 regulation they are also affected by other cell type specific mechanisms in these conditions. Using gene expression data from RNA-seq in the MPP cells we showed that MYC/MAX and HIF1A/ARNT heterodimers, as well as ITF2 and ZEB1 TFs, are the most expressed TFs (Figure 4.4 D), suggesting that the `diffTF` cluster 12 signal can be mostly influenced by the presence of these TFs. However, none of the TFs from this cluster were significantly differentially expressed between TET2 knockout and wild-type (Figure 4.4 E), thus revealing that TET2 absence is mostly affecting chromatin binding rather than TF expression.

Overall decrease of TF activity in the TET2 deficient state showed also cluster 8 TFs, containing TFs with homeobox domain (Figure 4.4 B). Summarized signal across all TFs of this cluster was negative in the AML and GMP condition, suggesting epigenetic changes of the regulation in normal and malignant hematopoiesis, and specifically CDX4, PO3F2 and EVX2 motifs have decreased activity in almost all cases. Interestingly, for *AML.Pro*, *MPP.Pro* and *ES.pro* comparisons TF activity of these factors were not significantly decreased, but rather increased, proposing that TET2 binding is affecting regulation of such TFs particularly in the enhancer regions. Similarly to the cluster 12 TFs, we didn't observe significant differences of the CDX4 and EVX2 TF expression between wildtype and TET2 knockout states.

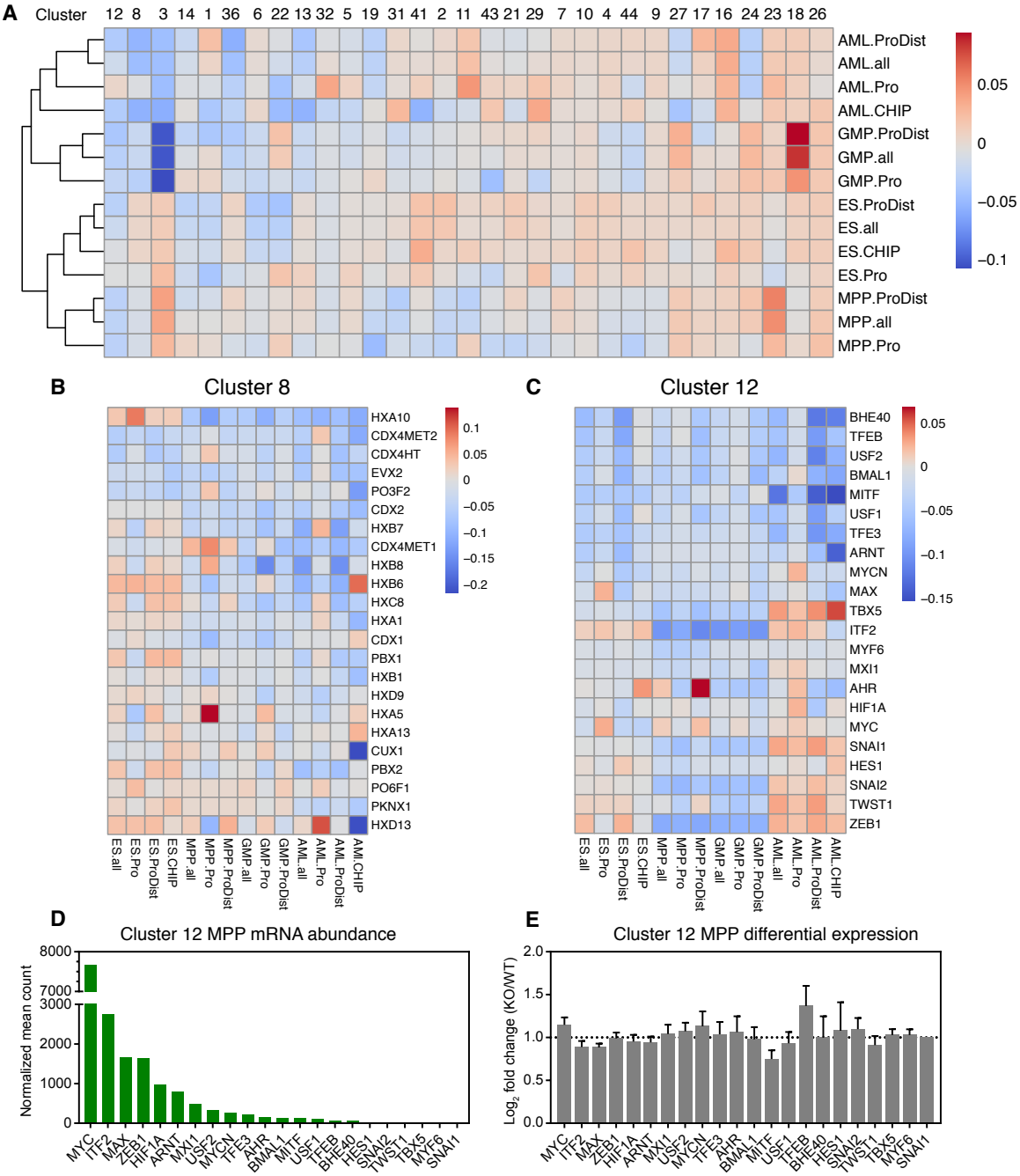

Figure 4.4: **Summary of `diffTF` analysis of multiple cell types with loss of TET2.** (A) The heatmap represents the summary of `diffTF` analyses between Tet2 knockout and wild type across multiple cell types. The color scale of the heatmap corresponds to the Z-scores of the weighted mean difference values (TF activity) obtained from diffTF. Only TF clusters with three or more TFs are shown. For each cell type, the analysis was run on different sets of peaks: promoter regions only (-1.5 kb/+500 bp from TSS; labeled celltype.Pro) and putative enhancer regions (+100 kb/-100 kb from TSS excluding promoter regions; celltype.ProDist), in addition to the full set of peaks (celltype.all). For the cell types for which we mapped TET2 binding sites by ChIP-seq (ES cells and AML cells), `diffTF` was also run using ATAC-seq peaks intersected with TET2 ChIP in the corresponding cell type (celltype.CHIP). Detailed heatmap for TF similarity clusters that showed consistent changes across investigated cell types are shown, cluster 8 (B) and cluster 12 (C). (D) Histogram showing mean normalized counts for Cluster 12 TFs obtained from RNA-seq data generated from wild-type and Tet2 knockout MPP cells. The TFs within Cluster 12 are ranked based on mean normalized abundance. (E) DESeq2 log2 fold changes of Cluster 12 TFs ranked as in C. This figure was produced by myself and part of it (panels A,D,E) is published in the (Rasmussen et al., 2019).

4.3.3 TFs mode of action are conserved upon TET2 knockout

In order to test validity of activator/repressors `diffTF` classification mode for the small size multiomics dataset (n=8), we compared TF specific changes in gene expression and chromatin accessibility for the GMP versus MPP comparison and GMP TET2 knockout versus wild-type case. We assigned TF mode of action relying only on the GMP and MPP samples with normal TET2 functioning (Figure 4.5 A), and observe expected highly significant correlation (ρ=0.69) for activators and negative correlation for the repressors (ρ=-0.58). Afterwards, we transferred such TF classification to the GMP specific TET2 knockout versus wild-type comparison (Figure 4.5 B) and found similar direction of the correlations (activators: ρ=0.44; repressors: ρ=-0.32). 63% of all expressed TFs (158) in GMP TET2 specific case have similar to the HSC differentiation relation between expression and activity.

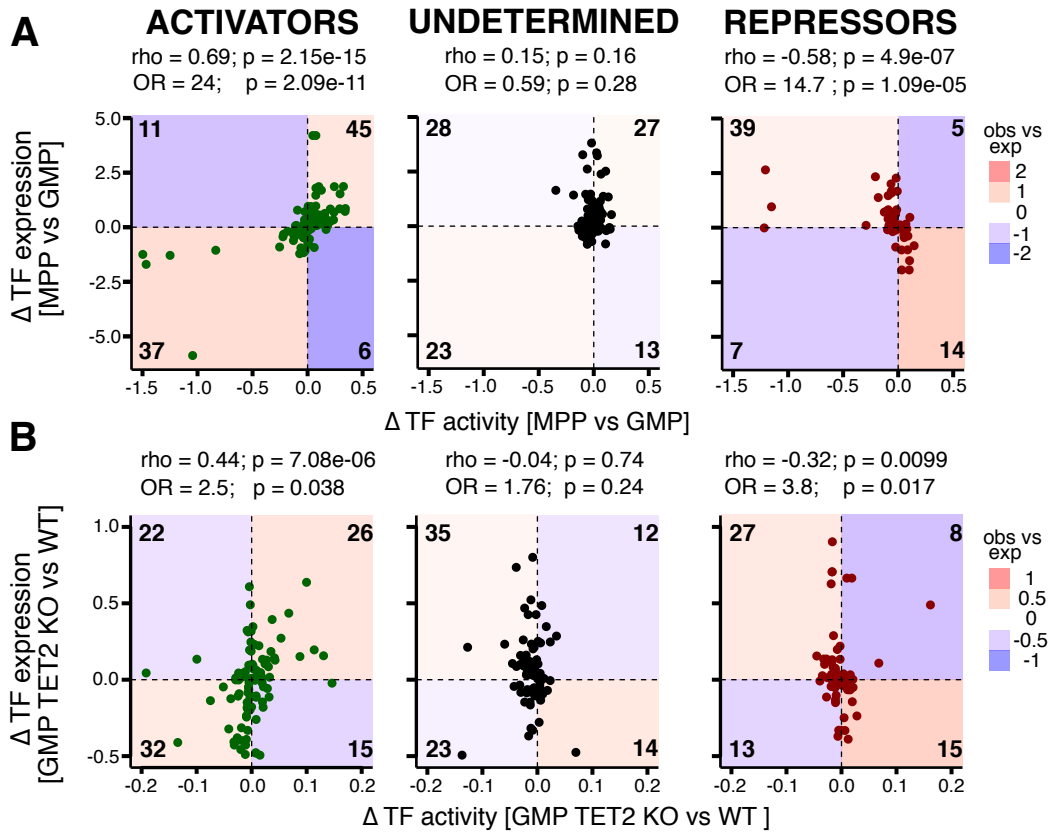

Figure 4.5: **TF classification through HSC differentiation.** Correlation of differential TF activity and differential expression (log2 fold changes obtained from DESeq) are shown for activators (left), undetermined TFs (middle), and repressors (right) for the comparison of GMP versus MPP (A) and GMP TET2 knockout versus wild type, which represents independent data (B). Color shadings in (A) and (B) indicate the observed versus expected log2 ratio for each quadrant (blue, less than expected; red, more than expected). TF classifications were obtained from the wild-type GMP and MPP samples. Spearman's rho and p value, as well as the odds ratio (OR) and p values of Fisher's exact test, are provided. The number of TFs is indicated in each quadrant. This figure was produced by myself and is published in the (Berest et al., 2019).

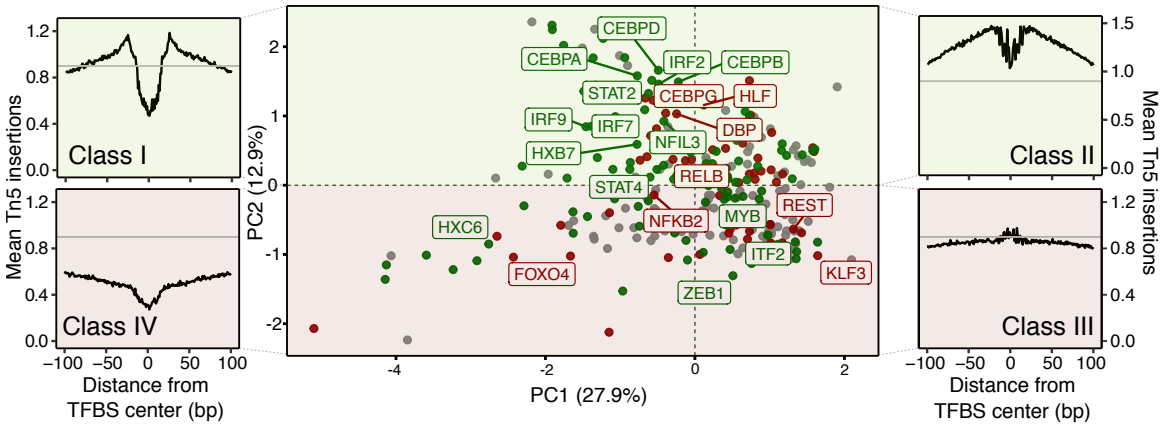

Figure 4.6: **GMP/MPP TF footprinting analysis.** Scatterplot comparing PC1 and PC2 from a PCA of the footprints of all expressed TFs (n = 268). The insets (classes I–IV) display the average footprint across TFs in that quadrant. This figure was produced by myself and is published in the (Berest et al., 2019).

We also used obtained from GMP/MPP comparison TF classification and performed TF footprinting analysis for all 268 expressed TFs in that conditions (Figure 4.6). Taking into consideration the first two PCs from the PCA analysis of footprinting data, we observed comparable to the human CLL data (Figure 3.4 A) separation of the mouse TFs. Similar to the described in the previous chapter results PC1 corresponds to the changes of chromatin accessibility in the motif center, and PC2 to the openness of the flanking regions. Such results are very interesting, suggesting that with footprinting analysis we characterize general TF biology, independent of conditions and species. Highlighted above condition specific CEBP and IRF TF families with huge differences in TF activity between GMP and MPP (Figure 4.2 A) are classified as activators and showed typical activator footprints. However, the strongest repressors JUN and KLF4 overall didn't have a huge range of differences in TF activity, therefore recapitulating the hypothesis that MPP to GMP transition is mostly occurring through the activation of TF regulatory network (Cheng et al., 2019). Surprisingly, TFs that were shown to be affected by TET2 regulation (MYB,ITF2 and ZEB1) were classified as activators, but had a repressor footprint.

Overall, described analyses once more validate `diffTF` classification mode and showed that it can be used even for datasets with low amounts of biological replicates.

4.4 Discussion

Despite multiple studies extensively characterizing the effect of TET2 on DNA methylation dynamics (Hon et al., 2014; Rasmussen et al., 2015; Yamazaki et al., 2015), there is still limited understanding of the regulatory mechanisms underlying the impact of TET2 on the normal and malignant hematopoietic differentiation. In this chapter we presented a novel, highly specific set of TET2 binding regions for the ES and AML cells. We observed that the majority of TET2 binding is occurring in enhancer regions and is highly overlapping with EP300 binding sites. Interaction on the molecular level between EP300 and TET2 were shown before (Zhang et al., 2017), as well as co-localization of EP300 binding sites with enhancer and super-enhancer regions (Ebrahimi et al., 2019). Generation of the refined genome-wide map of TET2 binding can lead to further linkage of DNA methylation and chromatin histone marks for the cell-type specific gene regulation. Based on the DNA methylation/hydroxymethylation data (Hon et al., 2014) we confirmed that TET2 knockout leads to the increase of the DNA methylation and decrease of the hydroxymethylation in the regions of binding. It is important to note that we didn't observe a linear relationship between ChIP-seq signal and DNA methylation signal, suggesting that, although TET2 plays a key role in the demethylation cycle, binding of it cannot fully determine methylation status. Low enrichment of CpG islands in the TET2 binding sites also highlight the importance of additional, independent of TET2, DNA methylation mechanisms regulating CpG islands function (Rasmussen and Helin, 2016).

To characterize further molecular mediators of the TET2 regulation in hematopoiesis we generated a complex dataset of paired chromatin accessibility and gene expression data in multiple cell types. We characterize genome-wide chromatin changes for multiple TFs at once upon TET2 knockout using differential TF activity measure from `diffTF`. Taking into account PWM similarity bias of different TF motifs we summarized the changes for each PWM cluster and identified groups of TFs that showed changes in activity on different states of HSC differentiation in the TET2 deficient state. Intriguingly, we didn't observe any consistent activity of the GATA TF family in all conditions, with decrease of the activity in the GMP and AML cell types and increased activity for MPP and ES cell types. GATA TFs were shown to bind methylated regions (Zhu et al., 2016), and we showed that upon TET2 knockout DNA methylation overall increases, therefore expecting an overall increase of GATA TFs activities. This discrepancy can be explained by the fact that

in GMP, and similarly in AML cells, GATA TFs are not expressed. CEBPB TFs have been reported to have a preference for binding to motifs with a methylated cytosine (Mann et al., 2013), and showed more consistent signal of increased activity upon TET2 knockout, with only difference in the ES cells.

Despite the cell-type specific signal, we found two potential TF families (HOX and bHLH family) that have overall a similarly decreased activity across all conditions upon TET2 knockout. Several studies suggested a potential relationship between members of bHLH TFs and TET2. It was shown that TET2 is recruited to c-MYC binding sites by SNIP1 (Chen et al., 2018), as well as increased DNA methylation inhibits MYC binding in the ES cells from triple TET knockout mice (Yin et al., 2017). Decreased activity of ITF2 and ZEB1 TFs can be explained by the involvement of MYC in the regulation of the Wnt signalling pathway (Wong et al., 2014). These observations in literature highlight dysregulation of activity for bHLH TFs in the TET2 knockout hematopoietic cells and, hopefully, will stimulate further research of their role in the malignant hematopoiesis. We also observed a decreased activity of HOX TFs, in correspondence with observations that TET2 activity is responsible for the expression of the HOXA cluster of genes (Bocker et al., 2012). Notably, as CDX4 was described to regulate early embryonic hematopoietic development (Wang et al., 2008), the hypothesized binding of the TET2 to it, supported by our findings, can be of a particular interest for future research. Together, these results suggest that loss of TET2 catalytic activity, and the resulting aberrant DNA methylation turnover, has pleiotropic consequences on chromatin accessibility and TF activity.

We also showed that upon genetic perturbation (TET2 knockout in this case)the majority of the TFs keep their global mode of action. Apart from this we found activator/repressor specific TF footprinting patterns of the surrounding motif flanking regions. The latter observation is very exciting, as it highlights different chromatin microenvironments for the activators and repressors. Additionally, with the increase of research highlighting the ability of TFs to bind nucleosomes and closed chromatin (Mayran et al., 2019), we need to refine further our footprinting analysis and integrate it with more epigenetic molecular phenotypes.

Chapter 5

Conclusions and future perspectives

Tremendous technological developments of sequencing technologies combined with novel experimental methods improved significantly our knowledge of chromatin biology. In my opinion, chromatin is a very direct measure and a key for understanding complex gene regulation mechanisms in the cell. Inevitably in the future, integration of data from various chromatin related molecular phenotypes, such as information about histone modifications or DNA methylation, will reveal fundamental basic biological logic behind cell functioning. Important roles in the regulatory pathways play TFs, as major source of reaction and adaptation of cellular internal encoded program to the extra- and intracellular changes.

In the presented dissertation I tackled the problem of inferring TF activity from the integrated multiomics data. We developed a novel computational tool `diffTF` (Berest et al., 2019) that summarizes changes in chromatin accessibility for the estimated sites of TF binding and through the extensive statistical workflow define differentially active TFs in pairwise comparisons. Additionally, `diffTF` uses gene expression data to predict global functional roles of TF in such comparisons. We benchmarked our algorithm on one of the biggest multiomics dataset and defined potential biases and limitations of our tool. Notably, `diffTF` is the first published computational tool that integrates chromatin accessibility and gene expression to define TF activity and function. It is important to note that on top of the development of methods for assessing biological systems, there is also constant evolution of programming algorithms and tools used in bioinformatics. We aimed to use up-to-date novel computational approaches for the construction of our pipeline, such as Snakemake workflow management system, that allowed to reduce `diffTF` complexity for the user.

We extensively used `diffTF` for many different projects, however, in this dissertation I focused on two main datasets. First we applied `diffTF` to the multiomics dataset from CLL patients(Rendeiro et al., 2016) to define differences in TF regulatory landscapes between two subtypes of CLL. Using our approach we found 68 differentially active TFs, part of which (∼40%) were already known to be involved in the differential regulation between U-CLL and M-CLL. Importantly, we found novel TFs and based on the published literature hypothesized about potential functions of them in the progression of each subtype. We validated *in silico* data-driven classification into activators and repressors. We also used an external CLL multiomics dataset for the experimental validations of the `diffTF` classification mode, and found that our classification is robust to the perturbations caused by the drug treatment. Defined TFs can be used in future to design novel chromatin-targeted strategies of personalized CLL treatment.

In the chapter 4, we also applied `diffTF` to the multiomics dataset characterizing the effects of DNA methylation turnover affected by TET2 on the chromatin changes during hematopoietic system differentiation (Rasmussen et al., 2019). Using the first high quality map of TET2 binding we defined two major classes of TFs regulated by TET2 (bHLH and Hox family). We also showed that TF activity changes are also occurring in AML cells, therefore confirming the vast impact of DNA methylation on chromatin changes during leukemogenesis. For both datasets we performed TF footprinting analysis, which highlight interesting dynamics between predicted activators and repressors. Increased chromatin accessibility of the flanking regions are specific for the activator TFs, whereas repressor TFs are mostly located in the closed chromatin. This interesting observation can potentially explain different mechanisms of binding of activator and repressors, possibly even dissecting the roles of TF in different chromatin context. These statements, however, need further careful evaluation and testing.

With the burst of the single cell technologies, we are aiming in the future to add `diffTF` functionality to work with the single cell data. This will require change of the technical aspects of the pipeline, as well as statistical methods behind the method. We are also planning to transfer the `diffTF` pipeline to the R/Bioconductor environment to make it easier to use for biologists around the world.

Appendix A

Original CLL dataset metadata

Patient ID	name	Gender	Age	IGVH, %	IGVH status	batch	RNA-seq
50	ATAC_50	F	84	94.00	mutated	1	-
244	ATAC_244_1	M	61	89.22	mutated	2	-
244	ATAC_244_2	M	69	89.22	mutated	2	-
552	ATAC_552_1	M	72	100.00	unmutated	3	-
552	ATAC_552_2	M	75	100.00	unmutated	3	-
653	ATAC_653_1	F	76	99.65	unmutated	2	-
653	ATAC_653_2	F	82	99.65	unmutated	2	-
680	ATAC_680_1	M	52	92.00	mutated	2	+
680	ATAC_680_2	M	57	92.00	mutated	2	-
981	ATAC_981_1	M	79	91.39	mutated	3	-
981	ATAC_981_2	M	83	91.39	mutated	4	-
1125	ATAC_1125_1	M	61	93.40	mutated	2	-
1125	ATAC_1125_2	M	65	93.40	mutated	2	-
1303	ATAC_1303	M	90	93.75	mutated	1	-
1781	ATAC_1781	F	75	91.00	mutated	5	+
2132	ATAC_2132_1	F	54	100.00	unmutated	3	-
2132	ATAC_2132_2	F	56	100.00	unmutated	3	-
2459	ATAC_2459_1	M	58	93.00	mutated	5	-
2459	ATAC_2459_2	M	63	93.00	mutated	5	-
2483	ATAC_2483	F	90	100.00	unmutated	5	-
2613	ATAC_2613_1	M	82	96.60	mutated	3	-
2613	ATAC_2613_2	M	85	96.60	mutated	3	-*
2886	ATAC_2886_1	M	76	92.00	mutated	5	-
2886	ATAC_2886_2	M	77	92.00	mutated	5	-
2938	ATAC_2938_1	M	74	97.60	mutated	2	-
2938	ATAC_2938_2	M	77	97.60	mutated	2	-
2938	ATAC_2938_3	M	80	97.60	mutated	2	-
2938	ATAC_2938_4	M	82	97.60	mutated	2	-
2938	ATAC_2938_5	M	83	97.60	mutated	6	-
2977	ATAC_2977_1	M	60	97.58	mutated	6	-
2977	ATAC_2977_2	M	67	97.58	mutated	2	-
3069	ATAC_3069_1	F	66	100.00	unmutated	1	-
3069	ATAC_3069_2	F	71	100.00	unmutated	3	-
3142	ATAC_3142	F	82	99.30	unmutated	5	-
3156	ATAC_3156_1	F	65	100.00	unmutated	2	-
3156	ATAC_3156_2	F	72	100.00	unmutated	3	-
3215	ATAC_3215_1	M	73	89.90	mutated	5	-
3215	ATAC_3215_2	M	77	89.90	mutated	5	-
3215	ATAC_3215_3	M	82	89.90	mutated	5	-
3215	ATAC_3215_4	M	83	89.90	mutated	5	-
3240	ATAC_3240	M	84	100.00	unmutated	5	+

3263	ATAC_3263_1	M	66	95.30	mutated	2	-
3263	ATAC_3263_2	M	73	95.30	mutated	2	-
3386	ATAC_3386	M	82	91.16	mutated	3	-
3439	ATAC_3439_1	F	88	96.88	mutated	5	-
3439	ATAC_3439_2	F	90	96.88	mutated	5	-
3439	ATAC_3439_3	F	95	96.88	mutated	5	-
3492	ATAC_3492_1	F	84	100.00	unmutated	5	-
3492	ATAC_3492_2	F	87	100.00	unmutated	6	-
3756	ATAC_3756	F	61	98.61	unmutated	6	-
3811	ATAC_3811	M	58	93.95	mutated	2	-
3823	ATAC_3823_1	M	66	95.92	mutated	3	-
3823	ATAC_3823_2	M	71	95.92	mutated	1	-
3873	ATAC_3873	F	49	98.00	unmutated	5	-*
3943	ATAC_3943	M	59	100.00	unmutated	5	-
3980	ATAC_3980	M	81	100.00	unmutated	5	-
4034	ATAC_4034	F	77	91.60	mutated	1	-
4078	ATAC_4078_1	F	77	100.00	unmutated	3	-
4078	ATAC_4078_2	F	80	100.00	unmutated	6	-
4080	ATAC_4080	M	79	90.50	mutated	2	+
4102	ATAC_4102_1	M	78	92.44	mutated	5	-
4102	ATAC_4102_2	M	82	92.44	mutated	5	-
4102	ATAC_4102_3	M	83	92.44	mutated	5	-
4189	ATAC_4189	M	84	100.00	unmutated	1	+
4251	ATAC_4251_1	M	61	100.00	unmutated	6	-
4251	ATAC_4251_2	M	64	100.00	unmutated	6	-
4333	ATAC_4333	M	54	100.00	unmutated	6	+
4621	ATAC_4621_1	M	66	91.32	mutated	3	-
4621	ATAC_4621_2	M	69	91.32	mutated	4	-
4621	ATAC_4621_3	M	69	91.32	mutated	3	-
4668	ATAC_4668_1	M	50	99.65	unmutated	5	-
4668	ATAC_4668_2	M	49	99.65	unmutated	6	-
4747	ATAC_4747	F	82	93.40	mutated	7	-
4784	ATAC_4784	F	82	93.40	mutated	6	+
4802	ATAC_4802	NA	NA	NA	NA	6	-
4963	ATAC_4963	M	68	86.11	mutated	1	-
4989	ATAC_4989	M	72	100.00	unmutated	5	-
5019	ATAC_5019	F	74	NA	NA	5	-
5044	ATAC_5044	M	73	97.22	mutated	6	-
5048	ATAC_5048	F	83	90.97	mutated	5	-
5129	ATAC_5129	F	86	100.00	unmutated	2	-
5147	ATAC_5147	M	47	100.00	unmutated	8	-
5170	ATAC_5170	F	88	NA	NA	5	-
5199	ATAC_5199	M	59	100.00	unmutated	5	-
5204	ATAC_5204	M	76	98.98	unmutated	1	-
5229	ATAC_5229	M	65	100.00	unmutated	5	-
5263	ATAC_5263	M	71	100.00	unmutated	2	+
5277	ATAC_5277	NA	NA	NA	NA	8	-

Appendix B

CLL TFs literature annotation

TF motifs	Mode of action	DBD domain	Function	CLL association
ARI3A.S, ARI3A.D	repressor	ARID/BRIGHT	Arid3a altering accessibility of IGVH genes to rearrangement in fetal development. Impact selection of B-cells.	no
ASCL2	activator	bHLH	It activates transcription by binding to the E box (5'-CANNTG-3').	no
CLOCK	repressor	bHLH	BMAL1:CLOCK regulate circadian cycles in CLL.	yes
EGR3	undetermined	C2H2 ZF	Part of the BCR signaling pathway. Neriched in hypomethilated CpG sites in CLL	yes
ERF	undetermined	ETS	ETS repressor factor.	no
ESR1.A	undetermined	Nuclear receptor	Is repressed by mir-18a that bind to it mRNA at the 3' UTR.	yes
GATA2, GATA3	undetermined	GATA	HSCs in CLL have cell-intrinsic propensity to generate clonal CLL-like B cells.	no
GCM1	undetermined	GCM	Key factor in the beta-catenin/GCMa/syncytin-1 pathway.	no
GFI1, GFIB	undetermined	C2H2 ZF	GFI-1 involved in B cell development	no
GLI1, GLI3	undetermined	C2H2 ZF	Part of the HH signaling pathway. Inhibition of it results in reduced apoptosis and bigger survival of the CLL cells.	yes
HESX1	undetermined	Homeodomain	HESX1 is differentially expressed gene comparing endothelocyte and B-cell CLL.	no
HXB7	undetermined	Homeodomain	HoxB7 repress expression of the death-associated protein kinase 1 (DAPK1), which result in heritable predisposition to CLL.	yes
KAISO.S	activator	C2H2 ZF	Overexpression of Kaiso significantly increased cell viability and inhibited hydrogen peroxide-induced apoptosis. Also is effected by miRNAs.	no
KLF16	repressor	C2H2 ZF	KLF16 is a novel regulator of metabolic genes by regulatable coupling to Sin3-histone deacetylase complexes.	no
MUSC	undetermined	bHLH	Suppressed activity by BLIMP-1(PRDM1). Repress E2A proteins. Downstream gene of the B-cell receptor signal transduction pathway. May play a role in regulating antigen-dependent B-cell differentiation.	no
MYBB	undetermined	Myb/SANT	Regulator of cell-cycle progression and cell survival.	yes
NF2L1	undetermined	bZIP	Related to PAK Pathway and ERK Signaling.	no
NR1D1	undetermined	Nuclear receptor	Circadian rhytms regulation. Repression of the BMAL1/CLOCK complex.	no
NR1H2	repressor	Nuclear receptor	Regulate lipid and cholesterol metabolism. Promote apoptosis and regulate cell proliferation.	yes
NR1I2.S	undetermined	Nuclear receptor	Transcriptional regulator of the cytochrome P450 gene CYP3A4. Hyperforin activates PXR and SXR and promote apoptosis in CLL.	yes
NR4A3	undetermined	Nuclear receptor	Affected by IgM stimulation and cAMP.	yes
OVOL1	undetermined	C2H2 ZF	Critical for Mesenchymal to Epithelial Transition (MET) plasticity.	no
P73.S	repressor	p53	Activate p53-mediated apoptosis mitochondrial apoptosis.	yes
PO3F1	undetermined	POU	Regulate neural commitment from ES cells.	no
PPARD	undetermined	Nuclear receptor	Promotes survival of CLL in energetically unfavorable conditions.	yes
PRRX1	undetermined	Homeodomain	Transcription co-activator, enhancing the DNA-binding activity of serum response factor, a protein required for the induction of genes by growth and differentiation factors.	no

RARG.C	undetermined	Nuclear receptor	Induce CD1d expression by CLL cells, which in combination with iNKT cell agonist glycolipids could sensitize CLL cells forlysis by iNKT cells.	yes
RORA	undetermined	Nuclear receptor	A potent negative regulator of NF-KB signaling.	yes
RREB1	activator	C2H2 ZF	Represses miR-143 and miR-145 Promoter.	no
SMAD2	undetermined	SMAD	Phosphorylated by activated type I receptor for TGF-$beta$ and associated with Activin (Act) pathway.	yes
TAL1.A	activator	bHLH	Key marker of the T-cell acute lymphoblastic leukemia.	no
ZBT49	undetermined	C2H2 ZF	Inhibits cell proliferation by activating either CDKN1A/p21 transcription or RB1 transcription.	no
ZBTB6	activator	C2H2 ZF	Involved in B-cell development have been associated with cell cycle and cell survival control.	no
ZFHX3	repressor	C2H2 ZF; Homeodomain	Negatively regulate c-Myb, and transactivate the cell cycle inhibitor cyclin-dependent kinase inhibitor 1A (also known as p21CIP1). This gene is reported to function as a tumor suppressor in several cancer.	no
ZKSC1	undetermined	C2H2 ZF	This encoded protein may function as a transcription factor that regulates the expression of GABA type-A receptors in the brain. Transcripts from this gene have been shown to form stable and abundant circular RNA.	no
ZN784	activator	C2H2 ZF	May be involved in transcriptional regulation.	no
BC11A	repressor	C2H2 ZF	T(2,14) translocation increase expression of the BC11A in U-CLL.	yes
BHE40	activator	bHLH	Controls the development and self-renewal of B-1a cells.	no
BMAL1	activator	bHLH	Circadian regulation; should be bigger in normal or in the treated with radiology samples.	no
CEBPB	repressor	bZIP	ATF6 and C/EBP-beta complex is required for the IFN-gamma-induced expression of DAPK1. Defects in this pathway fail to control growth of CLL.	yes
FOXO4	undetermined	Forkhead	TF target enrichment analysis (MsigDB database) on gene lists retrieved from DNA methylation data showed FOXO4 TFBS enriched for those genes.	no
IRF1, IRF2, IRF3, IRF4, IRF5, IRF7,IRF8, IRF9	activator, undetermined	IRF	IRF pathway is highly expressed and regulate cell proliferation and apoptosis. Ma et al.: IRF4 suscept-locus for CLL, IRF 4 (+/-) mice have shorter onset of CLL, increased B1 cell survival (precursors of CLL cells), accellerated cell renewal and resistnace to apoptosis. Injection of IRF4 inhibits their survival. Similar results in human. High IRF4 supress AKT activity. NFKB, JNK/p38, NF/IL6 and IRF pathways are intermediate to highly expressed (due to high expression of TLRs).	yes
MAX, MYC	activator, undetermined	bHLH	Several reports of c-Myc in CLL, high mRNA expression is associated with poor prognosis. Mostly in un-mutated.	yes
PAX2.S	undetermined	Homeodomain Paired box	PAX2 Protein Induces Expression of Cyclin D1 through Activating AP-1 Protein and Promotes Proliferation of Colon Cancer Cells.	no
PAX5.A	repressor	Paired box	Altered isoform expression of PAX5 is involved in lymphomagenesis.	yes
PAX6	activator	Homeodomain Paired box	PAX6 dependent regulation of BCL6.	no
PRDM1	repressor	C2H2 ZF	Suppresion of the PRDM1 supress differentiation.	yes
STAT2	activator	STAT	STAT activation is a common characteristic of leukemias.	yes
TAL1.S	activator	bHLH	Aberrant activation by interchromosomal interactions.	no
TF7L2	undetermined	HMG/Sox	Participates in the Wnt signaling pathway and modulates MYC expression by binding to its promoter in a sequence-specific manner.	no
TFAP4	undetermined	bHLH	Misregulation contribute to genomic instability and tumorigenesis.	no
TFE3, TFEB	activator, undetermined	bHLH	TFE-fusion proteins are highly overexpressed in cancers. Mainly because of the chromosomal rearrangements.	no
USF1, USF2	activator	bHLH	USF1 recruits histone modification complexes and is critical for maintenance of a chromatin barrier.	no
ZBT18	repressor	C2H2 ZF	Cell movement and cell migration in leukemias.	no

Appendix C

GO enrichment results for CLL TFs

Gene Ontology term name	AUC	PRC	TF count
defense response to other organism	0.918	0.563	9
response to interferon-gamma	0.886	0.729	9
interferon-gamma-mediated signaling pathway	0.886	0.729	9
cellular response to interferon-gamma	0.886	0.729	9
response to type I interferon	0.867	0.798	10
type I interferon signaling pathway	0.867	0.798	10
cellular response to type I interferon	0.867	0.798	10
response to biotic stimulus	0.827	0.516	16
response to external biotic stimulus	0.827	0.516	16
response to other organism	0.827	0.516	16
adaptive immune response	0.798	0.272	10
immune effector process	0.791	0.381	13
regulation of leukocyte differentiation	0.786	0.293	13
myeloid leukocyte differentiation	0.762	0.182	9
cytokine-mediated signaling pathway	0.760	0.664	13
immune response	0.756	0.642	26
innate immune response	0.751	0.636	17
regulation of leukocyte activation	0.740	0.207	11
regulation of cell activation	0.740	0.207	11
positive regulation of cytokine production	0.734	0.336	11
regulation of hemopoiesis	0.733	0.285	15
response to cytokine	0.726	0.582	18
cellular response to cytokine stimulus	0.726	0.582	18
immune system process	0.714	0.667	39
response to bacterium	0.714	0.185	9
response to molecule of bacterial origin	0.714	0.185	9
positive regulation of immune system process	0.714	0.321	17
regulation of innate immune response	0.710	0.294	10
symbiosis, encompassing mutualism through parasitism	0.697	0.413	16
interspecies interaction between organisms	0.697	0.413	16
regulation of immune response	0.693	0.296	14
regulation of immune system process	0.691	0.408	26
positive regulation of immune response	0.690	0.244	12
leukocyte differentiation	0.690	0.304	18
regulation of cytokine production	0.675	0.318	12
defense response	0.671	0.570	24
negative regulation of immune system process	0.670	0.155	10
fat cell differentiation	0.670	0.123	9
viral process	0.670	0.327	15
multi-organism cellular process	0.670	0.327	15
T cell activation	0.668	0.186	11

activation of immune response	0.667	0.225	10
immune response-activating signal transduction	0.667	0.225	10
immune response-regulating signaling pathway	0.667	0.225	10
immune system development	0.667	0.404	27
hemopoiesis	0.667	0.404	27
hematopoietic or lymphoid organ development	0.667	0.404	27
T cell differentiation	0.665	0.178	10
myeloid cell differentiation	0.647	0.234	14
cytokine production	0.644	0.307	13
leukocyte cell-cell adhesion	0.644	0.156	9
single organismal cell-cell adhesion	0.644	0.156	9
regulation of cell-cell adhesion	0.644	0.156	9
cell-cell adhesion	0.644	0.156	9
regulation of leukocyte cell-cell adhesion	0.644	0.156	9
positive regulation of defense response	0.642	0.223	9
regulation of defense response	0.639	0.286	15
blood vessel morphogenesis	0.638	0.135	10
regulation of body fluid levels	0.636	0.246	10
cell surface receptor signaling pathway	0.629	0.556	31
response to external stimulus	0.628	0.492	33
response to organic substance	0.627	0.666	47
lymphocyte activation	0.626	0.182	12
cell adhesion	0.626	0.179	12
biological adhesion	0.626	0.179	12
multi-organism process	0.620	0.487	31
lymphocyte differentiation	0.620	0.174	11
membrane-enclosed lumen	0.620	0.858	76
nuclear lumen	0.620	0.858	76
organelle lumen	0.620	0.858	76
intracellular organelle lumen	0.620	0.858	76
leukocyte activation	0.613	0.196	14
regulation of cell adhesion	0.612	0.167	11
regulation of chromosome organization	0.609	0.109	9
cytosol	0.607	0.550	33
multicellular organismal process	0.606	0.825	75
single-multicellular organism process	0.606	0.825	75
positive regulation of multicellular organismal process	0.602	0.458	36

Appendix D

Chapter 3 supplementary data

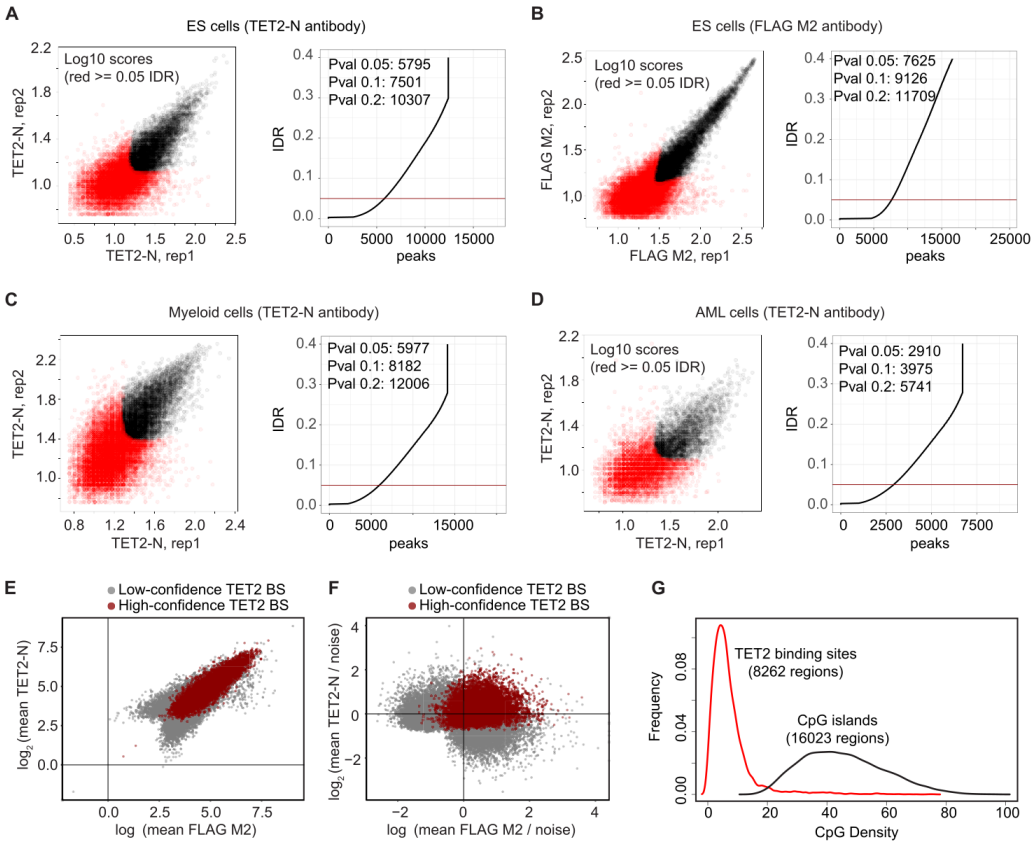

Figure D.1: **Quality controls for TET2 ChIP-seq.** (A-B) IDR analysis of replicate ES TET2 ChIP samples generated with (A) TET2-N, or FLAG M2 (B) antibodies. (C) IDR analysis of replicate TET2 ChIP samples from myeloid hematopoietic cells. (D) IDR analysis of replicate TET2 ChIP samples from AML cells. The number of reproducible peaks at different confidence levels are shown. (E) Scatterplot showing correlation of log2 transformed normalized average TET2 ChIP-seq read counts in peaks in ES cells from TET2-N and FLAG M2 ChIP experiments. High-confidence (red dots) TET2 binding sites are marked. (F) Same as (E), but for log2 transformed fold changes of TET2 ChIP-seq reads over the background signal (noise) measured in Tet2 knockout cells (for TET2-N) or empty cells (for FLAG M2). (G) Plot showing density of CpG sites in high-confidence TET2 binding sites as well as CpG islands in the mouse genome. For the TET2 bindign sites the CpG density is significantly lower as compared to CpG islands. This figure was produced by myself and is published in the (Rasmussen et al., 2019).

References

Ambrosini, G., Groux, R., and Bucher, P. (2018). PWMScan: A fast tool for scanning entire genomes with a position-specific weight matrix. Bioinformatics *34*, 2483–2484.

Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE blacklist: Identification of problematic regions of the genome. Scientific Reports *9*.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data.

Arendt, D., Musser, J.M., Baker, C.V.H., Bergman, A., Cepko, C., Erwin, D.H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M.D., et al. (2016). The origin and evolution of cell types. Nature Reviews Genetics *17*, 744–757.

Arvaniti, E., Ntoufa, S., Papakonstantinou, N., Touloumenidou, T., Laoutaris, N., Anagnostopoulos, A., Lamnissou, K., Caligaris-Cappio, F., Stamatopoulos, K., Ghia, P., et al. (2011). Toll-like receptor signaling pathway in chronic lymphocytic leukemia: Distinct gene expression profiles of potential pathogenic significance in specific subsets of patients. Haematologica *96*, 1644–1652.

Azofeifa, J.G., Allen, M.A., Hendrix, J.R., Read, T., Rubin, J.D., and Dowell, R.D. (2018). Enhancer RNA profiling predicts transcription factor activity. Genome Research *28*, 334–344.

Baek, S., Goldstein, I., and Hager, G.L. (2017). Bivariate genomic footprinting detects changes in transcription factor activity. Cell Reports *19*, 1710–1722.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: Tools for motif discovery and searching. Nucleic Acids Research *37*, W202–W208.

Beekman, R., Chapaprieta, V., Russiñol, N., Vilarrasa-Blasi, R., Verdaguer-Dot, N., Martens, J.H.A., Duran-Ferrer, M., Kulis, M., Serra, F., Javierre, B.M., et al. (2018). The reference epigenome and regulatory chromatin landscape of chronic lymphocytic

leukemia. Nature Medicine *24*, 868–880.

Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research *40*, e72.

Benjamini, Y., Y Hochberg - Journal of the royal statistical society. Series B, 1995, and al. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 289–300.

Berest, I., Arnold, C., Reyes-Palomares, A., Palla, G., Rasmussen, K.D., Giles, H., Bruch, P.-M., Huber, W., Dietrich, S., Helin, K., et al. (2019). Quantification of differential transcription factor activity and multiomics-based classification into activators and repressors: diffTF. Cell Reports *29*, 3147–3159.e12.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature Biotechnology *24*, 1429–1435.

Bocker, M.T., Tuorto, F., Raddatz, G., Musch, T., Yang, F.-C., Xu, M., Lyko, F., and Breiling, A. (2012). Hydroxylation of 5-methylcytosine by TET2 maintains the active state of the mammalian HOXA cluster. Nature Communications *3*.

Boettiger, A.N., Bintu, B., Moffitt, J.R., Wang, S., Beliveau, B.J., Fudenberg, G., Imakaev, M., Mirny, L.A., Wu, C.-t., and Zhuang, X. (2016). Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature *529*, 418–422.

Boileau, M., Shirinian, M., Gayden, T., Harutyunyan, A.S., Chen, C.C.L., Mikael, L.G., Duncan, H.M., Neumann, A.L., Arreba-Tutusaus, P., Jay, N.D., et al. (2019). Mutant h3 histones drive human pre-leukemic hematopoietic stem cell expansion and promote leukemic aggressiveness. Nature Communications *10*.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics *30*, 2114–2120.

Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I.M., Herr, W., Hernandez, N., Delorenzi, M., et al. (2014). Quantifying ChIP-seq data: A spiking method providing an internal reference for sample-to-sample normalization. Genome Research *24*, 1157–1168.

Boorsma, A., Lu, X.-J., Zakrzewska, A., Klis, F.M., and Bussemaker, H.J. (2008). In-

ferring condition-specific modulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression. Plos One *3*, e3112.

Buchberger, Reis, Lu, and Posnien (2019). Cloudy with a chance of insights: Context dependent gene regulation and implications for evolutionary studies. Genes *10*, 492.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods *10*, 1213–1218.

Bussemaker, H.J., Li, H., and Siggia, E.D. (2001). Regulatory element detection using correlation with expression. Nature Genetics *27*, 167–171.

Calviello, A.K., Hirsekorn, A., Wurmus, R., Yusuf, D., and Ohler, U. (2019). Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. Genome Biology *20*.

Castro-Mondragon, J.A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., and Helden, J. van (2017). RSAT matrix-clustering: Dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Research *45*, e119.

Chen, L.-L., Lin, H.-P., Zhou, W.-J., He, C.-X., Zhang, Z.-Y., Cheng, Z.-L., Song, J.-B., Liu, P., Chen, X.-Y., Xia, Y.-K., et al. (2018). SNIP1 recruits TET2 to regulate c-MYC target genes and cellular DNA damage response. Cell Reports *25*, 1485–1500.e4.

Cheng, H., Zheng, Z., and Cheng, T. (2019). New paradigms on hematopoietic stem cell differentiation. Protein & Cell *11*, 34–44.

Chevrier, S., Emslie, D., Shi, W., Kratina, T., Wellard, C., Karnowski, A., Erikci, E., Smyth, G.K., Chowdhury, K., Tarlinton, D., et al. (2014). The BTB-ZF transcription factor zbtb20 is driven by irf4 to promote plasma cell differentiation and longevity. The Journal of Experimental Medicine *211*, 827–840.

Cirovic, B., Schönheit, J., Kowenz-Leutz, E., Ivanovska, J., Klement, C., Pronina, N., Bégay, V., and Leutz, A. (2017). C/EBP-induced transdifferentiation reveals granulocyte-macrophage precursor-like plasticity of b cells. Stem Cell Reports *8*, 346–359.

Cordoba, R., Sanchez-Beato, M., Herreros, B., Domenech, E., Garcia-Marco, J., Garcia, J.-F., Martinez-Lopez, J., Rodriguez, A., Garcia-Raso, A., Llamas, P., et al. (2015). Two distinct molecular subtypes of chronic lymphocytic leukemia give new

insights on the pathogenesis of the disease and identify novel therapeutic targets. Leukemia & Lymphoma *57*, 134–142.

Cortini, R., and Filion, G.J. (2018). Theoretical principles of transcription factor traffic on folded chromatin. Nature Communications *9*.

Coscia, M., Pantaleoni, F., Riganti, C., Vitale, C., Rigoni, M., Peola, S., Castella, B., Foglietta, M., Griggio, V., Drandi, D., et al. (2011). IGHV unmutated CLL b cells are more prone to spontaneous apoptosis and subject to environmental prosurvival signals than mutated CLL b cells. Leukemia *25*, 828–837.

Damm, F., Mylonas, E., Cosson, A., Yoshida, K., Valle, V.D., Mouly, E., Diop, M., Scourzic, L., Shiraishi, Y., Chiba, K., et al. (2014). Acquired initiating mutations in early hematopoietic cells of CLL patients. Cancer Discovery *4*, 1088–1101.

D'Annibale, S., Kim, J., Magliozzi, R., Low, T.Y., Mohammed, S., Heck, A.J.R., and Guardavaccaro, D. (2014). Proteasome-dependent degradation of transcription factor activating enhancer-binding protein 4 (TFAP4) controls mitotic division. The Journal of Biological Chemistry *289*, 7730–7737.

Dawlaty, M.M., Breiling, A., Le, T., Barrasa, M.I., Raddatz, G., Gao, Q., Powell, B.E., Cheng, A.W., Faull, K.F., Lyko, F., et al. (2014). Loss of tet enzymes compromises proper differentiation of embryonic stem cells. Developmental Cell *29*, 102–111.

Delgado, M.D., and Leon, J. (2010). Myc roles in hematopoiesis and leukemia. Genes & Cancer *1*, 605–616.

Delhommeau, F., Dupont, S., Valle, V.D., James, C., Trannoy, S., Massé, A., Kosmider, O., Couedic, J.-P.L., Robert, F., Alberdi, A., et al. (2009). Mutation inTET2in myeloid cancers. New England Journal of Medicine *360*, 2289–2301.

Dietrich, S., Oleś, M., Lu, J., Sellner, L., Anders, S., Velten, B., Wu, B., Hüllein, J., Silva Liberio, M. da, Walther, T., et al. (2018). Drug-perturbation-based stratification of blood cancer. The Journal of Clinical Investigation *128*, 427–445.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Döhner, H., Stilgenbauer, S., Benner, A., Leupolt, E., Kröber, A., Bullinger, L., Döhner, K., Bentz, M., and Lichter, P. (2000). Genomic aberrations and survival in chronic lymphocytic leukemia. New England Journal of Medicine *343*, 1910–1916.

Ebrahimi, A., Sevinç, K., Sevinç, G.G., Cribbs, A.P., Philpott, M., Uyulur, F., Morova, T., Dunford, J.E., Göklemez, S., Arı, et al. (2019). Bromodomain inhibition of the coactivators CBP/EP300 facilitate cellular reprogramming. Nature Chemical Biology *15*, 519–528.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association *102*, 93–103.

El-Athman, R., and Relógio, A. (2018). Escaping circadian regulation: An emerging hallmark of cancer? Cell Systems *6*, 266–267.

Ernst, J., and Kellis, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. Nature Methods *9*, 215–216.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research *47*, D766–D773.

Fudenberg, G., and Pollard, K.S. (2019). Chromatin features constrain structural variation across evolutionary timescales. Proceedings of the National Academy of Sciences *116*, 2175–2180.

Furman, R.R., Sharman, J.P., Coutre, S.E., Cheson, B.D., Pagel, J.M., Hillmen, P., Barrientos, J.C., Zelenetz, A.D., Kipps, T.J., Flinn, I., et al. (2014). Idelalisib and rituximab in relapsed chronic lymphocytic leukemia. The New England Journal of Medicine *370*, 997–1007.

Galas, D.J., and Schmitz, A. (1978). DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Research *5*, 3157–3170.

Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Research *29*, 1363–1375.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. Nature Genetics *42*, 255–259.

Ghamlouch, H., Darwiche, W., Hodroge, A., Ouled-Haddou, H., Dupont, S., Singh, A.R., Guignant, C., Trudel, S., Royer, B., Gubler, B., et al. (2015). Factors involved in CLL pathogenesis and cell survival are disrupted by differentiation of CLL b-cells

into antibody-secreting cells. Oncotarget *6*, 18484–18503.

Griffon, A., Barbier, Q., Dalino, J., Helden, J. van, Spicuglia, S., and Ballester, B. (2015). Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Nucleic Acids Research *43*, e27.

Grubert, F., Zaugg, J.B., Kasowski, M., Ursu, O., Spacek, D.V., Martin, A.R., Greenside, P., Srivas, R., Phanstiel, D.H., Pekowska, A., et al. (2015). Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell *162*, 1051–1065.

Gu, T., Lin, X., Cullen, S.M., Luo, M., Jeong, M., Estecio, M., Shen, J., Hardikar, S., Sun, D., Su, J., et al. (2018). DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. Genome Biology *19*.

Guièze, R., and Wu, C.J. (2015). Genomic and epigenomic heterogeneity in chronic lymphocytic leukemia. Blood *126*, 445–453.

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C.Y., Lee, M., Kim, E., et al. (2018). TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Research *46*, D380–D386.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next generation. Cell *144*, 646–674.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: The reference human genome annotation for the ENCODE project. Genome Research *22*, 1760–1774.

Havelange, V., Pekarsky, Y., Nakamura, T., Palamarchuk, A., Alder, H., Rassenti, L., Kipps, T., and Croce, C.M. (2011). IRF4 mutations in chronic lymphocytic leukemia. Blood *118*, 2827–2829.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. Molecular Cell *38*, 576–589.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A phase separation model for transcriptional control. Cell *169*, 13–23.

Hon, G.C., Song, C.-X., Du, T., Jin, F., Selvaraj, S., Lee, A.Y., Yen, C.-a., Ye, Z., Mao, S.-Q., Wang, B.-A., et al. (2014). 5mC oxidation by tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation. Molecular Cell *56*, 286–297.

Hou, N., Ye, B., Li, X., Margulies, K.B., Xu, H., Wang, X., and Li, F. (2016). Transcription factor 7-like 2 mediates canonical wnt/β-catenin signaling and c-myc upregulation in heart failure. Circulation: Heart Failure *9*.

Huang, D., Petrykowska, H.M., Miller, B.F., Elnitski, L., and Ovcharenko, I. (2019). Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. Genome Research *29*, 657–667.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with bioconductor. Nature Methods *12*, 115–121.

Hunt, R.W., Mathelier, A., Peso, L. del, and Wasserman, W.W. (2014). Improving analysis of transcription factor binding sites within ChIP-seq data based on topological motif enrichment. BMC Genomics *15*, 472.

Jayaram, N., Usvyat, D., and Martin, A.C.R. (2016). Evaluating tools for transcription factor binding site prediction. BMC Bioinformatics *17*.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497–1502.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpaa, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Research *20*, 861–873.

Kasim, V., Xie, Y.-D., Wang, H.-M., Huang, C., Yan, X.-S., Nian, W.-Q., Zheng, X.-D., Miyagishi, M., and Wu, S.-R. (2017). Transcription factor yin yang 2 is a novel regulator of the p53/p21 axis. Oncotarget *8*, 54694–54707.

Kauffman, E.C., Ricketts, C.J., Rais-Bahrami, S., Yang, Y., Merino, M.J., Bottaro, D.P., Srinivasan, R., and Linehan, W.M. (2014). Molecular genetics and cellular features of TFE3 and TFEB fusion kidney cancers. Nature Reviews. Urology *11*, 465–475.

Kempfer, R., and Pombo, A. (2019). Methods for mapping 3D chromosome architecture. Nature Reviews Genetics.

Kern, D., Regl, G., Hofbauer, S.W., Altenhofer, P., Achatz, G., Dlugosz, A., Schnidar, H., Greil, R., Hartmann, T.N., and Aberger, F. (2015). Hedgehog/GLI and PI3K signaling in the initiation and maintenance of chronic lymphocytic leukemia. Oncogene *34*, 5341–5351.

Kikushige, Y., Ishikawa, F., Miyamoto, T., Shima, T., Urata, S., Yoshimoto, G., Mori, Y., Iino, T., Yamauchi, T., Eto, T., et al. (2011). Self-renewing hematopoietic stem cell is the primary target in pathogenesis of human chronic lymphocytic leukemia. Cancer Cell *20*, 246–259.

Kim, R., Emi, M., and Tanabe, K. (2007). Cancer immunoediting from immune surveillance to immune escape. Immunology *121*, 1–14.

Klein, D.C., and Hainer, S.J. (2019). Genomic methods in profiling DNA accessibility and factor localization. Chromosome Research.

Komatsu, M., Kurokawa, H., Waguri, S., Taguchi, K., Kobayashi, A., Ichimura, Y., Sou, Y.-S., Ueno, I., Sakamoto, A., Tong, K.I., et al. (2010). The selective autophagy substrate p62 activates the stress responsive transcription factor nrf2 through inactivation of keap1. Nature Cell Biology *12*, 213–223.

Köster, J., and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Kreslavsky, T., Vilagos, B., Tagoh, H., Poliakova, D.K., Schwickert, T.A., Wöhner, M., Jaritz, M., Weiss, S., Taneja, R., Rossner, M.J., et al. (2017). Essential role for the transcription factor bhlhe41 in regulating the development, self-renewal and BCR repertoire of b-1a cells. Nature Immunology *18*, 442–455.

Kribelbauer, J.F., Rastogi, C., Bussemaker, H.J., and Mann, R.S. (2019). Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. Annual Review of Cell and Developmental Biology *35*, 357–379.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Soboleva, A.V., Kasianov, A.S., Ashoor, H., Ba-alawi, W., Bajic, V.B., Medvedeva, Y.A., Kolpakov, F.A., et al. (2015). HOCOMOCO: Expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Research *44*, D116–D125.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Lai, X., Stigliani, A., Vachon, G., Carles, C., Smaczniak, C., Zubieta, C., Kaufmann, K., and Parcy, F. (2019). Building transcription factor binding site models to understand gene regulation in plants. Molecular Plant *12*, 743–763.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The human transcription factors. Cell *172*, 650–665.

Lambert, S.A., Yang, A.W.H., Sasse, A., Cowley, G., Albu, M., Caddick, M.X., Morris, Q.D., Weirauch, M.T., and Hughes, T.R. (2019). Similarity regression predicts evolution of transcription factor sequence specificity. Nature Genetics *51*, 981–989.

Landau, D.A., Tausch, E., Taylor-Weiner, A.N., Stewart, C., Reiter, J.G., Bahlo, J., Kluth, S., Bozic, I., Lawrence, M., Böttcher, S., et al. (2015). Mutations driving CLL and their evolution in progression and relapse. Nature *526*, 525–530.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research *22*, 1813–1831.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with bowtie 2. Nature Methods *9*, 357–359.

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., et al. (2014). Immunogenetics. Chromatin state dynamics during blood formation. Science *345*, 943–949.

Lee, B.H., Tothova, Z., Levine, R.L., Anderson, K., Buza-Vidas, N., Cullen, D.E., McDowell, E.P., Adelsperger, J., Fröhling, S., Huntly, B.J.P., et al. (2007). FLT3 mutations confer enhanced proliferation and survival properties to multipotent progenitors in a murine model of chronic myelomonocytic leukemia. Cancer Cell *12*, 367–380.

Li, Y., Hu, M., and Shen, Y. (2018). Gene regulation in the 3D genome. Human Molecular Genetics *27*, R228–R233.

Li, Y.-J., Sun, L., Shi, Y., Wang, G., Wang, X., Dunn, S.E., Iorio, C., Screaton, R.A., and Spaner, D.E. (2017). PPAR-delta promotes survival of chronic lymphocytic leukemia cells in energetically unfavorable conditions. Leukemia *31*, 1905–1914.

Liao, Y., Smyth, G.K., and Shi, W. (2019). The r package rsubread is easier, faster,

cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Research *47*, e47–e47.

Lieberman-Aiden, E., Berkum, N.L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lio, C.-W., Zhang, J., González-Avalos, E., Hogan, P.G., Chang, X., and Rao, A. (2016). Tet2 and tet3 cooperate with b-lineage transcription factors to regulate DNA modification and chromatin accessibility. eLife *5*.

Lio, C.-W.J., Yuita, H., and Rao, A. (2019). Dysregulation of the TET family of epigenetic regulators in lymphoid and myeloid malignancies. Blood *134*, 1487–1497.

Liu, L., Jin, G., and Zhou, X. (2015). Modeling the relationship of epigenetic modifications to transcription factor binding. Nucleic Acids Research *43*, 3873–3885.

Liu, Z., Lee, J., Krummey, S., Lu, W., Cai, H., and Lenardo, M.J. (2011). The kinase LRRK2 is a regulator of the transcription factor NFAT that modulates the severity of inflammatory bowel disease. Nature Immunology *12*, 1063–1070.

Long, H.K., Blackledge, N.P., and Klose, R.J. (2013). ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. Biochemical Society Transactions *41*, 727–740.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Lu, F., Liu, Y., Jiang, L., Yamaguchi, S., and Zhang, Y. (2014). Role of tet proteins in enhancer activity and telomere elongation. Genes & Development *28*, 2103–2119.

Lun, A.T.L., and Smyth, G.K. (2016). Csaw: A bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Research *44*, e45.

Mann, I.K., Chatterjee, R., Zhao, J., He, X., Weirauch, M.T., Hughes, T.R., and Vinson, C. (2013). CG methylated microarrays identify a novel methylated sequence bound by the CEBPB ATF4 heterodimer that is active in vivo. Genome Research *23*, 988–997.

Massari, M.E., and Murre, C. (2000). Helix-loop-helix proteins: Regulators of transcription in eucaryotic organisms. Molecular and Cellular Biology *20*, 429–440.

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.-Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2016). JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Research *44*, D110–5.

Mayran, A., and Drouin, J. (2018). Pioneer transcription factors shape the epigenetic landscape. Journal of Biological Chemistry *293*, 13795–13804.

Mayran, A., Sochodolsky, K., Khetchoumian, K., Harris, J., Gauthier, Y., Bemmo, A., Balsalobre, A., and Drouin, J. (2019). Pioneer and nonpioneer factor cooperation drives lineage specific chromatin opening. Nature Communications *10*.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research *20*, 1297–1303.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature Biotechnology *28*, 495–501.

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J.A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., et al. (2015). RSAT 2015: Regulatory sequence analysis tools. Nucleic Acids Research *43*, W50–6.

Mezger, A., Klemm, S., Mann, I., Brower, K., Mir, A., Bostick, M., Farmer, A., Fordyce, P., Linnarsson, S., and Greenleaf, W. (2018). High-throughput chromatin accessibility profiling at single-cell resolution. Nature Communications *9*, 3647.

Minami, Y., Oishi, I., Endo, M., and Nishita, M. (2010). Ror-family receptor tyrosine kinases in noncanonical wnt signaling: Their implications in developmental morphogenesis and human diseases. Developmental Dynamics *239*, 1–15.

Moran-Crusio, K., Reavie, L., Shih, A., Abdel-Wahab, O., Ndiaye-Lobry, D., Lobry, C., Figueroa, M.E., Vasanthakumar, A., Patel, J., Zhao, X., et al. (2011). Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. Cancer Cell *20*, 11–24.

Movva, R., Greenside, P., Marinov, G.K., Nair, S., Shrikumar, A., and Kundaje, A.

(2019). Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. PLOS ONE *14*, e0218073.

Neu, K.E., and Wilson, P.C. (2016). Taking the broad view on b cell affinity maturation. Immunity *44*, 518–520.

Oakes, C.C., Seifert, M., Assenov, Y., Gu, L., Przekopowitz, M., Ruppert, A.S., Wang, Q., Imbusch, C.D., Serva, A., Koser, S.D., et al. (2016). DNA methylation dynamics during b cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. Nature Genetics *48*, 253–264.

Ou, H.D., Phan, S., Deerinck, T.J., Thor, A., Ellisman, M.H., and O'Shea, C.C. (2017). ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. Science *357*, eaag0025.

Piper, J., Assi, S.A., Cauchy, P., Ladroue, C., Cockerill, P.N., Bonifer, C., and Ott, S. (2015). Wellington-bootstrap: Differential DNase-seq footprinting identifies cell-type determining transcription factors. BMC Genomics *16*, 1000.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Quivoron, C., Couronné, L., Valle, V.D., Lopez, C.K., Plo, I., Wagner-Ballon, O., Cruzeiro, M.D., Delhommeau, F., Arnulf, B., Stern, M.-H., et al. (2011). TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. Cancer Cell *20*, 25–38.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: A flexible platform for exploring deep-sequencing data. Nucleic Acids Research *42*, W187–91.

Rana, S., Munawar, M., Shahid, A., Malik, M., Ullah, H., Fatima, W., Mohsin, S., and Mahmood, S. (2013). Deregulated expression of circadian clock and clock-controlled cell cycle genes in chronic lymphocytic leukemia. Molecular Biology Reports *41*, 95–103.

Rasmussen, K.D., and Helin, K. (2016). Role of TET enzymes in DNA methylation, development, and cancer. Genes & Development *30*, 733–750.

Rasmussen, K.D., Jia, G., Johansen, J.V., Pedersen, M.T., Rapin, N., Bagger, F.O., Porse, B.T., Bernard, O.A., Christensen, J., and Helin, K. (2015). Loss ofTET2in

hematopoietic cells leads to DNA hypermethylation of active enhancers and induction of leukemogenesis. Genes & Development *29*, 910–922.

Rasmussen, K.D., Berest, I., Kessler, S., Nishimura, K., Simón-Carrasco, L., Vassiliou, G.S., Pedersen, M.T., Christensen, J., Zaugg, J.B., and Helin, K. (2019). TET2 binding to enhancers facilitates transcription factor recruitment in hematopoietic cells. Genome Research *29*, 564–575.

Rendeiro, A.F., Schmidl, C., Strefford, J.C., Walewska, R., Davis, Z., Farlik, M., Oscier, D., and Bock, C. (2016). Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. Nature Communications *7*, 11938.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research *43*, e47.

Ross, S.E., and Bogdanovic, O. (2019). TET enzymes, DNA demethylation and pluripotency. Biochemical Society Transactions *47*, 875–885.

Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature *481*, 389–393.

Saito, T., and Rehmsmeier, M. (2017). Precrec: Fast and accurate precision-recall and ROC curve calculations in r. Bioinformatics *33*, 145–147.

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proceedings of the National Academy of Sciences *112*, E6456–E6465.

Sardina, J.L., Collombet, S., Tian, T.V., Gómez, A., Stefano, B.D., Berenguer, C., Brumbaugh, J., Stadhouders, R., Segura-Morales, C., Gut, M., et al. (2018). Transcription factors drive tet2-mediated enhancer demethylation to reprogram cell fate. Cell Stem Cell *23*, 727–741.e9.

Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nature Methods *14*, 975–978.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887–898.

Scourzic, L., Mouly, E., and Bernard, O.A. (2015). TET proteins and the control of cytosine demethylation in cancer. Genome Medicine *7*, 9.

Seré, K.M., Lin, Q., Felker, P., Rehage, N., Klisch, T., Ortseifer, I., Hieronymus, T., Rose-John, S., and Zenke, M. (2012). Dendritic cell lineage commitment is instructed by distinct cytokine signals. European Journal of Cell Biology *91*, 515–523.

Slager, S.L., Caporaso, N.E., Sanjose, S. de, and Goldin, L.R. (2013). Genetic susceptibility to chronic lymphocytic leukemia. Seminars in Hematology *50*, 296–302.

Song, L., and Crawford, G.E. (2010). DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harbor Protocols *2010*, pdb.prot5384–pdb.prot5384.

Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of proteinDNA interactions. Nature Reviews Genetics *11*, 751–760.

Strati, P., and Shanafelt, T.D. (2015). Monoclonal b-cell lymphocytosis and early-stage chronic lymphocytic leukemia: Diagnosis, natural history, and risk stratification. Blood *126*, 454–462.

Tripodi, I., Allen, M., and Dowell, R. (2018). Detecting differential transcription factor activity from ATAC-seq data. Molecules *23*, 1136.

Vassiliou, G.S., Cooper, J.L., Rad, R., Li, J., Rice, S., Uren, A., Rad, L., Ellis, P., Andrews, R., Banerjee, R., et al. (2011). Mutant nucleophosmin and cooperating pathways drive leukemia initiation and progression in mice. Nature Genetics *43*, 470–475.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Research *22*, 1798–1812.

Wang, Y., Yabuuchi, A., McKinney-Freeman, S., Ducharme, D.M.K., Ray, M.K., Chawengsaksophak, K., Archer, T.K., and Daley, G.Q. (2008). Cdx gene deficiency compromises embryonic hematopoiesis in the mouse. Proceedings of the National

Academy of Sciences *105*, 7756–7761.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics *5*, 276–287.

Wei, B., Jolma, A., Sahu, B., Orre, L.M., Zhong, F., Zhu, F., Kivioja, T., Sur, I., Lehtiö, J., Taipale, M., et al. (2018). A protein activity assay to measure global transcription factor activity reveals determinants of chromatin accessibility. Nature Biotechnology *36*, 521–529.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

Weissmann, S., Alpermann, T., Grossmann, V., Kowarsch, A., Nadarajah, N., Eder, C., Dicker, F., Fasan, A., Haferlach, C., Haferlach, T., et al. (2011). Landscape of TET2 mutations in acute myeloid leukemia. Leukemia *26*, 934–942.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Wong, C., Chen, C., Wu, Q., Liu, Y., and Zheng, P. (2014). A critical role for the regulated wntMyc pathway in naive t cell survival. The Journal of Immunology *194*, 158–167.

Wu, F., Olson, B.G., and Yao, J. (2016). DamID-seq: Genome-wide mapping of protein-DNA interactions by high throughput sequencing of adenine-methylated DNA fragments. Journal of Visualized Experiments.

Yamazaki, J., Jelinek, J., Lu, Y., Cesaroni, M., Madzo, J., Neumann, F., He, R., Taby, R., Vasanthakumar, A., Macrae, T., et al. (2015). TET2 mutations affect non-CpG island DNA methylation at enhancers and transcription factor-binding sites in chronic myelomonocytic leukemia. Cancer Research *75*, 2833–2843.

Yang, X., Wong, M.P.M., and Ng, R.K. (2019). Aberrant DNA methylation in acute myeloid leukemia and its clinical implications. International Journal of Molecular Sciences *20*, 4576.

Yeomans, A., Thirdborough, S.M., Valle-Argos, B., Linley, A., Krysov, S., Hidalgo, M.S., Leonard, E., Ishfaq, M., Wagner, S.D., Willis, A.E., et al. (2016). Engagement of the

b-cell receptor of chronic lymphocytic leukemia cells drives global and MYC-specific mRNA translation. Blood *127*, 449–457.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.

Yu, G., Wang, L.-G., and He, Q.-Y. (2015). ChIPseeker: An r/bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics *31*, 2382–2383.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-seq (MACS). Genome Biology *9*, R137.

Zhang, Y.W., Wang, Z., Xie, W., Cai, Y., Xia, L., Easwaran, H., Luo, J., Yen, R.-W.C., Li, Y., and Baylin, S.B. (2017). Acetylation enhances TET2 function in protecting against abnormal DNA methylation during oxidative stress. Molecular Cell *65*, 323–335.

Zhou, Y., Li, Y.-S., Bandi, S.R., Tang, L., Shinton, S.A., Hayakawa, K., and Hardy, R.R. (2015). Lin28b promotes fetal b lymphopoiesis through the transcription factor arid3a. The Journal of Experimental Medicine *212*, 569–580.

Zhu, H., Wang, G., and Qian, J. (2016). Transcription factors as readers and effectors of DNA methylation. Nature Reviews Genetics *17*, 551–565.