

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PhD)

**Characterization of labelled regulatory elements in embryonic stem cells and macrophages using quantitative and qualitative methods**

by **Attila Horváth**

UNIVERSITY OF DEBRECEN  
DOCTORAL SCHOOL OF MOLECULAR CELL AND IMMUNE BIOLOGY

DEBRECEN, 2019

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PhD)

**Characterization of labelled regulatory elements in embryonic stem cells and macrophages using quantitative and qualitative methods**

by **Attila Horváth**

Supervisor: Prof. Dr. László Nagy

Co-Supervisor: Dr. Benedek Nagy



UNIVERSITY OF DEBRECEN  
DOCTORAL SCHOOL OF MOLECULAR CELL AND IMMUNE BIOLOGY

DEBRECEN, 2019

## TABLE OF CONTENT

<b>1. ABBREVIATIONS.....</b>	<b>6</b>
<b>2. INTRODUCTION.....</b>	<b>10</b>
Transcription regulation in Eukaryotes.....	10
The concept of enhancer .....	12
Identification of enhancer regions.....	13
Histone modifications.....	13
Classification of enhancer states based on epigenetic signatures.....	14
<b>3. LITERATURE REVIEW .....</b>	<b>16</b>
<i>3.1. Computational methods for modeling genomics data .....</i>	<i>16</i>
Introduction to Machine learning methods.....	16
Random Forest .....	16
Support Vector Regression.....	17
Network motifs.....	17
Turing machine and Finite State Automata.....	19
<i>3.2. Macrophages: as model system for the study of enhancer formation .....</i>	<i>22</i>
<i>3.3. PU.1 is a master regulator of macrophages.....</i>	<i>23</i>
Regulation of macrophage enhancers in the context of polarizations signals .....	25
<i>3.4. Embryonic stem cells: a model system to study developmental enhancers .....</i>	<i>28</i>
<i>3.5. OCT4 as one of the master regulators of pluripotency .....</i>	<i>29</i>
<b>4. AIMS AND HYPOTHESES.....</b>	<b>31</b>
<b>5. MATERIALS AND METHODS.....</b>	<b>34</b>

Differentiation of bone marrow derived macrophages .....	34
Embryonic stem cell culture.....	34
Ligands and Treatment .....	34
siRNA knockdown.....	34
Microarray analysis.....	35
RT-qPCR .....	35
ChIP-seq .....	36
Western blot analysis .....	36
GRO-seq .....	36
ChIP-seq, GRO-seq and ATAC-seq analyses.....	36
Machine learning.....	38
Data Availability.....	39
Author contributions to the wet-lab experiments.....	39
<b>6. RESULTS .....</b>	<b>40</b>
<i>6.1. Random Forest classification hints the existence of low accessible TF binding sites in macrophages....</i>	<i>40</i>
<i>6.2. PU.1-labelled regulatory elements are widespread in the macrophage genome.....</i>	<i>46</i>
<i>6.3. Key transcriptional regulators of macrophage form labelled regulatory elements.....</i>	<i>51</i>
<i>6.4. The role of PU.1 and IRF8 co-LREs in cellular response to IL-4 .....</i>	<i>54</i>
<i>6.5. IRF8 maintains low accessible chromatin structure at a subset of labelled regulatory elements .....</i>	<i>58</i>
<i>6.6. Labelled regulatory elements are dynamically utilized by macrophage polarization signals.....</i>	<i>60</i>
<i>6.7. Modelling enhancer states using Nondeterministic Finite State Automata .....</i>	<i>69</i>
<i>6.8. OCT4-LREs in the context of RA-induced neurogenesis.....</i>	<i>71</i>
<i>6.9. Modelling OCT4-related transcriptional circuits using network motifs.....</i>	<i>79</i>

7.	DISCUSSION .....	82
8.	SUMMARY .....	86
9.	ÖSSZEFOGLALÁS.....	88
10.	TABLE OF FIGURES.....	90
11.	REFERENCES .....	93
12.	LIST OF KEYWORDS .....	105
13.	KULCSSZAVAK LISTÁJA .....	106
14.	ACKNOWLEDGEMENTS .....	107
15.	APPENDIX.....	108

## 1. ABBREVIATIONS

AP-1	activator protein 1
APC	antigen presenting cell
Arg1	arginase 1
AUC	area under the curve
BMDM	bone marrow-derived macrophages
BRE	B recognition element
cDNA	Complementary DNA
CEBP	CCAAT/enhancer binding protein
ChIP-seq	chromatin immunoprecipitation sequencing
CTCF	CCCTC-binding Factor
CTF	collaborating transcription factor
DKO	double knock-out
DMEM	Dulbecco's Modified Eagle's Medium
DNA	deoxyribonucleic acid
DPE	downstream promoter element
EC	enhancer cluster
EICE	Ets-IRF composite element
EGR	early growth response protein
eRNA	enhancer RNA
ESC	embryonic stem cell
ETV	E26 transformation-specific
FAIRE	formaldehyde-assisted isolation of regulatory elements

FBS	fetal bovine serum
FC	fold-change
FDR	false discovery rate
FSA	finite state automaton
GB	gene body
GRO-seq	global run-on sequencing
H3K27ac	histone H3 lysine 27 acetylation
H3K4me1	histone H3 lysine 4 mono-methylation
Hbegf	heparin-binding EGF-like growth factor
IGV	integrative genomics viewer
IL12B	interleukin-12B
IL1B	interleukin-1B
IL-4	interleukin-4
Inr	initiator
IRF	interferon regulatory factor
ISRE	interferon stimulated response element
LDTF	lineage determining transcription factor
LIF	leukemia inhibitory factor
lncRNA	long non-coding RNA
LPS	lipopolysaccharide
LRE	labelled regulatory element
M-CSF	macrophage colony-stimulating factor
miRNA	micro RNA

MITF	microphthalmia-associated transcription factor
MNase	Micrococcal nuclease
mRNA	messenger RNA
NF-kB	nuclear factor-kB
NGS	next generation sequencing
OCR	open chromatin region
OCT4	octamer-binding transcription factor 4
PCR	polymerase chain reaction
PMEF	primary mouse embryonic fibroblast
PUER	PU.1 dna-binding domain fused with estrogen receptor ligand binding domain
RA	retinoic acid
RAR	retinoic acid receptor
RARE	retinoic acid response element
RD	read distribution
Retnla	resistin-like alpha
RNA	ribonucleic acid
RNAPII-pS2	RNA polymerase II phosphorylated at serine 2.
RNA-seq	RNA sequencing
ROC	receiver operating characteristics
RPKM	reads per kilobase per million mapped reads
rRNA	ribosomal RNA
RT-qPCR	reverse transcription-quantitative real-time polymerase chain reaction



RUNX	runt-related transcription factor
RXR	retinoid X receptor
SDTFs	signal dependent transcription factors
snRNA	small nuclear RNA
STAT6	signal transducer and activator of transcription
TF	transcription factor
TM	turing machine
TNF	tumor necrosis factor
tRNA	transfer RNA
TRE	12-O-tetradecanoylphorbol-13-acetate (TPA) response element
TSS	transcription start site
UTR	untranslated region
WT	wild type

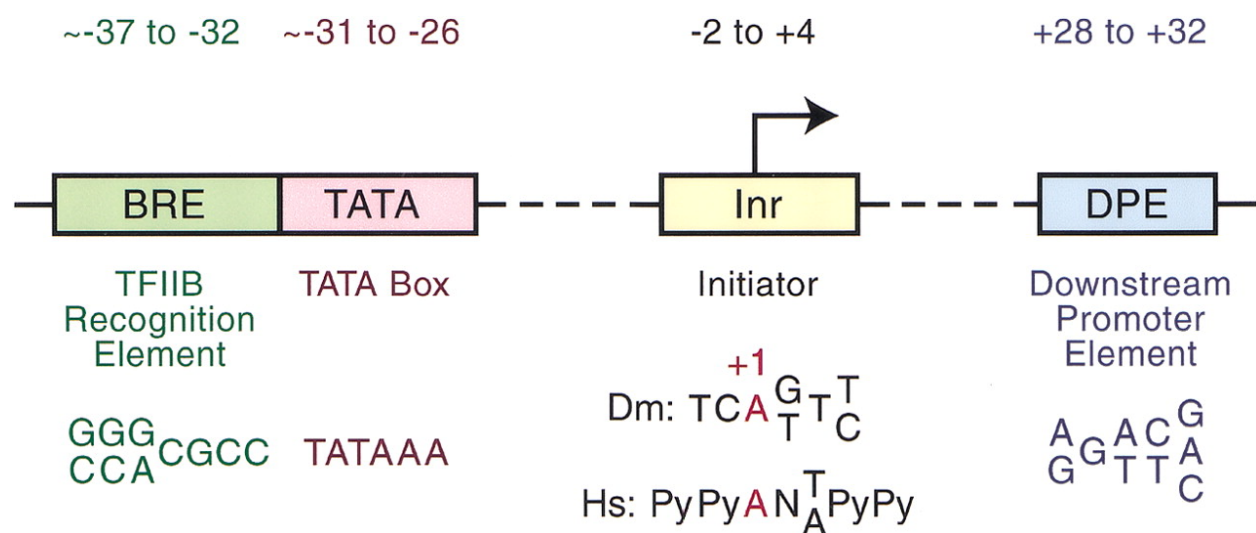
## **2. INTRODUCTION**

Cells are the basic structural and functional units of life. The human body consists of trillions of cells, which can be categorized into around 200 different cell types based on their morphological and/or functional characteristics. With a few notable exceptions, the genomic DNA sequence is identical in the nuclei of all diploid cells within an organism. Although genomic DNA contains all the necessary information for creating any of the organisms' cell types, the fate of an individual cell is defined by the local microenvironment during differentiation. The resulting unique gene expression pattern will define the mature cell's function. The genomic determinants of this unique gene expression profile and its functional role in the cell's response to external stimuli are two of the most exciting topics of the post-genomic era. The cell-type-specific gene expression profile is the result of a series of transcriptional regulatory processes, including transcriptional initiation, elongation, transcript processing and degradation (Cooper, 2000).

### **Transcription regulation in Eukaryotes**

Transcription is the process by which the genetic information stored in the DNA is copied into RNAs. Unlike prokaryotes, eukaryotic cells encode for multiple RNA polymerases, including Pol I, Pol II and Pol III, which synthesize different types of RNAs. Pol I transcribes ribosomal RNAs (rRNAs), Pol III carry out the synthesis of transfer RNAs (tRNAs), small RNAs, the 5S rRNA and most long non-coding RNAs (lncRNAs), while Pol II transcribes mRNAs, small nuclear RNAs (snRNAs) and micro RNA (miRNA) precursors. The correct positioning of Pol II and other proteins, forming the so-called initiation complex, is essential to initiate transcription at the start of the gene body (Cooper, 2000).

A promoter region is defined as the minimal DNA sequence which is sufficient for transcription initiation by an RNA polymerase. One of the best example for promoter-like elements is the TATA box, located ~25 base pairs upstream of coding regions for certain protein- or non-protein-coding (e.g. lncRNA) genes. Other promoter elements have also been discovered: the initiator (Inr), the center of which is located at +1 (one nucleotide downstream of the transcription start site), and the downstream promoter element (DPE), which is centered at +30. These DNA elements provide binding sites for the transcription initiation complex and will direct RNA polymerase where to begin transcription (Butler & Kadonaga, 2002; Krishnamurthy & Hampsey, 2009).



**Figure 1. Various type of promoter elements**

This figure is derived from (Butler & Kadonaga, 2002)

Tissue-specific gene expression variability in multicellular organisms rely on distal regulatory regions. These genomics regions, namely enhancers play a major role in the transcriptional initiation process, modulating the timing and rate of transcription of the associated genes.

## **The concept of enhancer**

Enhancers are short DNA sequences that can induce the expression level of a gene, upon binding one or more regulatory protein(s) called transcription factors (TFs) and looping to the promoter of the regulated gene. An enhancer can regulate multiple genes, and one gene can have multiple enhancers. It has also been described that the promoters of certain genes can act as enhancers of distal genes. In contrast to promoters, enhancers can exert their gene regulatory function in a distance- and direction-independent manner; they do not need to be located near to the transcription start of the gene.

Typically, tens of thousands of enhancer regions are engaged in transcriptional regulation in a given cell type, at a given time point. Enhancers provide gene regulatory flexibility by being able to anchor different sets of transcriptional regulators, which may be present under only certain circumstances, such as after the exposure to a certain external stimulus. An enhancer can exert its gene regulatory function by making a spatial contact with the initiation complex at the promoter region of the regulated gene. In theory, almost any distance in the nucleus can be bypassed by this so-called promoter-enhancer looping. However, the spatial organization of the chromatin, such as the presence of topologically associated domains (TADs), puts a constraint on the possible promoter-enhancer interactions. Moreover, it has been shown that enhancers can regulate different stages of the transcription cycle, including RNAPII recruitment, release of promoter-proximal pause and transcription elongation. Therefore, the detection and characterization of enhancers is critical for understanding how these elements contribute to cell-type specific functions (W. Li, Notani, & Rosenfeld, 2016).

## **Identification of enhancer regions**

Identification of the enhancers of a certain gene is challenging for multiple reasons. First, the regulatory elements can be located at great distances (up to 1 Mb) both upstream or downstream the transcription start site (TSS) of the regulated gene. Moreover, studies have shown that not only the non-coding genome, but intronic and exonic regions can also contain enhancer elements. Surprisingly, gene promoters can also act as distal regulatory elements. Genes often have more than one enhancers and one enhancer can regulate multiple genes. Finally, unlike promoters, there is no general sequence code for enhancers that could be predicted by in silico methods. A widely used strategy for enhancer prediction is to map histone modifications and other epigenomic marks using high-throughput sequencing-based techniques such as chromatin immunoprecipitation sequencing (ChIP-seq) (Pennacchio, Bickmore, Dean, Nobrega, & Bejerano, 2013).

## **Histone modifications**

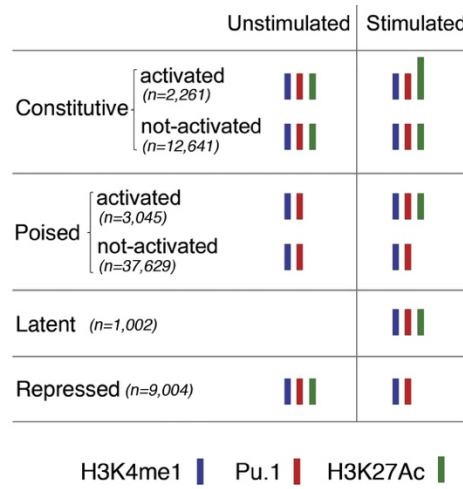
In the eukaryotic nucleus, DNA is packed and ordered by histones in structures called nucleosomes. There are five types of histones: H1, which is involved in higher-order structures of chromatin (known as linker histones), H2A, H2B, H3 and H4 (core histones constituting the so-called histone octamer). Nucleosomes comprise core histones and DNA wrapped around them. In human cells, there is about ~2 meters of DNA, but it is about 90 millimeters of chromatin in a condensed form (Kimura, 2013).

Histones can carry various post-translational modifications referred to as histone marks, which may regulate gene expression by making DNA less or more accessible to transcription. H3 is one of the most extensively modified protein among the core histones. The modifications of H3 can be used to distinguish between hetero or euchromatin states or identify different types

of regulatory elements. Trimethylated H3K4 (H3K4me<sub>3</sub>, three methyl groups to lysine 4 of histone H3) is enriched around transcription start sites (TSSs) of active genes. In contrast, gene bodies of actively transcribed genes are associated with trimethylated H3K36 (H3K36me<sub>3</sub>, three methyl groups to lysine 36 of histone H3). Monomethylated H3K4 (H3K4me<sub>1</sub>, one methyl group to lysine 4 of histone H3) has been shown to co-localize on regulatory regions with pioneer TFs, characterizing general enhancers (irrespective to transcriptional activity), and it is also enriched at promoter regions, although to a lesser extent. Finally, the histone H3K27ac (acetylation at the lysine 27 of histone H3) can be found both at promoters and enhancers, and marks active transcription (Calo & Wysocka, 2013).

### **Classification of enhancer states based on epigenetic signatures**

In macrophage biology, PU.1 and H3K4me<sub>1/2</sub> double-positive regions were considered as enhancer-like regulatory regions and active histone marks such as H3K27ac were used to distinguish between poised and active enhancers (Figure 2) (Creyghton et al., 2010). This classification revealed the existence of latent or de novo enhancers which are not marked either by H3K4me<sub>1/2</sub> or PU.1 binding prior to stimulus but they get activated by signal-dependent transcription factors (SDTFs) (Kaikkonen et al., 2013; Ostuni et al., 2013). Although these studies did not characterize the contribution of chromatin openness which might provide a better mechanistic insight into the different stages of enhancer formation.



**Figure 2. Classification of enhancer states in macrophages.**

This figure is derived from (Ostuni et al., 2013).

### **3. LITERATURE REVIEW**

#### **3.1. Computational methods for modeling genomics data**

##### **Introduction to Machine learning methods**

Machine learning methods are applied artificial intelligence algorithms that are able to recognize and infer during the so-called learning process. These learning algorithms need input data as examples that will be used for pattern recognition during the learning process. Having inferred discriminative features, they can make predictions on a previously unknown data set. Machine learning methods have become the part of our everyday life; they are used for traffic predictions, social media services, email spam filtering, customer support, etc.

Unsupervised machine learning algorithms are designed to infer similarities or associations among the observations without any prior knowledge, therefore they are primarily used for clustering data sets. On the contrary, in the case of supervised machine learning methods there is one or more labels that categorize the training set into distinct groups. In the latter case, the computational task is to find a function of the subset of input variables that discriminate the groups provided.

##### **Random Forest**

Supervised Random Forest is a widely used supervised machine learning method that can perform both classification task (the output variable is a finite set of labels) and regression task (the output variable is continuous). This algorithm combines multitudes of decision trees that are flowchart-like classifiers performing certain tests (logical or numerical) on an attribute at each branch. The final decision will be made based on the summarization of hundreds of decision trees by using some voting scheme such as the majority rule. Random Forest, since it is a linear method, can also estimate the relative contribution of the input variables to the classification.



## **Support Vector Regression**

As discussed above, classification algorithms need input data that are categorized into groups by which the machine learning methods will find the best matching function of discriminative features. However, in real-world applications, the output values to be predicted are often continuous. Support Vector Regressor is a special type of Support Vector Machines that can perform regression tasks. The basic principle of the algorithm is to map the input variables into a high-dimensional feature space by a non-linear (kernel) function and solve the tasks there by a linear function. Although SVR gives better performance, it does not provide the relative contribution of the input variables due to its nonlinear nature.

## **Network motifs**

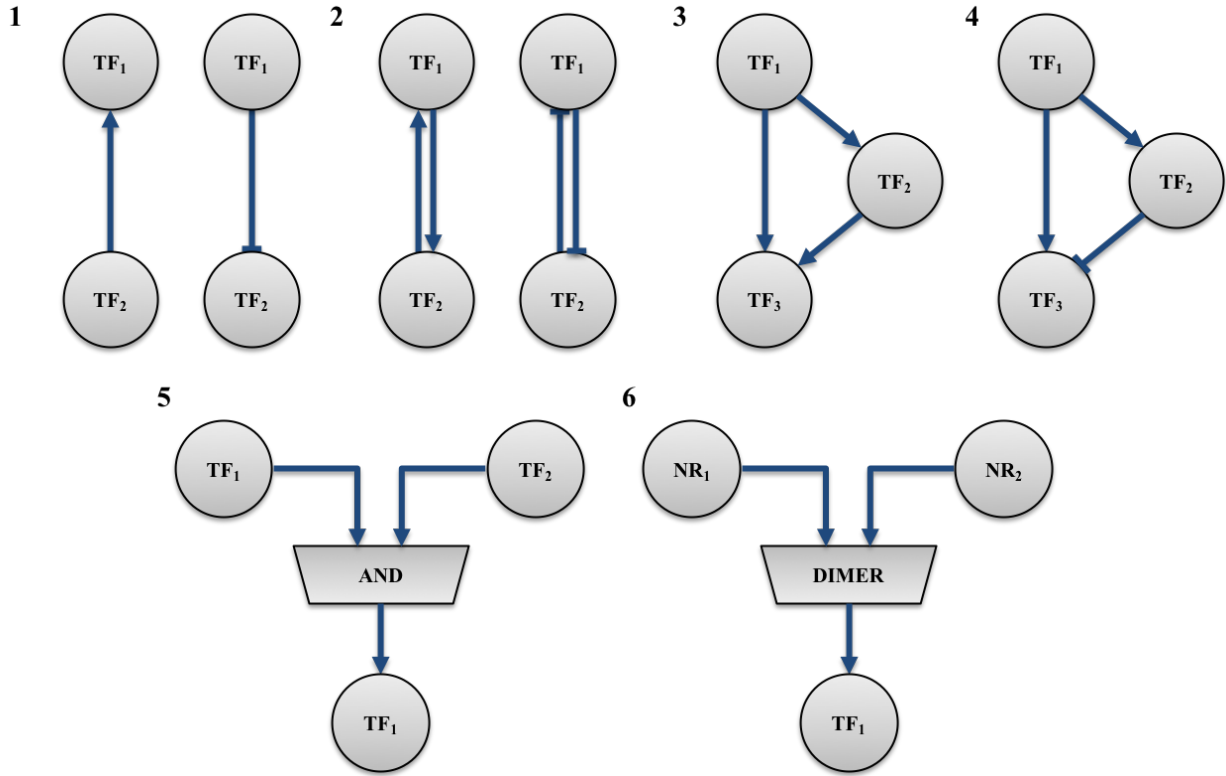
Network motifs are stereotypical sub-graphs that are considered to be functionally distinct units of the whole network. These topological patterns frequently referred to as “simple building blocks of complex networks” are examined in various fields, such as software design pattern diagrams, ecological, social and molecular networks. In biology, network motifs have been proven excellent tools to describe interactions among cells (e.g. neuronal networks), proteins (protein–protein interactions) or genes (transcription or gene regulation networks) (Wong, Baur, Quader, & Huang, 2012).

At the level of transcription, several network motifs have been characterized that perform partially independent tasks in the network. The simplest case is when the transcription factor  $TF_1$  up-regulates or down-regulates the level of  $TF_2$  (Figure 3). In the second type of networks, the so-called mutual activation/repression network motif, both  $TF_1$  and  $TF_2$  repress each other's expression level. Thus, these types of networks have only two stable states. In mutual activation network, both TFs are turned on or turned off. While in the case of mutual repression, the system

can only be in two mutually exclusive states: either  $TF_1$  is turned on and  $Y$  is turned off, or vice versa (Figure 3) (Alon, 2007).

Another important class of network motifs is feed-forward loops (FFLs), encompassing three TFs and three interactions. Consequently, there are eight types of FFLs. Although they are all widespread in many cells types and organisms, two types are overrepresented in *E. coli* and yeasts: Coherent Type 1 FFL (C1-FFL) in which  $TF_1$  positively regulates the expression level of  $TF_2$ , and  $TF_2$  activates  $TF_3$ . In contrast, in the case of Incoherent Type 1 FFL (I1-FFL),  $TF_2$  represses  $TF_3$ , meaning that  $TF_1$  activates  $TF_3$  directly, but down-regulates the expression of  $TF_3$  indirectly (Mangan & Alon, 2003).

Finally, we introduce two slightly different signal-integrating elements called gates that combine the values of input variables in a conditional manner. We will use AND gate when two input signals have to be present to turn on the output signal. This condition represents the collaborative binding of the TFs on a certain enhancer region. The special case of the AND gate is the so-called DIMER gate by which we can model the hetero- or homodimeric interactions of nuclear receptors.



**Figure 3. Basic types of network motifs**

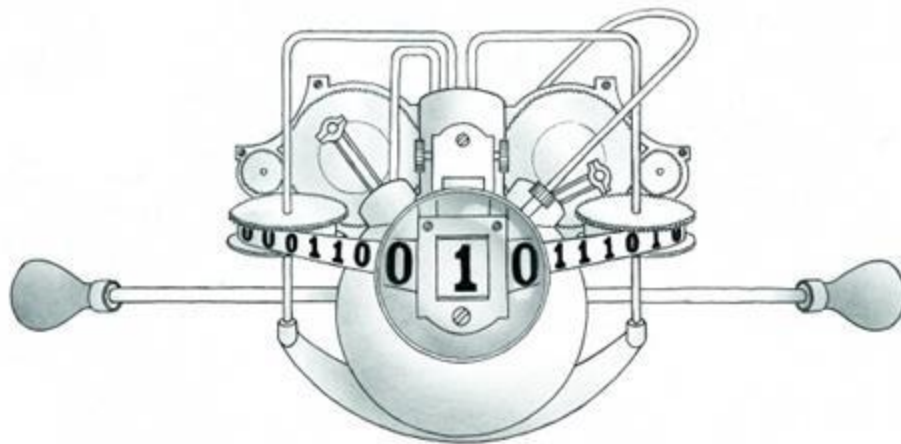
(1) Simple up/down-regulation motif (2) Mutual activation/repression motif (3) C1 FFLs (4) I1 FFLs (5) DIMER gate (6) AND gate

## Turing machine and Finite State Automata

The scientific community needs tools to conceptualize our current knowledge on complex biological processes. Theoretical scientists defined several types of mathematical ‘machines’ to precisely describe various biological phenomena (Kleene, 1956; Rozenberg, 1980). The most abstract way of describing processes is Turing machines (Turing, 1936). This type of model is considered as the most general and powerful tool to construct algorithms, because a Turing machine can carry out every effective computation as stated by the Turing-Church thesis (Hopcroft, Motwani, & Ullman, 2007). Formally a Turing machine is a septuple:  $TM = (Q, T, V, q_0, \#, d, F)$  where

- $Q$  is the nonempty (finite) set of the inner states of the machine
- $q_0 \in Q$  is the initial state
- $V$  is the nonempty set of tape alphabet symbols
- $T \subseteq V$  are the input alphabet symbols
- $\# \in (V \setminus T)$  is the *blank* symbol
- $F \subseteq Q$  are the set of finite states
- $\delta: Q \times V \rightarrow 2^{Q \times V \times \{L, R, N\}}$  is a partial function and  $2^{Q \times V \times \{L, R, N\}}$  denotes all possible subsets of  $Q \times V \times \{L, R, N\}$ , called *configuration*, where “L” means left shift, “R” means right shift. In case of “N” (no shift), the machine does not move the head.

If the current state of Turing machine (TM) is  $q \in Q$  and the tape head reads the symbol  $a \in V$ , then, if  $\delta(q, a)$  is defined, the  $\delta(q, a)$  gives the new configuration of the machine including the new inner state, the new tape symbol (not necessarily different from the previous one), and the moving direction. If  $\delta$  is not defined on the current state and the current tape symbol, then the machine halts (Pál Dömösi, János Falucska, Géza Horváth, Zoltán Mecsei, 2002).



**Figure 4. An artistic Turing machine**

<https://www.chemistryworld.com/opinion/turings-machine/4986.article>

Nonetheless, several other types of automata were elaborated for those problems that have specific restrictions for availability and/or accessibility of memory: finite automata or finite state machines (FSA) are relatively simple models and they have very pleasant computational properties (Rabin & Scott, 1959). Finite-state machines are formally defined as a tuple:  $A = (Q, q_0, T, \delta)$ , where

- $Q$  is a finite non-empty set of (inner) states
- $q_0 \in Q$  is the initial state
- $T$  is a finite non-empty set of input symbols
- $\delta: Q \times T \rightarrow 2^Q$  is the transition function.

Note that the main difference between Turing machine and FSA is the former has a ‘scratch’ memory (tape) whose content can be read and modified (e.g.: by moving the head left). While FSA can neither store nor change the input previous symbols, it can only read the current input symbol. Considering these restrictions FSA can recognize a restricted set of languages called *Regular languages* (Kleene, 1956). An example modeling a simple ATM can be seen in Figure 5.

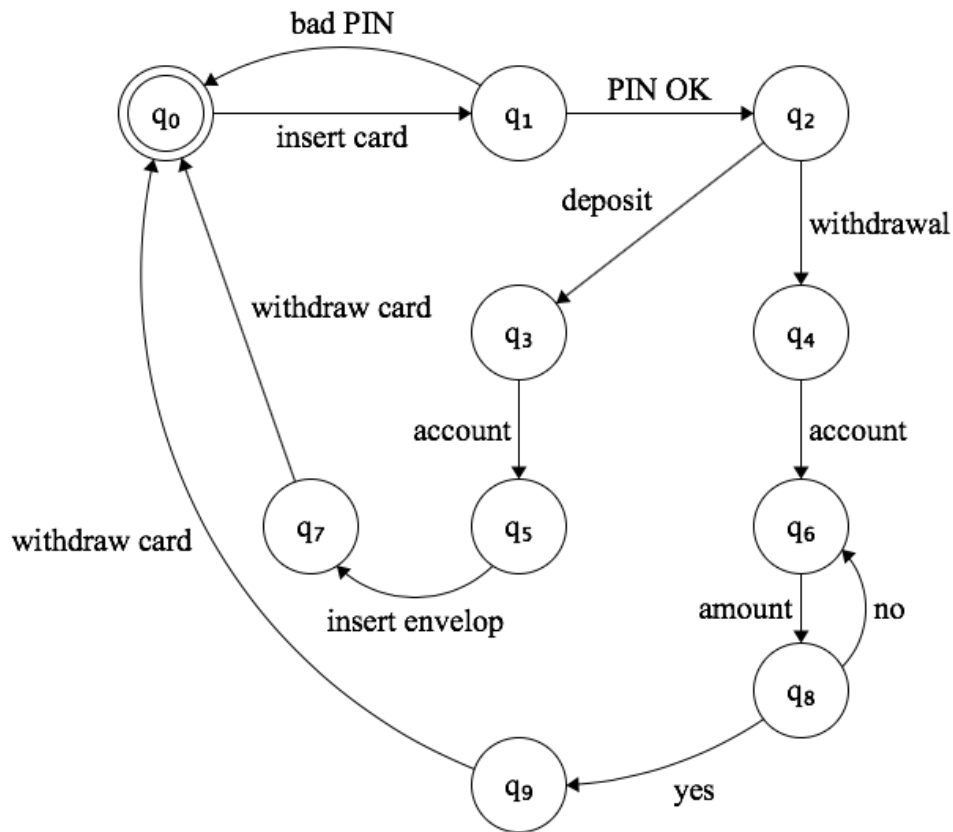


Figure 5. An example FSA modeling an ATM machine

This examples is derived from the webpage <https://people.engr.ncsu.edu/efg/210/s99/Notes/fsm/>

### 3.2. Macrophages: as model system for the study of enhancer formation

Macrophages have an essential role in immunity because they act as professional antigen-presenting cells (APCs), as phagocytes eliminating certain pathogens, and as immune modulators by secreting molecules which instruct other cells affecting their activity, proliferation, and migration (Aderem & Underhill, 1999; R. Z. Murray & Stow, 2014). Macrophage precursor cells are released into circulation as monocytes, and within a few days they seed tissues throughout the body (P. J. Murray & Wynn, 2011). Macrophages are involved in inflammatory responses, and contribute to immune tolerance and wound healing, as well as to the pathogenesis of inflammatory and degenerative diseases (Mosser & Edwards, 2008).

Macrophages are inherently heterogeneous, as they need to respond to a variety of signals in different tissue contexts. One level of heterogeneity originates from the anatomical location. Specialized tissue-resident macrophages include osteoclasts (bone), alveolar macrophages (lung), histiocytes (interstitial connective tissue), Kupffer cells (liver) and microglia (brain). The gut is populated with multiple types of macrophages, which have distinct phenotypes and functions, but work together with each other and dendritic cells to maintain tolerance to the gut flora and food (P. J. Murray & Wynn, 2011).

Another level of heterogeneity is derived from the capacity of macrophages to respond to special environmental cues. This remarkable plasticity enables monocytes and macrophages to change their physiology according to the host- and pathogen-derived signals (in situ polarization). In other words, the integration of environmental signals largely determines whether the macrophages will be involved in homeostatic activities, antimicrobial responses, host-defense against helminths, wound healing or immune regulation.

### **3.3. PU.1 is a master regulator of macrophages**

Both levels of macrophage heterogeneity are controlled mainly via transcriptional regulation. As early and more recent genome-wide studies demonstrated, PU.1 (also known as Sfp1 or Spi1) transcription factor (TF) plays the most complex role in macrophage biology. The three most commonly used terms for PU.1, namely “lineage determining” and “pioneer” transcription factor, as well as “master regulator”, accurately recapitulate its most important features (Sven Heinz et al., 2010; Kaikkonen et al., 2013; Nerlov & Graf, 1998).

First, PU.1 is a lineage determining transcription factor (LDTF) because it is indispensable for the development of macrophages. PU.1 is expressed specifically in myeloid and B-lymphoid cells of the hematopoietic system (Nerlov & Graf, 1998). In genetically modified mice, in which

the PU.1 gene has been inactivated, these lineages either completely fail to develop (Scott, Simon, Anastasi, & Singh, 1994) or the development is delayed and highly aberrant (McKercher et al., 1996).

Second, PU.1 is a pioneer TF because it can initiate events in closed chromatin (Iwafuchi-Doi & Zaret, 2014). Biochemical and genomic studies have shown that transcription factors with the highest reprogramming activity often have the special ability to bind to their target sites on DNA wrapped around the histone octamer (i. e., nucleosomal). Other reprogramming factors appear to be dependent on pioneer factors for binding nucleosomes and closed chromatin. PU.1 is most likely able to expand the linker region between nucleosomes, and promote the deposition of certain histone modifications on neighboring nucleosomes, likely contributing to its ability to enhance the binding of other transcriptional regulatory factors (K S Zaret & Carroll, 2011; Kenneth S Zaret & Mango, 2016).

Third, PU.1 is a global genomic organizer in macrophages, because the majority of the gene regulatory events (including the function of enhancers and transcriptional initiation) are actively supervised by PU.1. A series of genome-wide studies have demonstrated that PU.1 occupies the majority of active and inducible regulatory regions of the macrophage genome (Serena Ghisletti et al., 2010; Sven Heinz, Romanoski, Benner, & Glass, 2015; Lara-Astiaso et al., 2014; Lavin et al., 2014). These studies indicated that PU.1 is bound to the inducible enhancers of most regulated (i.e. not “housekeeping”) genes, including pro-inflammatory cytokines, such as IL12B, IL1b, IL6 and TNFs in resting macrophages, keeping these enhancers accessible for SDTFs, such as AP-1 dimers or NF-kB subunits.

Naturally, the three prominent features of PU.1 described above are interrelated. The facts that (1) PU.1 is highly expressed in macrophages, (2) it has the capacity to open chromatin (or



keep the enhancer accessible), and (3) it has a large number of putative binding sites in the macrophage genome make PU.1 eligible to be called as a global genomic organizer of macrophages.

### **Regulation of macrophage enhancers in the context of polarizations signals**

Cell type-specific functions are governed by lineage-determining transcription factors (LDTFs). Several LDTFs are known to act as pioneer factors that are capable of binding their specific DNA sequences and opening up condensed (closed) chromatin for other transcription factors (TFs) to bind (K S Zaret & Carroll, 2011). In macrophages, PU.1 is the master regulator of myeloid differentiation and a potential pioneer factor (Sven Heinz et al., 2015). It has been a longstanding theory that PU.1 alone is sufficient to establish the inducible enhancer repertoire in response to external stimuli (Sven Heinz et al., 2010). However, recent works in macrophages and B cells propose that PU.1 and other LDTFs need to bind and recruit chromatin remodelling factors in a collaborative manner to make these genomic regions available for the signal-dependent transcription factors (SDTFs), which will regulate their own specific gene sets (Mancino et al., 2015; Ostuni et al., 2013).

Using two mouse strains (C57BL/6J and BALB/cJ) and with a well-designed experimental layout in which naturally occurring mutations between the strains were utilized, it has been shown that PU.1 and CEBPA (CCAAT/enhancer-binding protein alpha), another key macrophage LDTF, largely affect the binding of each other; mutations in PU.1's binding motif altered the binding of CEBPA, and vice versa. Similarly, the binding of the SDTF nuclear factor- $\kappa$ B (NF- $\kappa$ B) largely relied on an intact PU.1 and CEBPA motif. On the contrary, mutations in the recognition sequence of NF- $\kappa$ B did not hinder the binding of PU.1 and CEBPA. These results demonstrate that PU.1 select regulatory regions via the binding of variably spaced

recognitions sequences in collaboration with other macrophage LDTFs (S. Heinz et al., 2013), and indicate that there might exist a binding hierarchy among the LDTFs and SDTFs. Moreover, SDTFs need open chromatin to be able bind a certain genomic region.

As we have seen above, the collaborative nature of the key macrophage TFs has been studied extensively. However, the association between macrophage-specific TFs and nucleosome free regions has remained elusive. It has been documented that the motifs of PU.1, CEBP, Interferon regulatory factor (IRF), Runt-related transcription factor (RUNX) and Activator protein 1 (AP-1) are enriched at macrophage-specific enhancers (Czimmerer et al., 2018a; Daniel et al., 2014a; S. Heinz et al., 2013). PU.1 is known to have an essential role in myeloid differentiation (Nerlov & Graf, 1998), high expression of CEBPA/B and PU.1 leads to the trans-differentiation of fibroblasts into macrophage-like cells from (Feng et al., 2008), RUNX1 is essential for the development of the hematopoietic stem cells (Samokhvalov, Samokhvalova, & Nishikawa, 2007), IRF8 is indispensable for monocyte differentiation in murine cells and is known to control pro-inflammatory genes induced by pro-classical (M1) polarization (Mancino et al., 2015; Thomas, Galligan, Newman, Fish, & Vogel, 2006). Finally, the AP-1 family member JUNB has a critical role in mediating classical and alternative macrophage polarization (Fontana et al., 2015). Consistently, recent studies indicate that specific TF combinations provide plasticity to the cells in response to single, synergistic or opposing external stimuli (K. Kang et al., 2017; Müller & Corthay, 2017; Piccolo et al., 2017). Finally, SDTFs can also form latent or de novo enhancers at regions lacking PU.1 binding, participating both in classical and alternative macrophage polarization (Kaikkonen et al., 2013; Ostuni et al., 2013).

From a computational perspective, the question arises whether chromatin openness and/or activity can be predicted using the binding of PU.1 alone or together with other key macrophage TFs. Machine learning approaches such as Random Forest or Support Vector Machines have been shown to be effective ways to extract and prioritize key features that determine TF binding and enhancer activity including histone marks and motif sequences (Chen et al., 2010; Gal et al., 2017; Liu, Jin, & Zhou, 2015; Tsai, Shiu, & Tsai, 2015) (Zhu et al., 2013). However, deterministic features of chromatin openness have not been evaluated in these works. Genome-wide maps of LDTF binding sites, profiling chromatin openness combined with machine learning approaches provide an excellent framework for exploring the connection between the binding of key TFs and chromatin openness. By using Random Forest and Support Vector Regressor methods, we build and evaluate models to predict chromatin openness from the genomic occupancy of key macrophage TFs (PU.1 CEBPA, IRF8, RUNX1 and JUNB). The Random Forest method provides not only prediction accuracies but also determines the relative contribution of the key TFs in predicting chromatin openness at the genome-wide level.

Here we demonstrate that Random Forest and Support Vector Regressor methods can predict chromatin openness from the occupancy of PU.1 and other key TFs. Using these approaches, we could not only assess the accuracy of predictions on chromatin openness, but we could reveal a novel type of enhancer class that is widespread in the macrophage genome. These genomic regions termed ‘labelled regulatory elements’ (LREs) are associated with low accessible or closed chromatin. The majority of key macrophage TFs can form LREs and these genomic regions can be transformed into functional enhancers upon polarization signals. In summary, these findings extend our understanding of the connection between TF binding and

chromatin openness, and points to the existence of LREs, which may represent a dormant state of enhancer regions and contribute to gene expression regulation in a stimulus-specific manner.

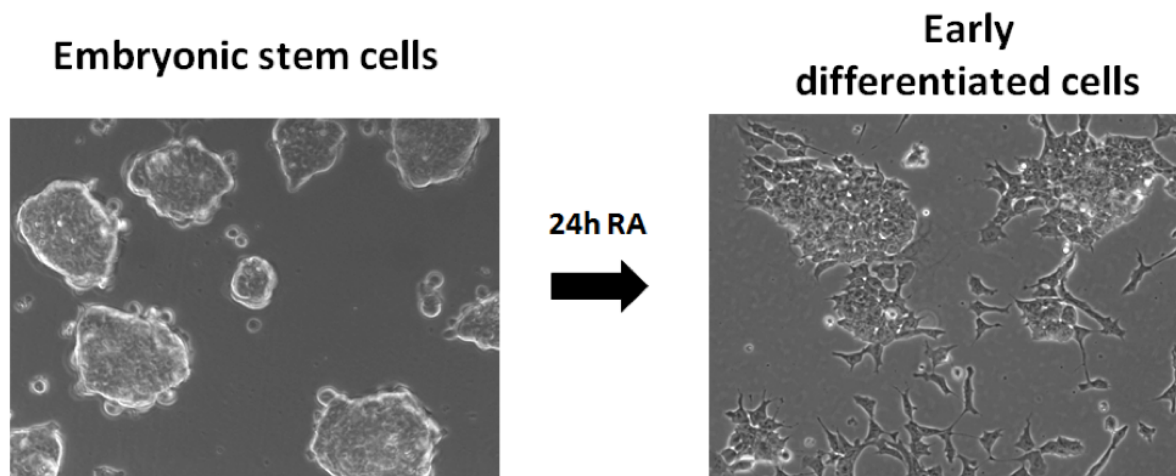
The integration of these approaches allowed us to characterize, classify and catalogue the activated regulatory elements based on Binding, Openness and Activity information, which led to the conceptualization of our understanding about the states and transitions of activated regulatory sites by building a Nondeterministic Finite Automaton termed Regulation Automaton.

### **3.4. Embryonic stem cells: a model system to study developmental enhancers**

Embryonic stem cells (ESCs) are pluripotent stem cells derived from early mammalian embryos, namely from the inner cell mass of the blastocyst. Unlike adult multipotent stem cells, which can only produce a limited number of cell types, ESCs are capable of differentiating into any of the cell types of an organism. The route of differentiation for each cell depends on external signals. When the appropriate signals are provided, ESCs can differentiate into all the three primary germ layers: ectoderm, endoderm, and mesoderm, which include each of the ~200 cell types in the adult human body (Thomson, 1998).

Retinoic acid (RA), the main active vitamin A metabolite, is a well-known regulator of embryonic development. RA is a commonly used *in vitro* model to differentiate ESCs into neurons. Moreover, it has been shown that RA deficiency leads to less effective neuronal differentiation within the granular cell layer in the dentate gyrus in adult mice, confirming the *in vivo* relevance of this model system. RA contributes to the early steps of neurogenesis, since its depletion also causes a decrease in the number of newborn cells that specifically express an immature neuronal marker (Jacobs et al., 2006; Simandi, Horvath, et al., 2016).

RA is an agonist ligand of the retinoic acid receptor (RAR), which can act as a transcription factor. RAR is known to form a non-permissive heterodimer with retinoid X receptor (RXR), that is, they can only be activated from the side of RAR. This heterodimer binds to cis-acting retinoic acid response elements (RAREs), and can induce or repress transcription of a nearby gene. RA is also known as a morphogen regulating homeobox-containing Hox genes such as Hoxa1, Hoxb1, Hoxb4 and Hoxd4 via RAREs within the regulatory elements of these genes, controlling anterior/posterior patterning in the early stages of developmental processes (Figure 6). When the ligand is not present, the RAR/RXR dimer binds to RAREs complexed with corepressor proteins. In contrast, the binding of RA to RAR results in the replacement of corepressors with coactivator proteins, that will induce the transcription of target genes (Simandi, Horvath, et al., 2016).



**Figure 6. Early stage of RA-induced neurogenesis**

### **3.5. OCT4 as one of the master regulators of pluripotency**

OCT4 (octamer-binding transcription factor 4) is a homeodomain transcription factor required for the maintenance of the undifferentiated state of mammalian ESCs. OCT4 has also

been shown to play a critical role in reprogramming somatic cells to induced pluripotent stem cells. The breakthrough discovery of the Yamanaka group has shown that by adding only four defined factors (OCT4, Sox2, c-Myc, and Klf4) to fibroblast cultures leads to the formation of pluripotent stem cells. The expression level of OCT4 has a critical role in the cell fate determination of ESCs. Low or no expression of OCT4 can trigger the differentiation towards trophoblasts, while high dosage of OCT4 leads to the establishment of primitive endodermal and mesodermal lineages. The above observations suggest that OCT4 has a fundamental role in guiding cell fate decision at the early stages of mammalian development (Radziskeuskaya et al., 2013).

The molecular mechanisms by which OCT4 fulfills its role as a gatekeeper of pluripotency, as well as a potential regulator of early cell fate decisions, remain largely uncharacterized. OCT4 binds to an octameric DNA motif (consensus: AGTCAAAT) in promoters and distal gene regulatory elements, leading to the activation or repression of certain sets of genes. Although numerous studies characterized the cisome and gene regulatory activity of OCT4 in ESCs, we still lack a mechanistic insight into how these functions are executed (Frum et al., 2013; Niwa, Burdon, Chambers, & Smith, 1998; Young, 2011). Growing number of evidence suggests that OCT4 binds to some of its target regions together with other transcription factors, controlling the gene expression program of ES cells in a collaborative fashion. However, the interaction between OCT4 and retinoic acid signaling, which is a key differentiation signal for stem cells, has not yet been characterized.

## **4. AIMS AND HYPOTHESES**

### **Aim 1. Evaluation of the contribution of key macrophage TFs to chromatin openness and enhancer formation in steady state and polarized mouse macrophages**

In the past decade, the working model of macrophage enhancer formation was that PU.1 alone is necessary and sufficient to establish and maintain the available enhancer repertoire for most SDTFs activated by external stimuli. However, recent works in macrophages and in other immune cells indicate that PU.1, and other key TFs need to bind collaboratively to make these genomic regions available for SDTFs, which will, in turn, initiate their own genomic programs. Using bioinformatics tools combined with computational methods such as machine learning approaches promise to make high accuracy predictions and find hidden correlations and features in vast data sets. In this work, we aimed to

- Map the genomic binding sites of key TFs, including PU.1, IRF8, JUNB, RUNX1 and CEBPA, in order to examine the interrelation and hierarchy among them
- Apply machine-learning approaches such as Random Forest and Support Vector Regression to build predictive models evaluating the contribution of the binding pattern of the studied TFs to chromatin openness and enhancer activation
- Classify the cistromes of the studied TFs based on chromatin openness and perform loss/gain of function experiments to investigate the deterministic role of low accessible regions in regulating gene expression in the steady state and/or in response to external stimuli

- Reveal whether there exist distinct TF modules binding low accessible genomic regions that regulate specific gene expression programs triggered by various macrophage polarizing stimuli
- Build a formal model using Automata Theory that describe the possible states of enhancer formation and transitions among them

## **Aim 2. Examination of the role of OCT4 in the early steps of RA-induced neurogenesis**

OCT4 is a homeodomain transcription factor that is essential for the maintenance of the pluripotent state in mammalian ESCs. OCT4 also play a critical role in reprogramming somatic cells to induced pluripotent stem cells, as revealed by the groundbreaking discovery of the Yamanaka-group. It has been shown that a change in the expression level of OCT4 trigger differentiation towards trophoblasts (low dosage), or lead to the establishment of primitive endodermal and mesodermal lineages (high dosage) suggesting that OCT4 may also have a contribution to fate decisions at the early stages of mammalian differentiation. However, the molecular mechanisms by which OCT4 acts as a gatekeeper of pluripotency, and as a potential regulator of early cell fate decisions remain poorly understood.

In this work, our purpose is to

- Map the OCT4 cistrome and investigate its relation to chromatin openness in both naïve and ground-state ESCs
- Characterize the interaction between the low accessible OCT4 binding sites and the retinoic acid signaling pathway, which plays a critical role in the early steps of neuronal differentiation



- Perform loss of function experiments for OCT4 to delineate its deterministic role in forming differentiation-related enhancers and regulate retinoic acid target genes
- Construct a composite network using network motifs to describe the dual role of OCT4 in maintaining pluripotency and in the context of early steps of RA-induced neurogenesis.

## **5. MATERIALS AND METHODS**

### **Differentiation of bone marrow derived macrophages**

Bone-marrow was flushed from the femur of wild-type C57BI6/J male animals. Cells were purified through a Ficoll-Paque gradient (Amersham Biosciences, Arlington Heights, IL) and cultured in DMEM containing 20% endotoxin-reduced fetal bovine serum and 30% L929 conditioned medium for 5 days. Isolated bone marrow-derived cells were differentiated for 6 days in the presence of L929 supernatant. At the 6<sup>th</sup> day of differentiation, cells were exposed to IL-4 (20ng/ml) and LPS (100 ng/ml) for the indicated period of time.

### **Embryonic stem cell culture**

Mouse E14 embryonic stem cells were cultured on primary mouse embryonic fibroblast (PMEF) feeder cells (5% CO<sub>2</sub> at 37 °C). ESC medium was prepared by supplementing DMEM Glutamax (Gibco) with 15% FBS (Hyclone), 1,000 U of LIF, penicillin/streptomycin, non-essential amino acids, and 2-mercaptoethanol. 2i ESCs were adapted for a minimum of five passages to grow in serum-free N2B27-based medium supplemented with LIF, PD0326901 (1 mM), and CHIR99021 (3 mM).

### **Ligands and Treatment**

ESCs were treated with vehicle (DMSO) or with all-trans RA (Sigma, 1 mM stock in DMSO, 1/1,000 dilution).

### **siRNA knockdown**

OCT4 and control siRNAs were obtained from Thermo Fisher. Mouse E14 cells were plated on gelatinized plates 12 hr before transfection. siRNA transfection was carried out with Lipofectamine 3000 (Invitrogen). OCT4 stealth RNAi oligo sequences: siOCT4\_A\_Fw: 5'-

AUG CUA GUU CGC UUU CUC UUC CGG G-3', 5'-CCCGGAAGAGAAAGCGAACUAGCAU-3', siOCT4\_A\_Rev: 5'-CCC GGA AGA GAA AGC GAA CUA GCA U-3', siOCT4\_B\_Fw: 5'-ACC UUC UCC AAC UUC ACG GCA UUG G-3', siOCT4\_B\_Rev: 5'- CCAAUGCCGUGAAGUUGGAGAAGGU-3'. Transfected cells were cultured in embryonic stem (ES) medium for 24 hr prior to experiments. Cells were ligand treated for 3-24 hrs before harvesting for eRNA and mRNA experiments, meaning that cells were harvested 24+3 or 24+24 hrs following the transfections. Similarly, ChIP experiments were performed 24 hrs following the transfection or the latest 48 hrs if ligand treatment was applied.

### **Microarray analysis**

Control, Nanog RNAi and OCT4 RNAi CEL files of GSE4189 microarray series were downloaded from NCBI/GEO database and imported into GeneSpring (version 13.0). Gene Level Experiment was carried out using the following parameters: Threshold raw signals to 1.0, Normalization algorithm: quantile, Baseline to median of control samples (GSM94856-GSM94860), Average over replicates in conditions. Entity lists containing pre-selected RA target genes or components of the retinoic acid signaling pathway were used for heatmap analysis.

### **RT-qPCR**

RNA was isolated with Trizol reagent (Ambion). RNA was reverse transcribed with High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) according to manufacturer's protocol. Transcript quantification was performed by qPCR reactions using SYBR green master mix (BioRad). Transcript levels were normalized to *Ppia*.

## **ChIP-seq**

ChIP was performed according to (Barish et al., 2010; Daniel et al., 2014b; Siersbaek et al., 2012), with minor modifications. The following antibodies were used: OCT4 (sc-8628), PU.1 (sc-352), IRF8 (sc-6058x), JUNB (sc-46x), CEBPA (sc-61x), RUNX1 (sc-8563x), STAT6 (sc-981x), p65 (sc-372), H4ac (millipore 06-866), H3K4me1 (ab8895), H3K27ac (ab4729), IgG (Millipore, 12-370), OCT4 (Santa Cruz, sc-8628), RXR (Santa Cruz, sc-774), RAR (Santa Cruz, sc-773), P300 (Santa Cruz, sc-585), H3K27me3 (Millipore, 07-449) and RNAPII-pS2 (ab5095). Libraries preparation was performed by Ovation Ultralow Library Systems (Nugen) from two biological replicates according to the manufacturer's instructions.

## **Western blot analysis**

20 µg protein whole cell or nuclear extracts were separated by electrophoresis in 10 or 12.5% polyacrylamide gels and then transferred to Immobilon-P Transfer Membrane (Millipore Crp., Billerica, Massachusetts). Membranes were probed with anti-Oct3/4 (Santa Cruz; sc-5278), anti-RAR (Santa Cruz, sc-773), or anti-GAPDH (Santa Cruz, sc-32233) antibodies according to the manufacturer's recommendations.

## **GRO-seq**

Global Run-On sequencing and library preparation was performed as described earlier (Core, Waterfall, & Lis, 2008). Libraries were made from two biological replicates of BMDMs. Libraries were sequenced with Illumina HiScanSQ sequencer.

## **ChIP-seq, GRO-seq and ATAC-seq analyses**

Primary analysis of the ChIP-seq, GRO-seq, and ATAC-seq raw reads was carried out using a ChIP-seq analyze command line pipeline (Barta, 2011). Briefly, Burrows-Wheeler Alignment Tool (H. Li & Durbin, 2009) was used to align the reads to mm10 genome assembly (GRCm38)

with default parameters. MACS2 2.0.10 (Zhang et al., 2008) was used for predicting TF peaks (q-value  $\leq 0.01$ ) and findPeaks.pl (with '-size 1000' and '-minDist 2500' options) for histone regions with the option '-style histone'. Artifacts were removed using the ENCODE blacklist (Consortium, 2012). 'Intergenic' and 'Intron' regions were considered as distal elements from HOMER (v4.2) annotation. Reads per kilobase per million mapped reads (RPKM) values of the predicted peaks was calculated using BedTools coverageBed and bash scripts. DiffBind v2.8.0 (Ross-Innes et al., 2012) was used to infer differential binding sites from duplicates of STAT6 and p65 from CTR, 1h IL-4 and 1h LPS treated cells (p-value  $< 0.05$ ), respectively, and from RNAPII-pS2 ChIP-seq time course experiments (p-value  $< 0.05$  & FC $>2$ ) measured on distal regions (normalized DiffBind occupancy  $> 30$ ) and on gene bodies (normalized DiffBind occupancy  $> 50$ ), using untreated samples as controls. K-means clustering of RNAPII-pS2 regions both on distal elements and gene bodies (mm10 RefSeq) was performed using kmeans() function from the R package stats. GO analyses were performed using the clusterProfiler R package. Intersections, subtractions, and merging of the predicted peaks were done with BedTools (v2.23.0). Regions were considered overlapping if there was at least one common nucleotide. Consensus sets were defined by merging overlapping regions (in at least 2 samples). Proportional Venn diagrams were generated with VennMaster. Genome coverage files (BedGraphs) for visualization purposes were generated by makeUCSCfile.pl, and then converted into tdf files using igvtools (IGV2.3, Broad Institute) with the 'toTDF' option. Genomic distribution was analyzed using HOMER categories provided by annotatePeaks.pl (UTR regions were merged). De novo motif discovery was performed in the 150 bp vicinity of the peak summits using findMotifsGenome.pl with options '-length -len '12,14,16,18,20,22'' and '-size 200' on the repeat-masked mouse genome (mm10r) from HOMER. Integrative

Genomics Viewer (IGV2.3, Broad Institute) was used for data browsing (Thorvaldsdottir, Robinson, & Mesirov, 2012) and creating representative snapshots. Normalized tag counts for Meta histograms and read distribution heatmaps (RD plots) were generated by `annotatePeaks.pl` with ‘-ghist’ and ‘-hist 25’ options from HOMER on one representative example of duplicates and then visualized by R (`ggplot2`) or Java TreeView. Motif matrices were remapped using `annotatePeaks.pl` with the ‘-mscore’ option. Summits used for centering RD plots and motif remapping were identical to the summit of the peak with the highest MACS score from those used for deriving the consensus region.

## **Machine learning**

Machine learning analyses were performed in R using the packages `randomForest`, `e1071` and custom scripts. For training sets, 1,000 sites were randomly chosen from both labelled and ‘HighAcc’ categories (based on the probability values given by the machine learning methods) for Random Forest and Support Vector Machine models (repeat-masked mouse genome `mm10r`). To avoid the well-known issue of ‘overfitting’ in data mining, all models were built using a k-fold ( $k=10$ ) cross validation. In total, 31 Random Forest models were generated from all possible TF combination (one TF only,  $n=5$ ; two TFs,  $n=10$ ; three TFs,  $n=10$ ; four TFs,  $n=5$ ; all five TFs - ‘full model’,  $n=1$ ). For validation sets, another 1,000 sites were randomly chosen from both labelled and ‘HighAcc’ categories that were not used for learning processes. Contribution scores (`MeanDecreaseGini`) were calculated using `randomForest` function with the ‘importance = T’ option. Sensitivity, Specificity and ROC values were calculated with the `caret` and `ROCR` packages. Boxplot of the 31 models and the ROC plot were generated using `ggplot2` R package. For SVM model, Pearson correlation coefficient was calculated on an independent validation set using stats packages in R.

## **Data Availability**

All data from this study have been submitted to Sequence Read Archive (SRA - NCBI; <https://www.ncbi.nlm.nih.gov/bioproject>) and to NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) under accession number PRJNA302640 (ESC data sets), PRJNA194083, PRJNA318630 and GSE106706 (macrophage data sets), respectively. In-house scripts used to generate data and figures can be downloaded from the GitHub repository [https://github.com/ahorvath/horvath\\_et\\_al\\_2019\\_labelled\\_enhancers](https://github.com/ahorvath/horvath_et_al_2019_labelled_enhancers).

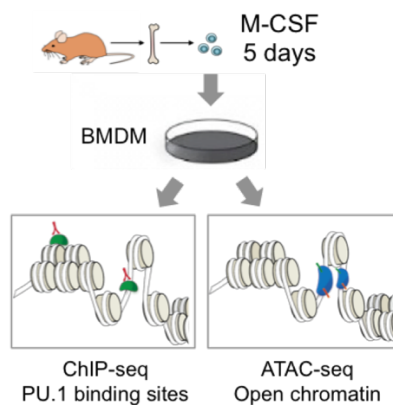
## **Author contributions to the wet-lab experiments**

The experiments were carried out by Lajos Széles, Bence Dániel (in macrophages) and Zoltán Simándi (in ESCs).

## 6. RESULTS

### 6.1. Random Forest classification hints the existence of low accessible TF binding sites in macrophages

In order to examine the interrelationship between the key macrophage TFs and chromatin openness, first we thoroughly characterized the open chromatin regions (OCRs) in macrophages. Although the cisomes of PU.1 and other key TFs have been extensively characterized in mouse macrophages (Czimmerer et al., 2018b; S Ghisletti et al., 2010; Sven Heinz et al., 2010; Link et al., 2018; Ostuni et al., 2013), the relationship of PU.1 and other key TFs with chromatin openness have remained poorly understood. One intriguing question is whether the binding of the key TFs overlap with OCRs or they are associated with low accessible regions as well. First, we profiled OCRs by performing Assay for Transposase Accessible Chromatin coupled to sequencing (ATAC-seq) in bone marrow-derived macrophages (BMDMs) (Figure 7).









**Figure 7. BMDM model system**

A scheme showing bone marrow-derived macrophage (BMDM) differentiation.



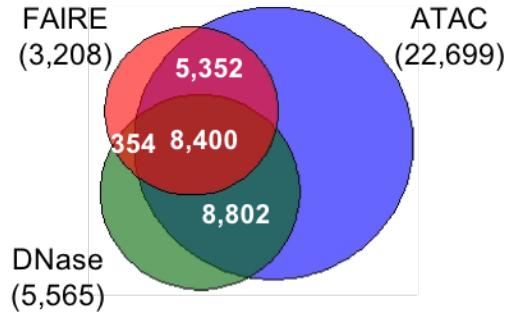
Next, we carried out a de novo motif analysis on the distal OCRs (over 32,000 regions) which identified the motif of PU.1 (ETS) (37.5%), TRE (28.4%) known to bound by a subset of the AP-1 family (Garces de Los Fayos Alonso et al., 2018), CTCF (6.4%), CEBP (31.9%), RUNX (12.4%) and a composite element called EICE (PU.1-IRF) (4.5%) (Figure 8).

Motif logo	P-value	% of Target	% of Bg	Db. Annotation
	1e-4057	37.48	8.03	PU.1
	1e-3257	28.35	5.42	TRE
	1e-1199	6.39	0.56	CTCF
	1e-588	31.81	18.75	CEBP
	1e-402	12.43	5.61	RUNX
	1e-299	4.53	1.31	EICE

**Figure 8. De novo motif analysis of distal OCRs in BMDMs**

The highly enriched de novo motifs under the distal OCRs. Motif sequence logos, P-value of enrichment, percentage of target sequences with motif (Targets (%)), and percentage of background sequences with motif (Bg. (%)) are shown.

To compare the sensitivity of the ATAC-seq technique to other available methods we analyzed two literature BMDM data sets; FAIRE-seq (Ostuni et al., 2013) and DNase-seq (Leddin et al., 2011). Based on our comparison, ATAC-seq could detect more than twice as many regions than FAIRE-seq or DNase-seq (Figure 9).

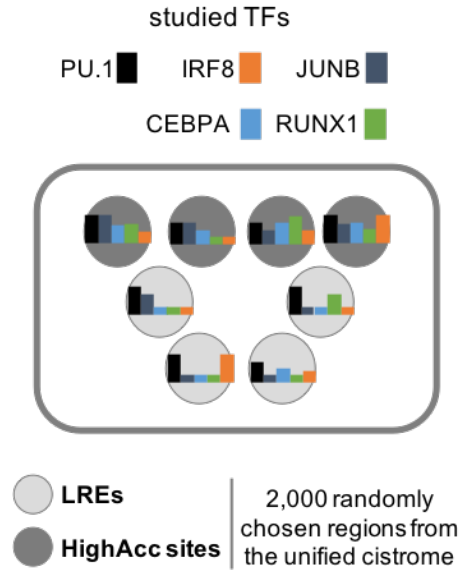


Experiment	Detected(Nr)	Detected(%)
ATAC	45,253/54,380	83.2%
DNase	23,121/54,380	42.5%
FAIRE	17,314/54,380	31.8%

**Figure 9. Comparison of methods profiling open chromatin regions**

Venn diagram showing overlap among the OCRs predicted by the three techniques (ATAC-seq (this study), DNase-seq (Leddin et al., 2011); and FAIRE-seq (Ostuni et al., 2013).

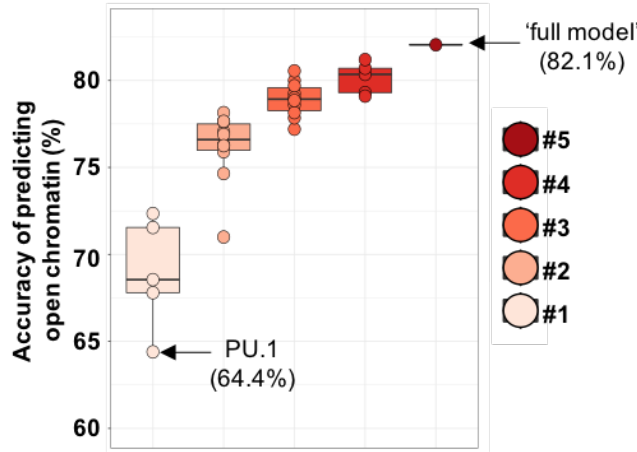
Next, we asked whether PU.1 binding alone is sufficient to determine chromatin openness or other key TFs need to be taken into account as well. To answer this question, we used a machine learning approach and examined the predictive accuracy and relative contribution of the binding pattern of the key macrophage TFs in establishing OCRs. First, we performed ChIP-seq experiments for the TFs PU.1, IRF8, JUNB, CEBPA and RUNX1, as their motifs were highly enriched based on our de novo motif analyses and these TFs were documented as key TFs in macrophages. After predicting the genomic binding sites of each TFs, we unified these genomic regions. We randomly chose 1,000 regions that did not overlap with OCRs, termed ‘labelled’ regulatory elements (LREs), and 1,000 regions that overlapped with OCRs which we termed ‘HighAcc’ sites (Figure 10).



**Figure 10. Flow chart of the machine learning approach.**

Labeled regulatory elements (LREs) are defined as one or more transcription factors (TFs) bound to low accessible chromatin. Highly accessible sites (HighAcc sites) are bound by one or more TFs and exhibit high ATAC-seq signals.

We systematically assessed all combinations of TFs by using Random Forest method and measured their prediction accuracy. Interestingly, the model where only PU.1 binding pattern was used showed weak prediction accuracy of openness (64.4%), only marginally higher than expected by chance. Interestingly, all the other models which accommodated only one of the other key TFs resulted in similarly weak predictive powers, however, slightly better than PU.1. Comparison of models revealed that the higher the number of included TFs, the higher the accuracy became (Figure 11).



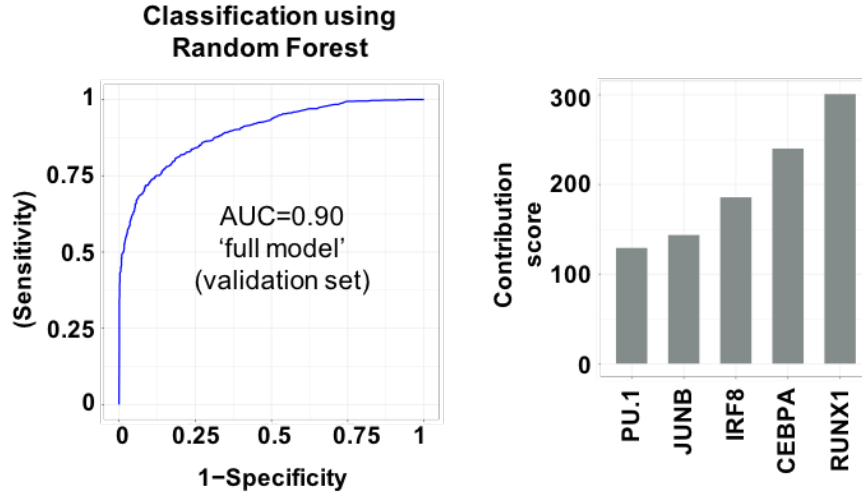
**Figure 11. Prediction accuracies of the tested models**

Box plots showing the accuracies of models in predicting open chromatin regions built with one TF (n=5), two (n=10), three (n=10), four (n=5) or five (n=1, 'full model') TF(s).

The model in which all the TFs were used as input ('full model'), resulted in a high accuracy prediction on an independent validation set (Accuracy = 0.82, Sensitivity = 0.88, Specificity = 0.77, AUC = 0.90, Table 1 and Figure 12) which were inferred similarly as the training set but not used for during the learning process. The applied Random Forest method provides not only prediction accuracies but also estimates the contribution of the input variable (TF binding) in predicting chromatin openness. This analysis revealed that all the other TFs had higher contribution than PU.1 in defining chromatin openness (Figure 12).

	Size	Sensitivity	Specificity	Accuracy
<b>Training Set</b>	2000	0.82	0.77	0.80
<b>Validation Set</b>	2000	0.88	0.77	0.82

**Table 1. The results of Random Forest approach on the Training and Validation sets**



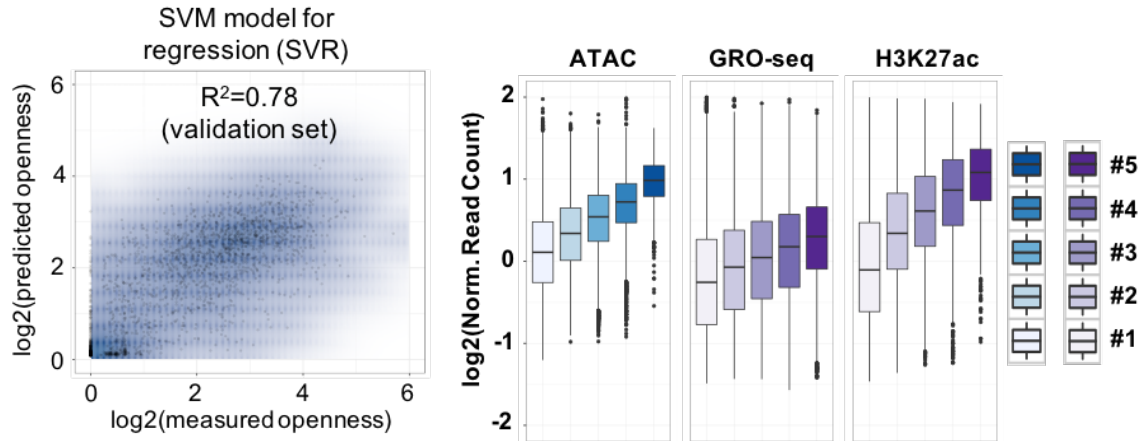
**Figure 12. The results of Random Forest approach on the ‘full model’**

(1) ROC (Receiver Operating Characteristic) curve of ‘full model’ (False Positive Rate and True Positive Rate are plotted on the x-axis and y-axis, respectively). AUC (area under the ROC curve) shown was calculated on the validation set. (2) Bar plot showing the relative importance of PU.1 and the other four TFs assessed in the ‘full model’.

In addition, using the Support Vector Regressor method, chromatin openness can be predicted not only qualitatively (classification task) but also quantitatively with a high correlation between the measured vs. predicted ATAC-seq signal (Figure 13). Finally, In consistence with our previous finding, we addressed the question whether and how chromatin openness and enhancer activation rely on the number of co-bound TFs. After partitioning the unified cistromes into disjunct subsets based on the number of co-bound TFs, SVR revealed that the more TFs bound to a particular genomic region, the higher chromatin openness (ATAC-seq), enhancer RNA level (GRO-seq) and chromatin activity (H3K27ac) can be measured (Figure 13).

Collectively, our results demonstrate that 1) using machine learning methods, chromatin openness can be predicted with high accuracy, using the binding pattern of all five TFs,

confirming that the most important macrophage TFs have been identified and included in the model; 2) in the ‘full model’ the contribution of PU.1 binding is the lowest among the studied TFs, which indicates that PU.1 binds low accessible genomics regions.



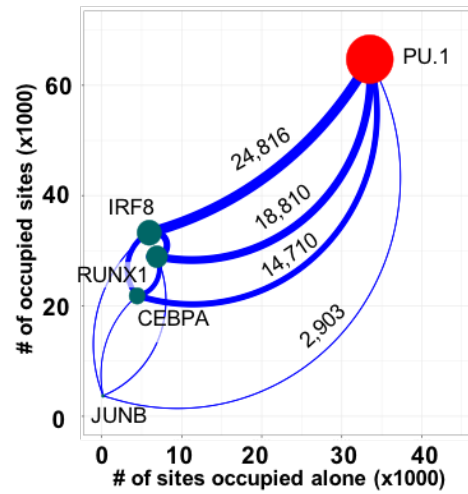
**Figure 13. The results of Support Vector Regressor analysis**

- (1) Scatter plot showing the Pearson correlation coefficient of predicted and measured openness calculated on the validation set. (2) Box plot representation of ATAC-seq, GRO-seq, and H3K27ac signals on TF co-bound regions. #1-5 labels indicate the number of co-bound TFs present at the given genomic loci.

## 6.2. PU.1-labelled regulatory elements are widespread in the macrophage genome

We examined the relationship between the cistromes of the studied TFs, and their ability to bind low accessible sites alone. Our analyses showed that PU.1 possessed the biggest cistrome (64,728 sites) and had the highest portion of the cistrome that did not overlap with any other TFs’ cistrome (51.8%). On the contrary, lower fractions of IRF8, RUNX1 and CEBPA (17.9%, 23.8% and 20.3%, respectively) had solo binding. Interestingly, only 4.5% of the total JUNB cistrome could be characterized by solo binding, suggesting that JUNB strongly relies on the binding of other TFs’. IRF8 had the most highly overlapping cistrome with PU.1 (24,816 sites), and RUNX1 and CEBPA also had an extensively shared cistrome with PU.1 (18,810 and 14,710

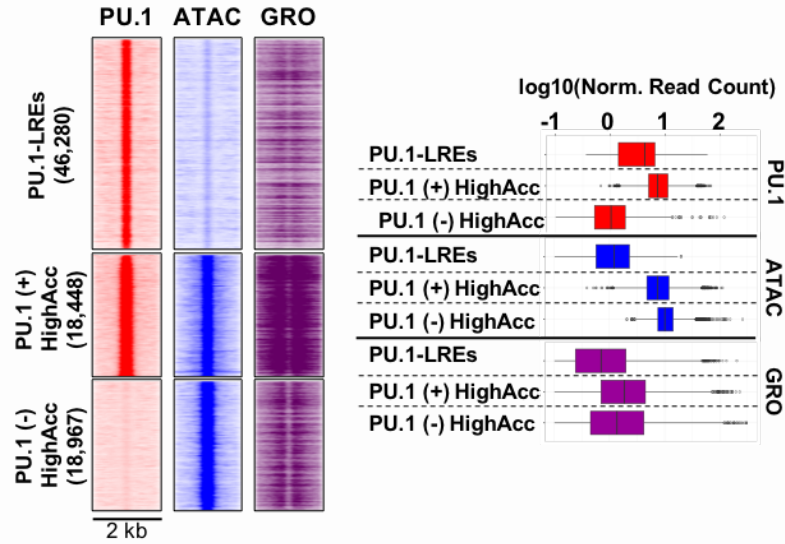
sites, respectively). Notably, PU.1 bound a very high fraction (78%) of the JUNB cistrome (2,903) (Figure 14).



**Figure 14. Binding hierarchy of PU.1 and other key TFs in macrophages**

Scatter plot showing the total number of binding regions for each transcription factor (TF) cistromes (y-axis), and the number of sites where the particular TF does not overlap with any other TF (x-axis). The overlap of the PU.1 cistrome with all the presented TF cistromes are represented by the thickness of the connecting lines on which the actual numbers of overlapping genomic regions are shown. Dot sizes are proportional to the sizes of the TFs' cistromes.

Based on these analyses we classified the PU.1-bound sites into three classes: (1) binding sites that were not associated with predicted OCRs, termed 'PU.1-labelled' regulatory elements (PU.1-LREs, 46,280 sites); (2) regions where PU.1 binding overlapped with open chromatin regions, termed 'PU.1 pos. HighAcc' sites (18,448 sites); and (3) PU.1-negative highly accessible regions, termed 'PU.1 neg. HighAcc' sites (18,967 sites). Strikingly, only one-third of the PU.1-bound genomic regions overlapped with OCRs (Figure 15).



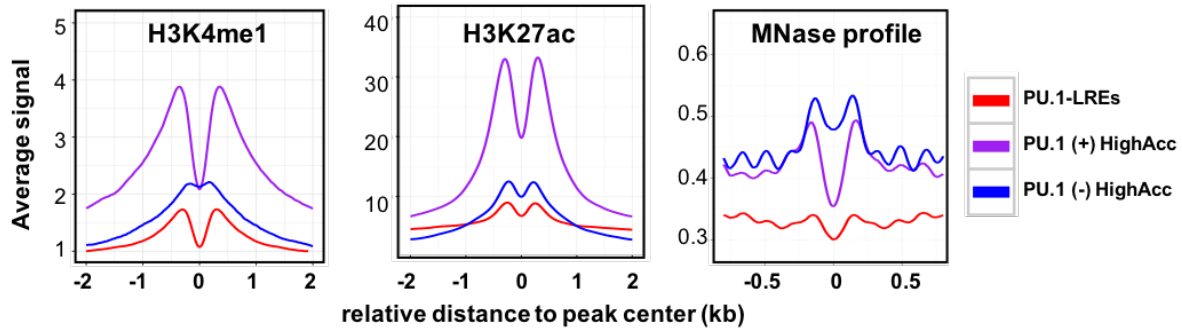
**Figure 15. Interrelationship of chromatin openness and PU.1 binding in BMDMs**

Read distribution (RD) plots showing PU.1 occupancies, chromatin openness (ATAC-seq) and nascent RNA transcription (GRO-seq) in the three groups clustered based on PU.1 occupancy and ATAC-seq signal. The following nomenclature was used for the three clusters: PU.1-labelled regulatory elements (PU.1-LREs), PU.1 positive highly accessible regions (PU.1 (+) HighAcc) and PU.1 negative highly accessible regions (PU.1 (-) HighAcc). RD plots show the signals in 2-kb windows around the summit of PU.1 or ATAC peaks. Box plot representations of PU.1 occupancy and chromatin accessibility in the three clusters from ChIP-seq and ATAC-seq experiments are also shown.

Next, we compared these three categories regarding GRO-seq signal and the level of histone marks H3K4me1 (general enhancer) and H3K27ac (active enhancer mark) (Figure 16). Both histone marks showed high enrichment on ‘PU.1 pos. HighAcc’ sites, while ‘PU.1 neg. HighAcc’ sites were much less enriched, and PU.1-LREs exhibited the lowest signals for these features. These results indicate that these classes are indeed functionally distinct, and PU.1-LREs may represent a context where PU.1 requires other TF(s) that are not present in the unstimulated state to maintain or establish open chromatin regions. Next, we re-analyzed a publicly available MNase-seq data set (Barozzi et al., 2014), which demonstrated that the



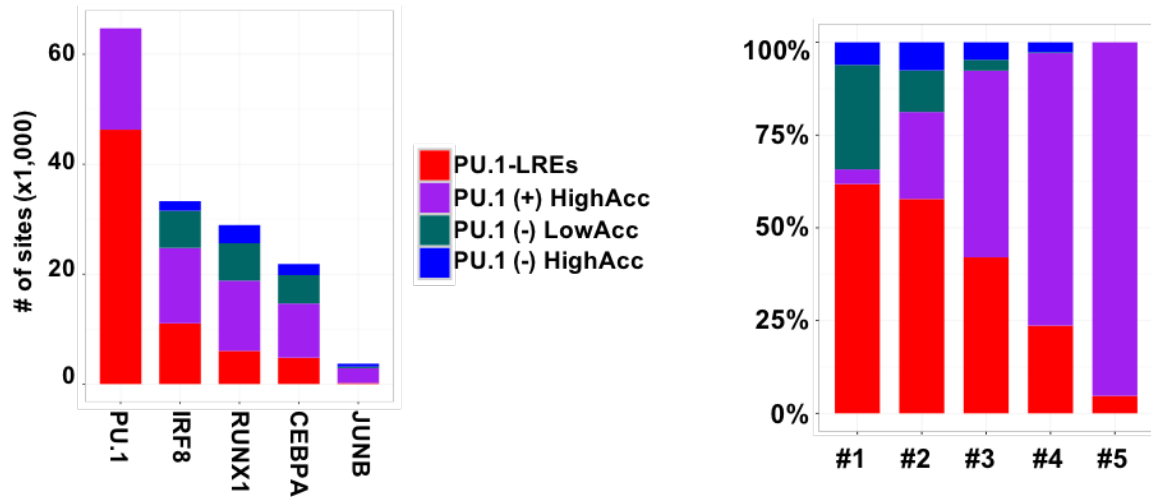
surrounding nucleosomes of ‘PU.1 neg. HighAcc’ and ‘PU.1 pos. HighAcc’ sites showed higher MNase signal compared to the ‘PU.1-labelled’ sites (Figure 16), confirming low accessible (closed) chromatin environment and also suggesting less stable PU.1 binding at labelled sites.



**Figure 16. Various epigenomics marks across the three groups**

Meta histograms showing H3K4me1, H3K27ac signals and MNase-seq (Barozzi et al., 2014) signals for the three categories. Signals were measured around the summits of PU.1 (PU.1-LREs and PU.1 (+) HighAcc) or ATAC-seq signals (PU.1 (-) HighAcc).

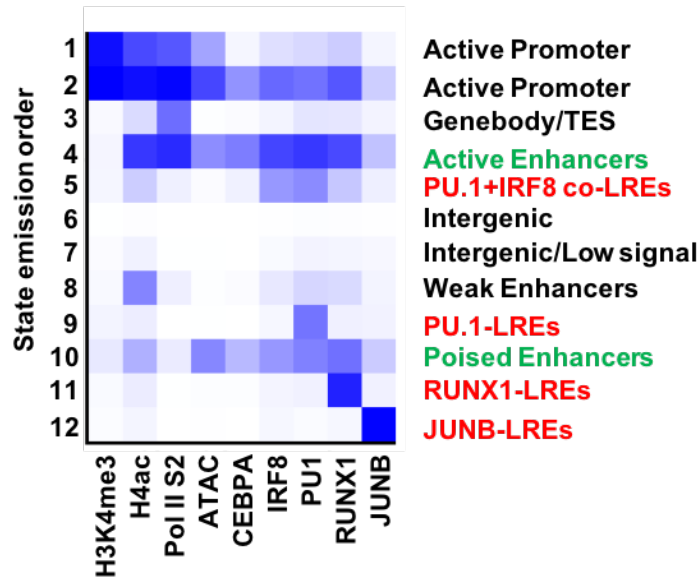
Next, we compared the TFs’ cistromes. This analysis revealed that more than half of the TFs’ cistromes were associated with the cistrome of PU.1, and the majority of these sites overlapped with ‘PU.1 pos. HighAcc’ sites (Figure 17). Notably, a remarkable portion of sites for IRF8, RUNX1 and CEBPA were neither associated with PU.1 nor with open chromatin regions (‘PU.1 neg. LowAcc’), suggesting the presence of LREs characterized by the binding of these other TFs. This classification also showed a good correlation with our previously characterized co-binding events. The fraction of the ‘PU.1 pos. HighAcc’ sites progressively increased, while the fraction of PU.1-LREs and PU.1 neg. LowAcc (labelled by other TFs) genomic regions gradually decreased with the number of co-bound TFs.



**Figure 17. Comparison of the four categories with the number of co-bound TFs**

(1) Stacked bar plots with the number of binding regions from each cistrome using PU.1, RUNX1, CEBPA, IRF8 and JUNB ChIP-seq experiments in the context of the identified categories (PU.1-labelled regulatory elements (PU.1-LREs), ‘PU.1 (+) HighAcc’, ‘PU.1 (-) LowAcc’ and ‘PU.1 (-) HighAcc’). (2) Stacked bar plot showing the percentage of binding sites in each cistromic category presented on panel (1) as a function of the number of co-binding factors present at the given genomic loci (#1-5 labeling indicates the number of co-bound TFs for each cistrome).

Finally, we used ChromHMM, a Hidden Markov Model-based method, as an alternative to investigate the pervasiveness of LREs (Ernst & Kellis, 2012). Using the studied TFs’ binding, ATAC-seq and RNAPII-pS2 ChIP-seq data (elongation-specific polymerase II), as well as H3K4me3 (active promoter mark) and H4ac (active histone mark) signals we could confirm our finding, as various types of LREs were predicted such as PU.1-LREs, PU.1+IRF8 co-LREs and RUNX1-LREs (Figure 18).



**Figure 18. ChromHMM analysis of the key macrophage TFs and various histone marks**

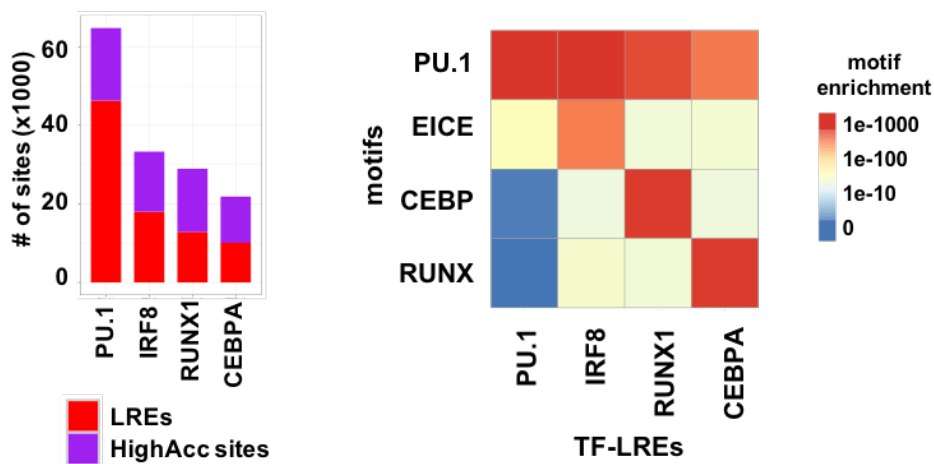
Heatmaps showing the result of ChromHMM (Ernst & Kellis, 2012) analysis. Genomic regions were clustered into 12 clusters from ATAC-seq, H3K4me3, H4ac, RNAPII-pS2, PU.1, IRF8, JUNB, RUNX1 and CEBPA ChIP-seq data as input variables (left panel) and the annotation of the inferred clusters according to RefSeq annotation (right panel).

Taken together, these results show that two-thirds of PU.1-bound regions are associated with low accessible chromatin and the lack of transcriptional activity; we termed these sites LREs. Moreover, both the characterization of ‘PU.1 neg. LowAcc’ sites and the ChromHMM-based analysis suggest that IRF8, RUNX1 and CEBPA also have labelled fractions and there might be unique co-LRE TF modules.

### **6.3. Key transcriptional regulators of macrophage form labelled regulatory elements**

Next, we asked whether only PU.1 has the ability to form LREs or it is a general phenomenon among the macrophage TFs. Therefore we investigated how the promoter-distal fractions of the IRF8, CEBPA and RUNX1 cistromes overlap with OCRs. We excluded JUNB

from further analyses as it is not eligible to be called LDTF due to its relatively small cistrome in the unstimulated state, which has been also pointed out by other groups (Ostuni et al., 2013), and its low fraction of LREs (Figure 17). Although IRF8, CEBPA and RUNX1 had cistrome sizes comparable to PU.1, nearly ~50% of their cistromes could be associated with low accessible chromatin. Therefore, we categorized these sites as LREs (Figure 19). Next, we analyzed the p-values of the motifs enriched at each type of LRE (Figure 19).

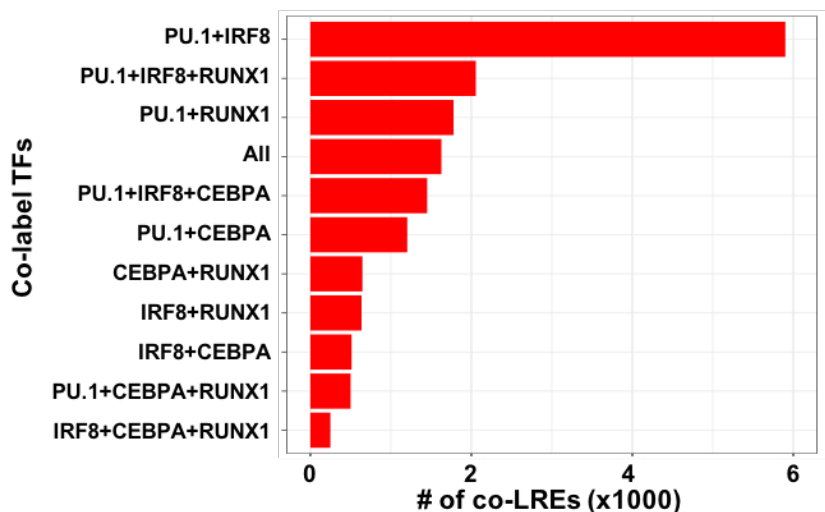


**Figure 19. Characterization of LREs on the cistromes of key TFs**

(1) Stacked bar plot showing the distribution of the transcription factor (TF) cistromes (PU.1, IRF8, RUNX1, CEBPA) on labelled regulatory elements (LREs) and highly accessible chromatin regions (HighAcc sites). (2) Heatmap showing the p-values of motif (PU.1, EICE, CEBP and RUNX) enrichments of the redundant set of the different TF-LREs (PU.1, IRF8, RUNX1 and CEBPA)

As anticipated, the specific motif of the TFs had high p-value in the corresponding TF-LRE groups. Interestingly, the PU.1 motif was highly enriched in all the TF-LRE groups, confirming the genome organizing role of PU.1 in macrophages. This result can provide an explanation why the labelled IRF8, CEBPA and RUNX1-labelled cistromes overlap with PU.1-labelled sites. In contrast, these three TFs were only moderately enriched for the motifs of the other TFs.

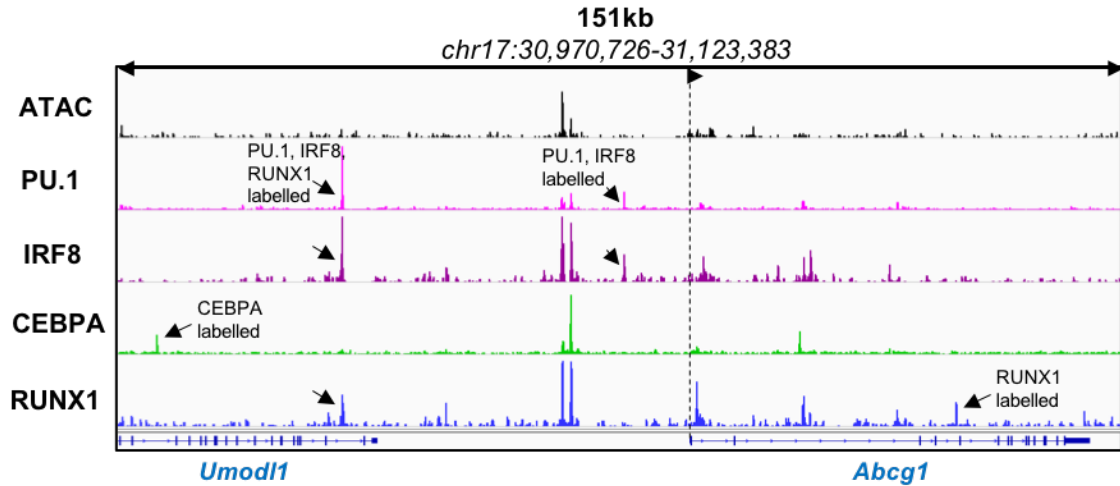
On the contrary, PU.1-LREs were less enriched for the motifs of the other three TFs, which might be due to the higher number of the PU.1 binding sites at genome-wide level. Interestingly, the EICE motif were highly enriched at PU.1-LREs, indicating that IRF8 strongly collaborates with PU.1 on PU.1-LREs (Figure 19). Next, we investigated the co-localization of TFs at different types of LREs (Figure 20).



**Figure 20. The key regulators form Co-LREs are widespread in macrophages**

Bar plot showing the number of different subsets of co-LREs bound by the TFs studied.

Our results show that PU.1+IRF8 co-LREs were the most frequent (5,901 sites) among the groups. In addition, the top six most prevalent combinations contained either PU.1 or IRF8, or both. Consistently with this, the highly accessible sites contained a high number of co-bound TFs, while we detected a relatively low number of co-LREs that overlapped with three or all four TFs.



**Figure 21. Examples for different co-LREs**

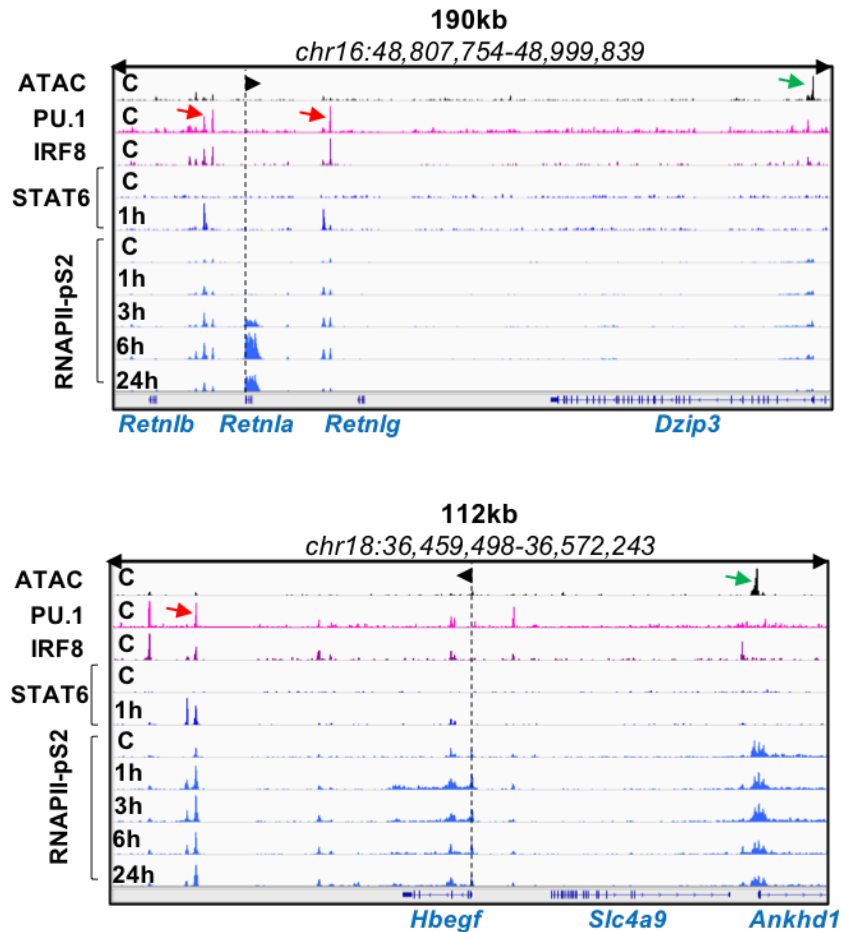
IGV snapshot of *Abcg1* locus with representative examples for different type of TF co-LREs. Black arrows highlight the different LREs and dashed line with an arrow head represent the transcription start site and the direction of the gene.

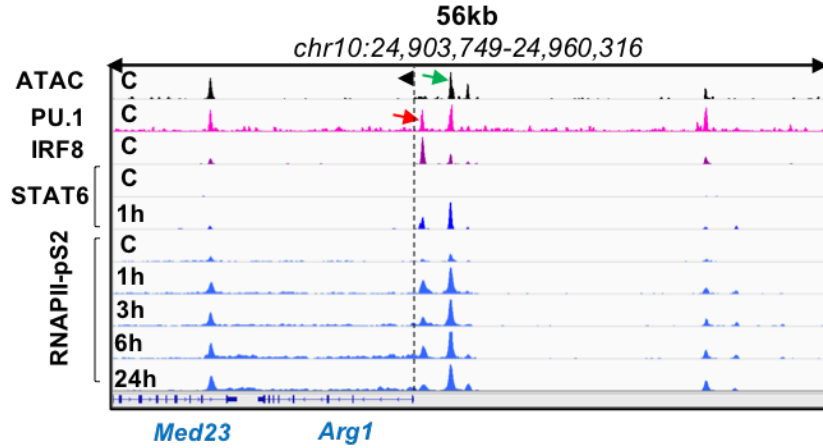
In conclusion, LREs are widespread in cistromes of the studied TFs, and PU.1-labelled sites are the most prevalent LREs because PU.1 has the largest cistrome and its ability to bind alone. Thus, the labelled portion of IRF8, CEBPA and RUNX1 often overlap with PU.1 and they form co-LREs (Figure 21). Finally, among the possible combinations of co-LREs, PU.1+IRF8 co-LREs are the most abundant, having around three times as many binding sites than the second most abundant combination, which also contain the TFs PU.1 and IRF8.

#### **6.4. The role of PU.1 and IRF8 co-LREs in cellular response to IL-4**

Based on our analyses, there is a high number of PU.1+IRF8 co-LREs in the unstimulated state, therefore we investigated the possible function of these genomic regions in IL-4-mediated macrophage polarization (alternative macrophage polarization). Signal Transducer and Activator of Transcription 6 (STAT6) almost exclusively governs the early steps of the polarization process (Gordon & Martinez, 2010), binding to thousands of regions in the genome

within minutes of IL-4 treatment, triggering a robust gene expressional program. Therefore we treated the macrophages for 1h with IL-4. We were wondering whether there might be PU.1+IRF8 co-LREs playing guiding roles in genomic programs at marker genes of the polarization process. After mapping the activated distal regulatory regions associated with LREs using time-course RNAPII-pS2 ChIP-seq experiments in IL-4-treated BMDMs (1h, 3h, 6h and 24h), we observed that PU.1+IRF8 co-LREs can often be associated to important IL-4 regulated genes such as *Retnla*, *Hbegf*, and *Arg1* (Figure 22).





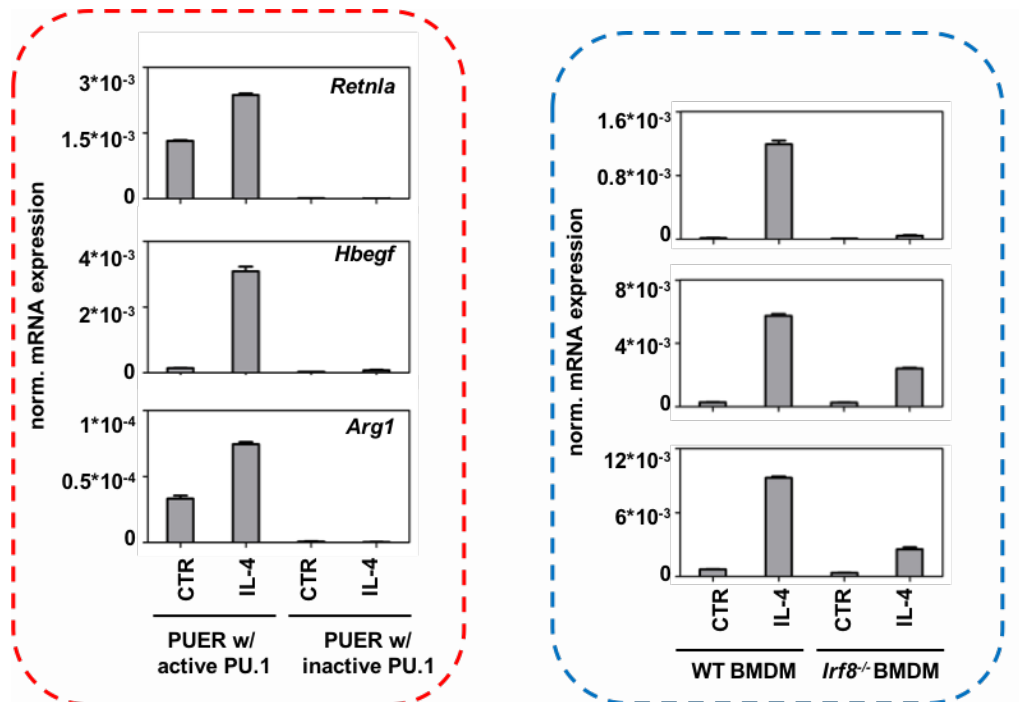
**Figure 22. IGV snapshots of alternative macrophage polarization marker genes**

IGV snapshots of three alternative macrophage polarization specific marker gene loci (*Retnla*, *Hbegf* and *Arg1*) with PU.1+IRF8 co-labelled regulatory elements (PU.1+IRF8 co-LREs) that bind STAT6 upon IL-4 treatment and recruit the elongation-specific form of RNA polymerase II (RNAPII-pS2). Green arrows represent highly accessible chromatin regions, while red arrows depict PU.1+IRF8 co-LREs. Dashed lines with arrowheads represent the transcription start site and the direction of the gene.

Our findings highlight that PU.1+IRF8 co-LREs can be transformed into active enhancers activated by IL-4 stimulus. Next, we asked whether both PU.1 and IRF8 have essential role at these sites in properly mediating the gene expression program of STAT6. Using gain and loss of function experimental systems we set out to reveal the possible role of PU.1+IRF8 co-LREs in gene regulation. First, we measured the mRNA level of three well-known IL-4 regulated genes (*Retnla*, *Hbegf*, and *Arg1*) in PUER cells (Pu.1<sup>-/-</sup> myeloid progenitor cells containing a PU.1-estrogen receptor ligand binding domain fusion protein), whose transcriptional activity can be turned on by adding tamoxifen. Our results in the PUER system demonstrated that transcriptionally active PU.1 up-regulated the expression of the three IL-4-target genes in the unstimulated state providing the context for efficient IL-4-mediated induction as well (Figure 23). To test the role of IRF8 in IL-4 regulated expression of the same three genes, we used wild



type (WT) and *Irf8*<sup>-/-</sup> BMDMs. Our result showed that in the lack of *Irf8* the inducibility by IL-4 treatment was partially (*Hbegf*, and *Arg1*) or completely (*Retnla*) impaired (Figure 23).



**Figure 23. Loss/gain of function for PU.1 and IRF8**

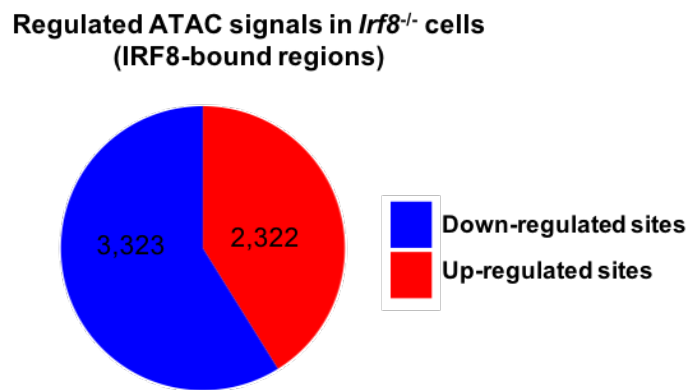
(1) Gene expression measurements by RT-qPCR at the mRNA level on the three alternative polarization marker genes (*Retnla*, *Hbegf* and *Arg1*) from PUER cells. PUER cells were exposed to tamoxifen for 24 hours (PUER w/ active PU.1) or left untreated (PUER w/ inactive PU.1) followed by 3 hours of IL-4 stimulation. The expression of mRNAs are normalized by the housekeeping gene *Ppia*. (2) Gene expression measurements of alternative polarization marker genes (*Retnla*, *Hbegf* and *Arg1*) by RT-qPCR at the mRNA level from wild type bone marrow-derived macrophages (WT BMDM) or *Irf8* deficient BMDMs (*Irf8*<sup>-/-</sup> BMDM) after exposed to IL-4 for 3 hours. Expression values are normalized to *Ppia*.

Taken together, PU.1+IRF8 co-LREs are present in the vicinity of alternative polarization marker genes that these sites have crucial role in IL-4 mediated gene activation. Both PU.1 and IRF8 had significant contribution in IL-4 induced gene expression. Future studies utilizing

genome editing techniques need to provide the mechanistic role of these co-LREs in mediating gene expression.

## 6.5. IRF8 maintains low accessible chromatin structure at a subset of labelled regulatory elements

Our previous results demonstrated that the IRF8 cistrome has a significant fraction of binding sites that bind to low accessible chromatin regions. We have also provided evidence on three genes that IRF8 is essential for IL-4-mediated gene regulation, but we had not investigated the consequence of its loss on chromatin openness. To answer this question, we carried out ATAC-seq experiments in unstimulated *Irf8*<sup>-/-</sup> macrophages. Analysis of distal IRF8-bound sites identified 3,323 regions with low accessible chromatin, while 2,322 sites gained openness in the absence of *Irf8* (Figure 24).
















**Figure 24. Up- and Down-regulated sites ATAC signals**

Pie chart showing the number of IRF8-bound sites that were up- or down-regulated in *Irf8*<sup>-/-</sup> cells. DiffBind was used to infer differential binding sites (p-value < 0.05) from duplicates using WT samples as control.

De novo motif analysis of the regions with decreased chromatin openness resulted in the emergence of typical macrophage-specific motifs we have seen in the wild type: TRE, PU.1, RUNX, CEBP, and a stronger EICE motif. In contrast, at the genomic regions where openness

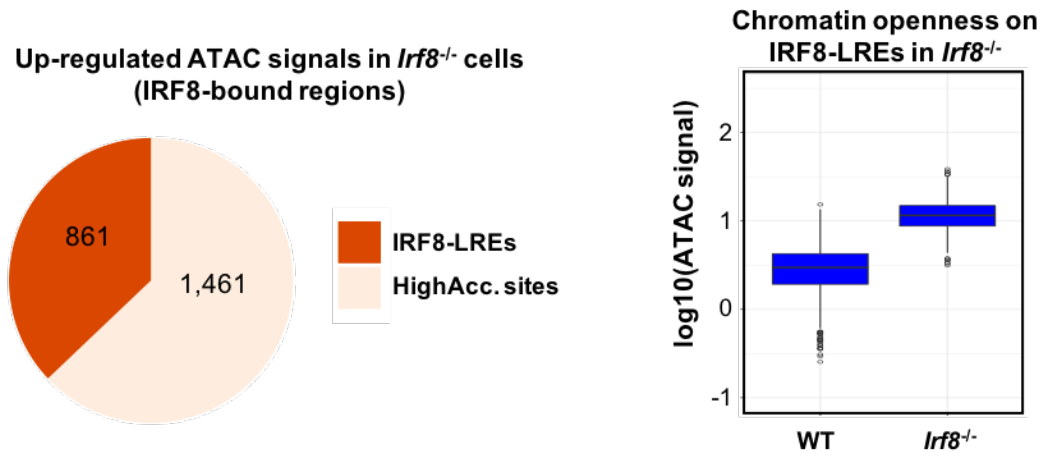
was increased, we neither identified this composite motif nor any other ISRE-like motifs. In contrast, CEBP, ETV (ETS variant 3) and NF- $\kappa$ B motifs showed moderate enrichment at these sets (typically with much weaker p-values) (Figure 25).

IRF8-bound regions		Motif logo	P-value	% of Target	% of Bg	Db. Annotation
Downregulated in IRF8 <sup>-/-</sup> cells			1e-787	46.18	5.91	TRE
			1e-517	40.01	7.18	PU.1
			1e-143	6.63	0.46	EICE
			1e-76	18.79	7.80	RUNX
			1e-61	7.64	1.93	MEF
			1e-58	12.05	4.47	CRE
			1e-25	8.21	3.79	CEBP
Upregulated in IRF8 <sup>-/-</sup> cells			1e-602	58.87	10.45	PU.1
			1e-247	36.19	8.75	CEBP
			1e-77	5.14	0.37	CTCF
			1e-76	15.88	4.74	RUNX
			1e-32	5.90	1.55	ETV
			1e-29	5.39	1.45	NfκB

**Figure 25. De novo motif analyses of up- and down-regulated sites in *Irf8*<sup>-/-</sup> cells**

The most enriched *de novo* motifs under IRF8-bound sites that were up- or down-regulated in *Irf8*<sup>-/-</sup> cells. Motif sequence logos, P-value of enrichment, percentage of target sequences with motif (Targets (%)), and percentage of background sequences with motif (Bg. (%)) are shown.

Interestingly, 37% of the sites that gained openness in the absence of *Irf8* overlapped with IRF8-LREs, suggesting that IRF8 binding is indispensable to prevent chromatin opening at these sites. This analysis indicates that IRF8 might have a chromatin compacting effect on a subset of loci, potentially through either indirect binding or binding to non-canonical motifs (Figure 26).



**Figure 26. Distribution of up-regulated ATAC signals in *Irf8*<sup>-/-</sup> cells**

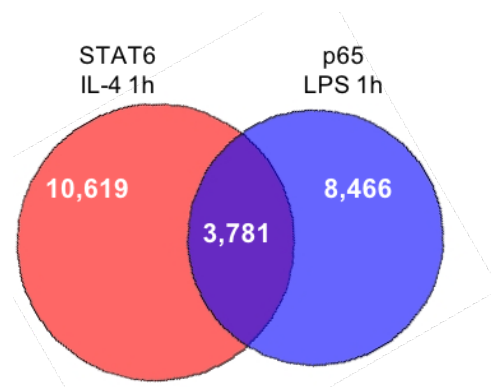
(1) Pie chart showing the number of IRF8-LREs and HighAcc. sites among IRF8-bound regions gaining openness in *Irf8*<sup>-/-</sup> cells. DiffBind was used to infer differential binding sites (p-value < 0.05) from duplicates using WT samples as control. (2) Box plot showing chromatin accessibility of IRF8-LREs in WT and *Irf8*<sup>-/-</sup> cells.

These findings highlight the importance of IRF8 as a regulator of chromatin and hints the existence of PU.1+IRF8-co-LREs and IRF8-LREs, where IRF8 may maintain low accessible chromatin, stabilizing the LRE state. Further studies need to investigate whether and how SDTFs act at these LREs in the absence of IRF8.

## **6.6. Labelled regulatory elements are dynamically utilized by macrophage polarization signals**

Finally, we asked whether and how classical and alternative macrophage polarization signals utilize LREs. To address this question, we compared the STAT6 (1h IL-4) and p65 (1h LPS) cistromes, the two main SDTFs of the alternative and classical polarization signals, respectively. The TF p65, a member of the NF-κB complex, which switches on the pro-

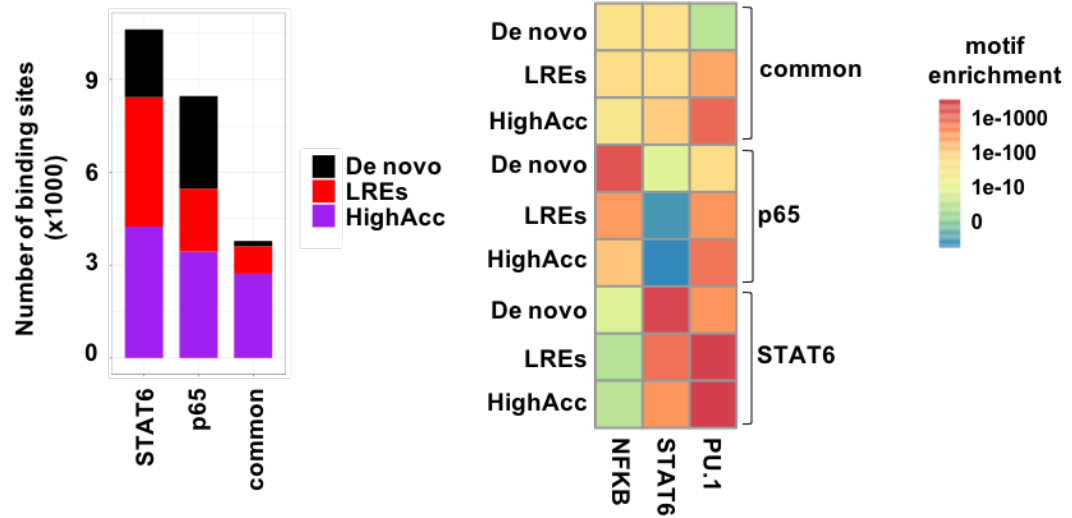
inflammatory program of macrophages by bacterial (LPS – lipopolysaccharide) stimulation. Comparison of STAT6 and p65 cistromes (Figure 27) revealed 10,619 STAT6-specific, 8,466 p65-specific regions and 3,781 co-bound regulatory elements that are utilized by both TFs upon the certain activating signals.



**Figure 27. Overlap between STAT6 (1h IL-4) and p65 (1h LPS) binding sites**

Venn diagram showing the overlap between STAT6 and p65 cistromes induced by IL-4 or LPS, respectively. DiffBind analysis were performed separately for STAT6 and p65 experiments to infer differential binding sites (p-value < 0.05) from duplicates using untreated samples as control.

Both cistromes of TF had similar fractions of de novo, LRE- and highly accessible regions. Co-bound genomic regions showed high overlap with highly accessible sites (2,720), but only 892 labelled and a very small set of de novo sites (169) were also identified (Figure 28).



**Figure 28. Characterization of STAT6 and p65 binding sites**

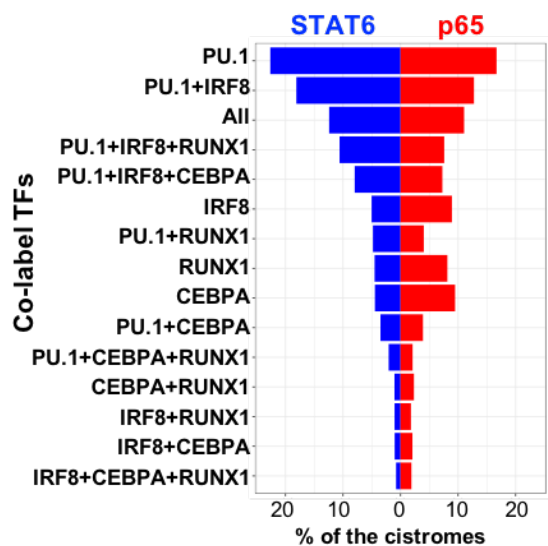
(1) Stacked bar plot showing the distribution of *de novo*, labelled regulatory elements (LREs) and highly accessible chromatin regions (HighAcc) across STAT6-specific, p65-specific and commonly-bound genomic regions by the two signal dependent transcription factors (SDTFs). (2) Heatmap showing the p-values of the motif (NFκB, STAT6 and PU.1) enrichments for the sets of *de novo*, LREs and HighAcc chromatin regions.

Motif analysis of NF-κB (represented by p65), STAT6 and PU.1 highlighted that (Figure 28): 1) the motif of PU.1 was weaker in almost all the cases at *de novo* sites and LREs compared to highly accessible sites, with no regard as to whether it overlapped by STAT6, p65 or both; 2) as expected, the p-value of the NF-κB motif was lower at p65-specific sites, and the *de novo* regions had the strongest motifs. Moreover, co-bound sites possessed a much weaker NF-κB motif, while this motif was virtually not detected at STAT6-only sites; 3) the STAT6 motif had very similar characteristics to the p65 motif, showing specificity to STAT6-bound regions with the highest p-values for this motif at the *de novo* sites.

LREs harbored similarly strong motifs to the *de novo* sites. In contrast, highly accessible chromatin regions had the least specific motifs. 4) Co-bound genomic regions by STAT6 and

p65 showed the weakest motifs, while 72% of these sites were highly accessible and 24% of these sites were LREs and only 4% were de novo.

These findings indicate that the capacity of SDTFs to bind and open chromatin is not completely sequence-specific, and high accessibility promotes additional TFs to bind LREs when the p-value of the motif is lower. Next we focused our attention to analyzing how LREs are utilized by STAT6 and p65. To address this issue, we mapped the binding sites of the two SDTFs to the identified TF-LREs (TF-LREs) (Figure 29).



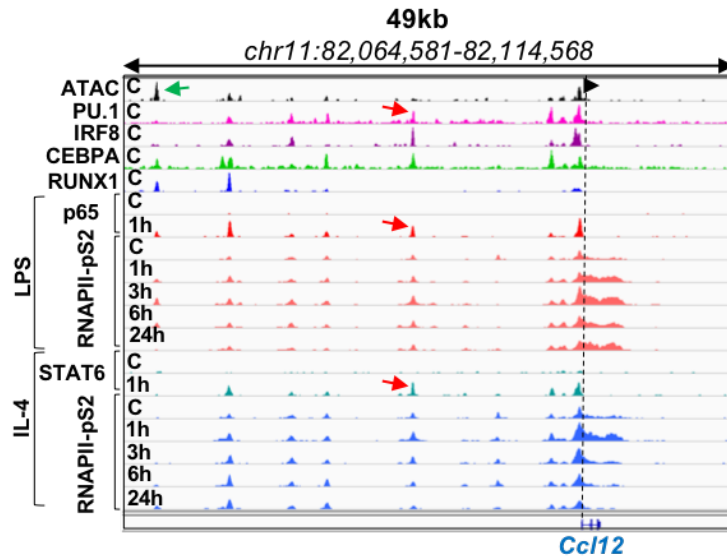
**Figure 29. Distribution of co-LREs on STAT6 and p65 binding sites**

Balance bar plot showing the distribution of the STAT6 and p65 cistromes on the different subsets of transcription factor co-labelled regulatory element groups. TF combinations that co-label these LREs are presented on the y-axis, while the x-axis depicts the percentages of the STAT6 and p65 cistromes that overlap with them.

The cistromes of both p65 and STAT6 highly overlapped with PU.1-, PU.1+IRF8- and PU.1+IRF8+CEPBA+RUNX1-LREs (“All” group), although a higher portion of the STAT6 cistrome overlapped with these groups compared to p65’s. Interestingly, p65 showed higher enrichment for IRF8-, RUNX1- and CEBPA-LREs than STAT6, suggesting that in certain

cases, SDTFs preferentially use certain LRE groups (without regard to the number of the co-bound TFs). Collectively, these findings highlight that the SDTFs of the two main polarization programs bind LREs, where the SDTFs do not necessarily prefer co-LREs, and show preferential binding to certain TF modules (TF combinations) at LREs.

Our results highlight that LREs have a contribution to short-term polarization programs. To address the question whether co-LREs also have any role in intermediate or long-term transcriptional responses, we investigated the activated enhancer network of classical and alternative macrophage polarization programs using fine resolution, time course RNAPII-pS2 ChIP-seq experiments: we treated BMDMs with IL-4 and LPS for 1h, 3h, 6h and 24h. (see an example for the *Ccl12* locus in Figure 30, where an LRE is activated by both IL-4 and LPS.

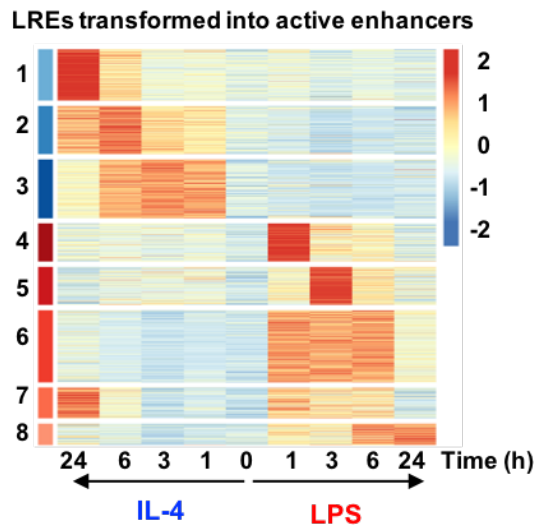


**Figure 30. IGV snapshot of *Ccl12* locus in the context of 1h IL-4 or LPS treatment**

IGV snapshot of *Ccl12* locus with a TF co-labelled regulatory element that can be used by both SDTFs (STAT6 and p65) and each of these two recruit the elongation-specific form of RNA polymerase II (RNAPII-pS2) with different kinetics upon LPS and IL-4 mediated macrophage polarization.



At the global scale, we identified 6,356 LREs that were up-regulated upon IL-4 or LPS treatment at least at one time point ( $FC > 2$  &  $p\text{-value} < 0.05$ , compared to the steady state BMDMs). Next, we clustered the up-regulated LREs into eight enhancer clusters (ECs), all of them with different transcriptional kinetics: EC1, EC2 and EC3 were IL4-specific, while clusters EC4-EC8 were specific for LPS (Figure 31). Having analyzed the cluster sizes, we found that LREs nearly equally contribute to early, intermediate and late induced TF modules.

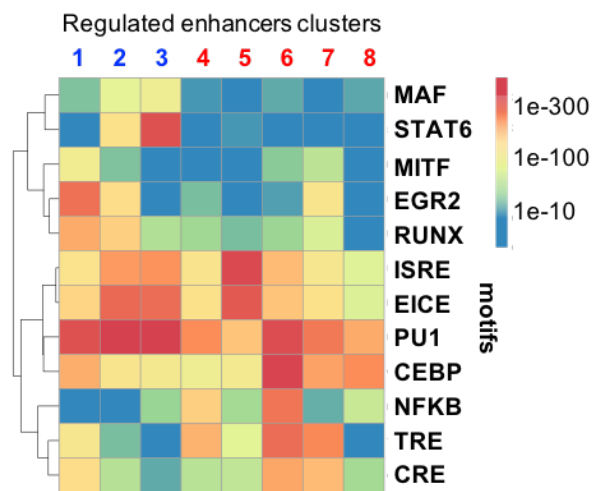


**Figure 31. Clustering of enhancers overlapping with LREs up-regulated by IL-4 or LPS**

Heatmap representation of the transformation of labelled regulatory elements (LREs) to active enhancers based on RNAPII-pS2 recruitment upon LPS or IL-4 mediated macrophage polarization on the time course of 1, 3, 6 and 24 hours. Regulated enhancer clusters (ECs) ( $FC > 2$  &  $p\text{-value} < 0.05$ ) and normalized Z-scores of RNAPII-pS2 signals are shown.

Next, we were wondering whether there are specific TF motifs that can be associated with certain TF modules. De novo motif analysis of the ECs revealed that the PU.1 motif showed high enrichment in all clusters. In contrast, the STAT6 motif was overrepresented in EC2 and EC3, the early and intermediate IL-4-specific TF modules, and NF- $\kappa$ B for EC4 and EC6, which were rapidly up-regulated by LPS at 1h. We also detected motifs of the EGR family in EC1

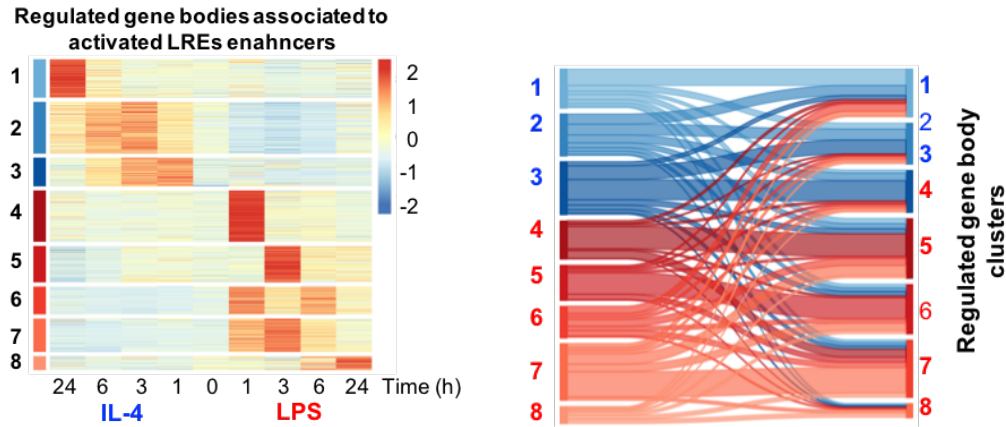
(late responsive IL-4 clusters) and the Microphthalmia-associated (MAF) TF family in EC2 and EC3 (IL-4-specific early and intermediate responsive modules), confirmed by recent works (Daniel et al., 2018; K. K. Kang et al., 2017), as well. We also identified MITF and RUNX under the IL-4 specific ECs. On the contrary, the motifs of TRE and CRE bound by certain AP-1 family members were highly enriched in the LPS intermediate (EC6 and EC7) modules while the motif of CEBP showed highly enrichment in the late responsive modules (EC8) as well. ISRE and EICE were overrepresented both in EC2 and EC3 (IL-4 specific early and intermediate ECs) and in EC5, an intermediate LPS-specific module (Figure 32).



**Figure 32. Motif enrichment of the regulated enhancer clusters**

Heatmap showing the p-values of the motifs enriched under at least one of the identified ECs.

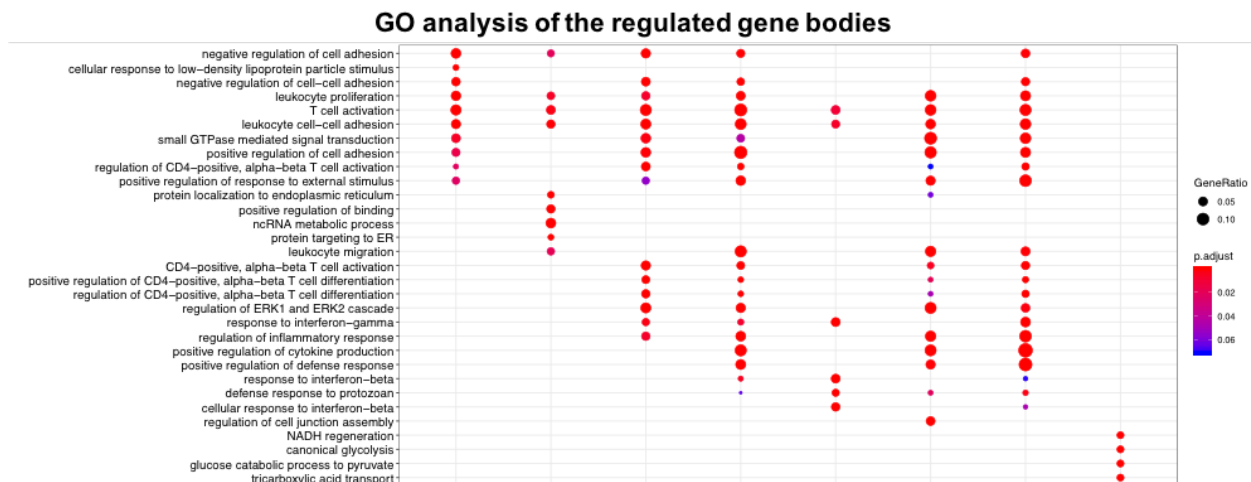
To reveal whether these co-LRE-associated ECs regulate gene expression with similar kinetics, we filtered the genes based on two conditions: 1) on the gene body, Pol II S2 signal was induced upon IL-4 or LPS compared to the resting BMDM state (1,735, FC>2 & p-value < 0.05) and 2) their TSSs was in the 100-kb vicinity of at least one regulated EC. Next, we clustered these genes into eight clusters (GBs), and they showed similar transcriptional kinetics (Figure 33).



**Figure 33. Association of regulated enhancers to regulated gene bodies**

(1) Heatmap showing the differentially RNAPII-pS2 enriched gene bodies (GBs) following IL-4 or LPS stimulations over the time course presented. DiffBind was used to infer differentially regulated ( $FC > 2$  &  $p\text{-value} < 0.05$ ) RNAPII-pS2 regions (1h, 3h, 6h, 24h IL4 and 1h, 3h, 6h, 24h LPS treatment) from duplicates using untreated samples as control. Normalized Z-scores of RNAPII-pS2 signals are shown. (2) Sankey plot showing the annotated gene-enhancer pairs from the 8 clusters defined for both the enhancers and gene bodies based on their different RNAPII-pS2 recruitment kinetics. The transcription start site of a gene body was associated to the regulated enhancer in a 200-kb-wide window.

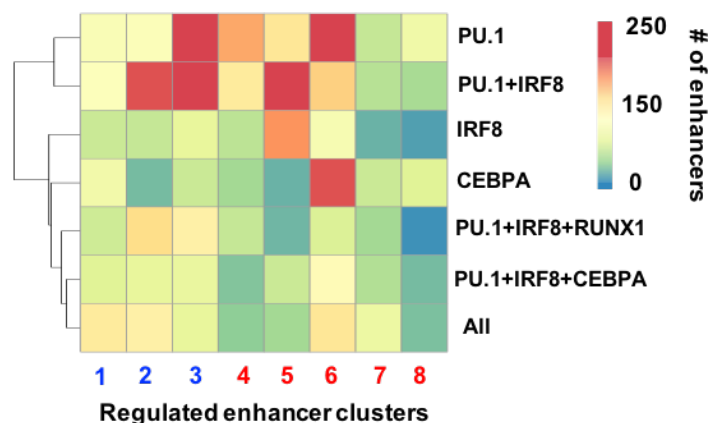
Finally we aimed to functionally annotate the regulated GBs using Biological Functions terms from the Gene Ontology database. GO analysis of GBs uncovered that IL-4-specific GBs showed enrichment for terms related to immune system modulation, such as ‘regulation of cell adhesion’ or ‘leukocyte migration’, more general terms including ‘response to external stimulus’ and selectively enriched for terms related to alternative polarization such as ‘CD4-positive alpha-beta T cell activation’. On the contrary, LPS-specific GBs showed enrichment for terms related to classical activation such as ‘regulation to interferon beta’ and ‘defense response’ (Figure 34).



**Figure 34. GO analysis of the regulated gene bodies**

Heatmap showing the enriched GO terms (Biological Processes) across GBs. The blue-red scale color contour represent p-values of the enrichments and the sizes of proportional to the ratio of the genes enriched for a certain term.

The motif enrichments were in agreement with the binding patterns of the highly enriched LREs (at least with 100 overlapping sites with any ECs): PU.1+IRF8+RUNX1 and IL-4 specific clusters had more co-bound regions and PU.1+IRF8+CEBPA and extensively overlapped with LPS-specific cluster (EC5), while ‘All’ LREs were moderately enriched for both IL-4 and LPS ECs. On the contrary, PU.1 and PU.1+IRF8 LREs binding showed high and specific enrichments in both IL-4 and LPS early and intermediate modules (EC2,3 and EC4,5,6, respectively). Interestingly, IRF8-LREs were highly enriched for EC5 (LPS-specific early responsive module). Similarly, CEBPA-LREs were overrepresented in the LPS-specific intermediate module (EC6) (Figure 35).



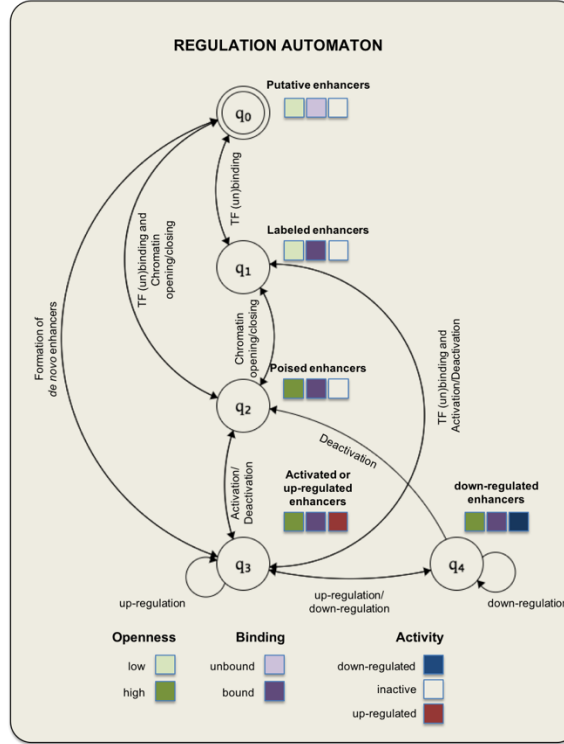
**Figure 35. Annotation of the regulated enhancers with co-LREs**

Heatmap depicting the highly enriched TF-LREs in the polarization induced enhancer clusters determined on panel E. TF-LREs at least with 100 overlapping sites are shown.

Collectively, our study confirms that LREs are dynamically utilized upon SDTF binding and activation both in the classical and alternative polarization programs of macrophages. In summary, LREs are widespread both macrophages, extending the working model of enhancer formation and providing a novel class of the regulatory elements which contribute to dynamic gene expression regulation.

## 6.7. Modelling enhancer states using Nondeterministic Finite State Automata

To provide a formal model to describe the states of the chromatin and the transitions among them, we defined a Nondeterministic Finite Automaton. In our case, Q is the set of the possible states of chromatin which are determined by three components: The first one is Binding, representing whether or not the given region is bound by any transcription factors ('bound', 'unbound'), the second component is Openness ('low' or 'high' accessibility) determined by ATAC-seq signal and the third one is Activity that is represented by the change in active histone mark signal ('down-regulated', 'inactive', 'up-regulated'), as shown in Figure 36.



**Figure 36. Regulation Automaton.**

An abstract model of enhancer formation. A Nondeterministic Finite state machine for modeling enhancer formation.  $q_i$  represent the different states and lines represent the possible transitions between them. The possible states are determined by three components: Binding ('bound', 'unbound'), Openness (low or highly accessible sites) and Activity ('down-regulated', 'inactive', 'up-regulated').

In the Regulation Automaton, input symbols represent the possible transcriptional events that could be connected to state changes of enhancer regions. Note that as the Binding component does not distinguish LDTFs and SDTFs, this automaton allows us to model the steady state as well as the IL-4-stimulated state. For instance,  $q_0$  represents the potential de novo enhancers that can be converted into  $q_1$  labelled enhancers upon PU.1 binding. Then, these enhancers become labelled sites characterized by low accessible chromatin and TF binding but inactive state. This can be further activated by binding collaborating TFs (CTFs), leading to recruitment of co-factors, RNA Pol II, etc. Generally,  $q_0$  represents the potential enhancers

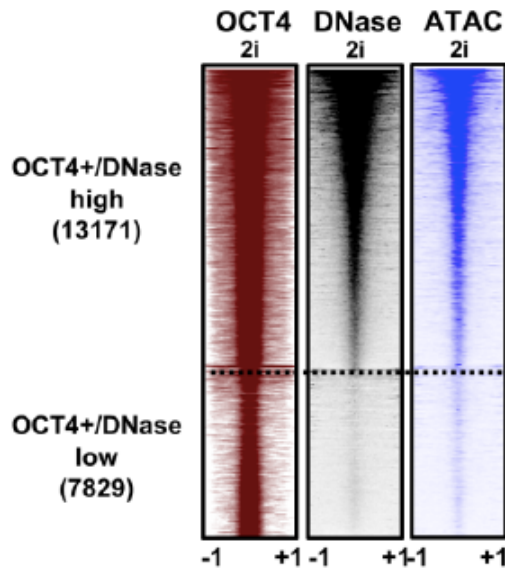
being activated by a certain signal. From this state, TF-binding events can lead to state  $q_1$ . In our case these sites would be called labelled sites. Further binding events of CTFs can convert these sites to the state  $q_2$ , which is still inactive but has the feature of open chromatin. This is termed primed or poised state (Creyghton et al., 2010). Further up-regulation can turn this state to  $q_3$ , which is a fully active binding site characterized by H4ac or other active marks as a result of co-activator recruitment. From this state, chromatin can turn to  $q_0$  or get down-regulated by repressive factors to  $q_4$  (Ostuni et al., 2013). This automaton can describe the interactions between the LDTF(s) and SDTF(s) in macrophages, and the results could serve as a new-paradigm which can be tested in other mammalian model systems.

## **6.8. OCT4-LREs in the context of RA-induced neurogenesis**

Our previous findings revealed that LREs are widespread in the macrophage genome and play an important role in both classical and alternative polarization programs. This finding raised the intriguing question whether LREs exist in a different cellular context as well and if so, these regulatory elements can also be activated in a signal specific manner. As we have seen before, OCT4 is one of the Yamanaka factors responsible for maintaining self-renewal and pluripotency in stem cells, similarly to PU.1 which, among others, has an essential role in maintaining cell identity in macrophages and in other immune cells. We hypothesized that there might exist a fraction of the OCT4 cistrome that is not associated with highly accessible chromatin in the undifferentiated state. ESCs can be cultivated in a “ground state” (2i inhibitor [2i]) or in a “naïve state” (serum + leukemia-inhibiting factor [LIF]) (Nichols & Smith, 2009), having unique gene expressional programs (Guo et al., 2016; Kolodziejczyk et al., 2015).

We sought to investigate this scenario by classifying OCT4 binding sites based on chromatin openness both in naïve (“serum”) and ground state (“2i”) ESCs. We identified

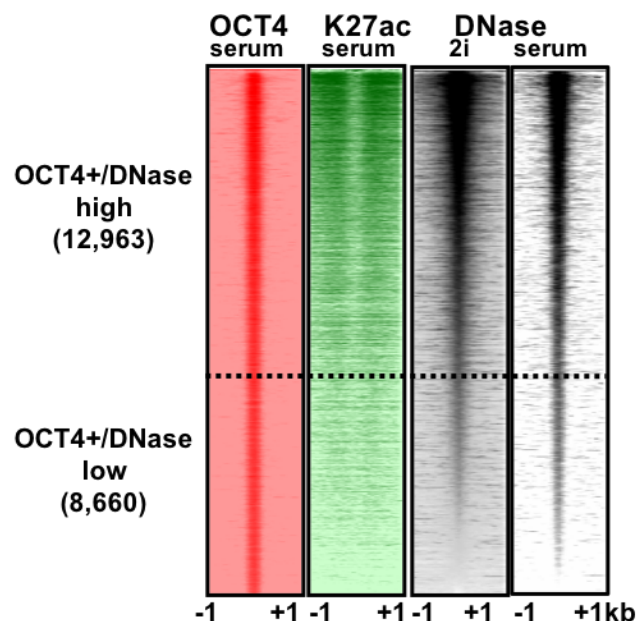
~21,000 OCT4 binding sites in both systems. Next we re-analyzed publicly available DNase-seq and ATAC-seq experiments and classified the OCT4 cistrome based on chromatin openness predicted by DNase-seq (Figure 37). Notably, 37% of the OCT4 cistrome in ground state was associated to low accessible chromatin (OCT4+/DNase high and OCT4+/DNase low). The DNase signal showed good correlation with ATAC-seq signals and this classification was in a good agreement in the naïve state as well (Figure 38).



**Figure 37. Classification of OCT4 binding sites based on chromatin openness (“2i”)**

Read distribution (RD) plots showing OCT4 occupancies, chromatin openness (DNase-seq, ATAC-seq) in the two groups clustered based on OCT4 and DNase-seq signal. The following nomenclature was used for the two clusters: OCT4+/DNase low (OCT4-LREs) and OCT4+/DNase high (OCT4 (+) HighAcc). RD plots show the signals in 2-kb windows around the summit of OCT4 or DNase peaks. OCT4 binding sites were classified in ground state of ESCs.





**Figure 38. Classification of OCT4 binding sites based on chromatin openness (“serum”)**

Read distribution (RD) plots showing OCT4 occupancies, chromatin openness (DNase-seq) and H3K27ac in the two groups clustered based on OCT4 and DNase-seq signal. The following nomenclature was used for the two clusters: OCT4+/DNase low (OCT4-LREs) and OCT4+/DNase high (OCT4 (+) HighAcc). RD plots show the signals in 2-kb windows around the summit of OCT4 or DNase peaks. OCT4 binding sites were classified in naïve state of ESCs.

To functionally annotate the newly identified OCT4-LREs, we performed a pathway analysis on the gene set associated with the OCT4+/DNase low genomic regions and found that these sites were enriched for WNT/b-catenin, Axonal Guidance Signaling, Epithelial Adherens Junction Signaling, PTEN Signaling, Signaling by Rho Family GTPases and RAR signaling pathways including WNT/b-catenin target genes such as Wnt3 and T/Brachyury and RXR:RAR including Hoxa1, Prmt8 and Cdx2 were enriched for OCT4 (Table 2).

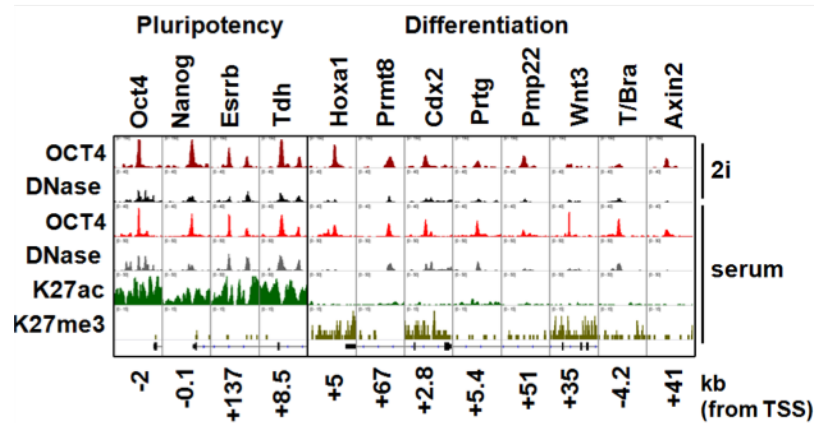
Canonical Pathways (OCT4+/DNase Low)	p-value
Wnt/ $\beta$ -catenin Signaling	3.33e-10

Axonal Guidance Signaling	1.36e-08
Epithelial Adherens Junction Signaling	1.36e-08
PTEN Signaling	3.15e-08
Signaling by Rho Family GTPases	4.52e-07
RAR Activation	8.00e-07

**Table 2. Pathway analysis of genes associated to OCT4+/DNase Low regions**

Upstream regulator analysis of genes associated to the OCT4/DNase Low regions. Pathway analysis was performed by Ingenuity.

These regions showed enrichment for the H3K27me3 (repressive mark) while they were negative for H3K27ac as shown in Figure 39, indicating that the differentiation-related genes are transcriptionally inactive in the unstimulated state.

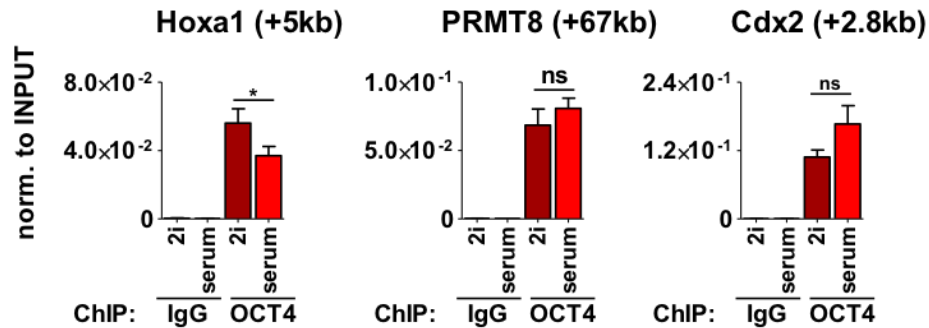


**Figure 39. IGV snapshots of putative OCT4 target genes**

Integrative Genomics Viewer (IGV) snapshot of regulatory regions in the vicinity of pluripotency- and differentiation related genes. Distances of the regulatory regions from the TSS of the genes are also indicated.

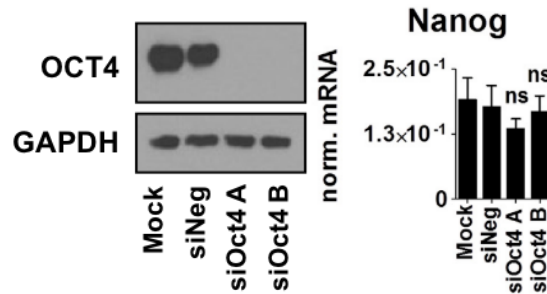
Next, we validated the OCT4 binding on the putative “OCT4/DNase Low” regulatory regions by ChIP-qPCR both in naïve and ground states. Our measurements confirmed the

binding of OCT4 to these regions (Figure 40) and hints the existence of a yet unknown role of OCT4 in regulating differentiation-specific pathways such as RAR signaling.



**Figure 40. ChIP-qPCR measurements of putative enhancer regions of OCT4 target genes**

To test whether OCT4 is necessary for the proper regulation of these genes, first we performed loss of function experiments using small interfering RNA (siRNA) knockdown of OCT4. The efficiency of OCT4 knockdown was high and the depletion of OCT4 did not substantially affect the mRNA level of NANOG.

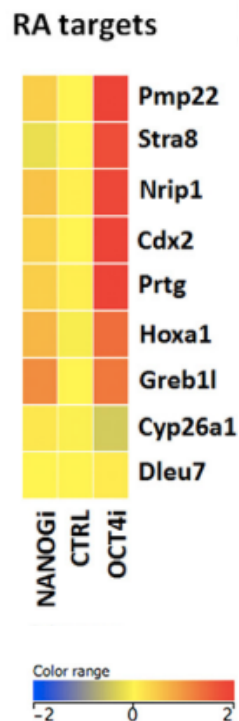


**Figure 41. Western blot validation of efficiency of OCT4 siRNA experiments**

(1) Western blot assay showing protein expression of OCT4 from untransfected cells (Mock) or transfected cells with siNeg (non-specific) or siOCT4 A/B (OCT4-specific siRNAs). (2) Bar plot showing the mRNA level of Nanog measured by qPCR 24 hr after RNAi transfection.

To examine the biological function of OCT4 at the labelled regulatory regions in the vicinity of RXR:RAR target genes, we investigated the expression changes of RA target genes upon OCT4 depletion. Next, we analyzed publicly available OCT4i and NANOGi microarray data

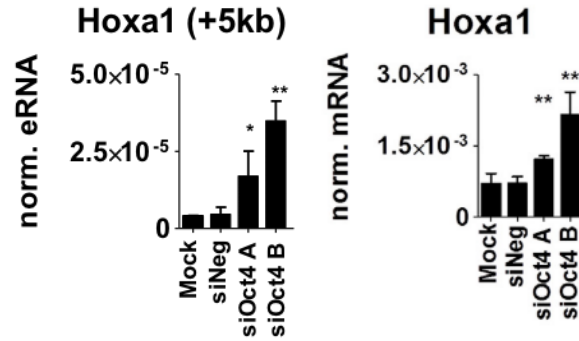
sets. Interestingly, upon the knockdown of OCT4 RA target genes, higher basal mRNA expression could be observed (Figure 42). In contrast, the depletion of Nanog did not substantially increase the mRNA level of these genes. This result confirms that there exists an OCT4-related mechanism by which RA-regulated genes such as Hoxa1 are suppressed, and this regulation mechanism is independent from the canonical pluripotency gene expression network.



**Figure 42. The effect of NANOG or OCT4 depletion on RA target genes**

Heatmap showing the quantile normalized expression values of RA target genes in NANOGi, OCT4i and control samples. Average values over replicates are shown.

qPCR measurements confirmed that OCT4 represses the mRNA level of the Hoxa1 gene and the eRNA of a putative enhancer 5 kb upstream of the TSS of Hoxa1 bound by OCT4 in the pluripotent state (Figure 43).

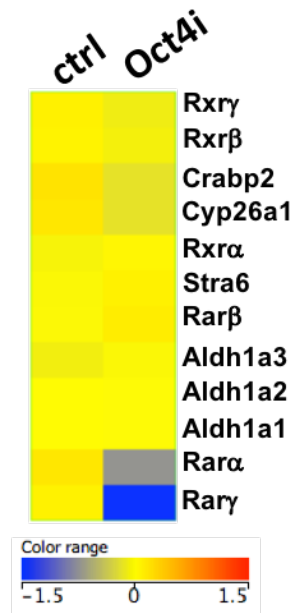


**Figure 43. Regulation of the mRNA level of Hoxa1 gene and its putative enhancer region**

Bar plot showing the mRNA level of Rary and eRNA level of its putative enhancer region measured by qPCR 24 hr after RNAi transfection.

RA signaling is known to regulate the Hox gene cluster, which is involved in the placement of hindbrain segments and specify the positional identities on the anterior-posterior axis for the cells during differentiation (Mallo, Wellik, & Deschamps, 2010). Due to its important function in neurogenesis, we chose Hoxa1 as a model gene for further investigation.

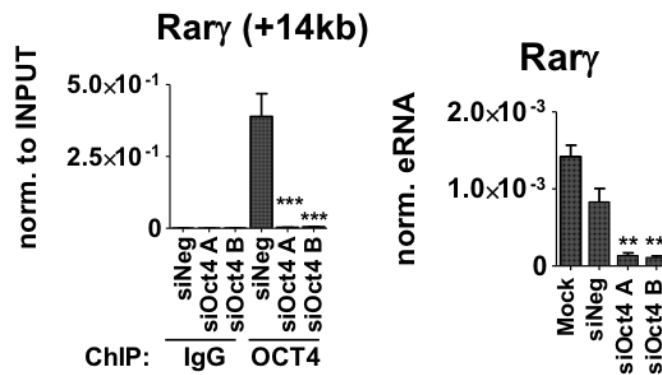
Next, we aimed to characterize the interaction between OCT4 and the RA pathway components. As shown in Figure 44, only Rar $\alpha$  and Rar $\gamma$  were down-regulated at the mRNA level upon the depletion of OCT4, however the fold change was much higher for Rar $\gamma$ .



**Figure 44. RA signaling pathway components in WT and OCT4i cells**

Heatmap showing the quantile normalized expression values of RA signaling pathway components in OCT4i and control samples. Average values over replicates are shown.

qPCR measurements of OCT4 binding and the eRNA level at the putative regulatory region of Rary after RNAi transfection revealed that OCT4 regulates Rary in the pluripotent state via direct binding to its putative enhancer region (Figure 45).

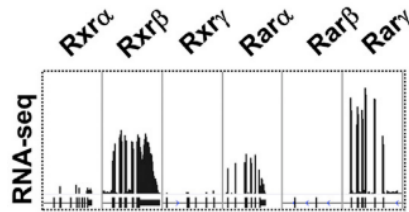


**Figure 45. Regulation of the mRNA level of Rary gene and its putative enhancer region**

Bar plot showing the OCT4 binding and the eRNA level of its putative enhancer region measured by qPCR 24 hr after RNAi transfection.

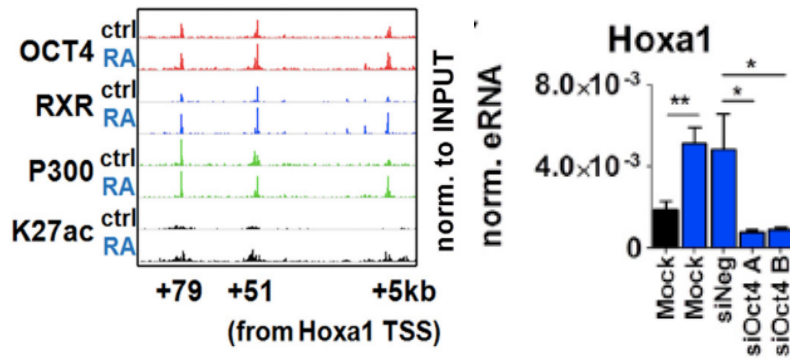
## 6.9. Modelling OCT4-related transcriptional circuits using network motifs

As we have seen that OCT4 positively contributes to gene expression in the early steps of RA-induced neurogenesis besides its critical role in maintaining the pluripotent state of ESCs. Strikingly, the depletion of OCT4 selectively down-regulated the mRNA level of Rar $\gamma$  from the RA signaling pathway. RAR forms heterodimers with RXR; having analyzed the mRNA level of the possible RXR and RAR isotypes, Rxr $\beta$  and Rar $\gamma$  were found to be the dominant heterodimer partners that most likely mediate the RA signaling in ESCs (Figure 46).



**Figure 46. mRNA level of Rxr and Rar isotypes in ESCs**

Moreover, our results show that RA up-regulate the mRNA level of Hoxa1 gene via distal regulatory elements in collaboration with OCT4. These enhancer regions are often low accessible regions in the pluripotent state and get activated by external stimuli such as RA treatment. Some enhancers were enriched for p300 (coactivator with histone acetyl transferase activity) prior to stimulus, and the activated RXR:RAR heterodimer further activated them (Figure 47).



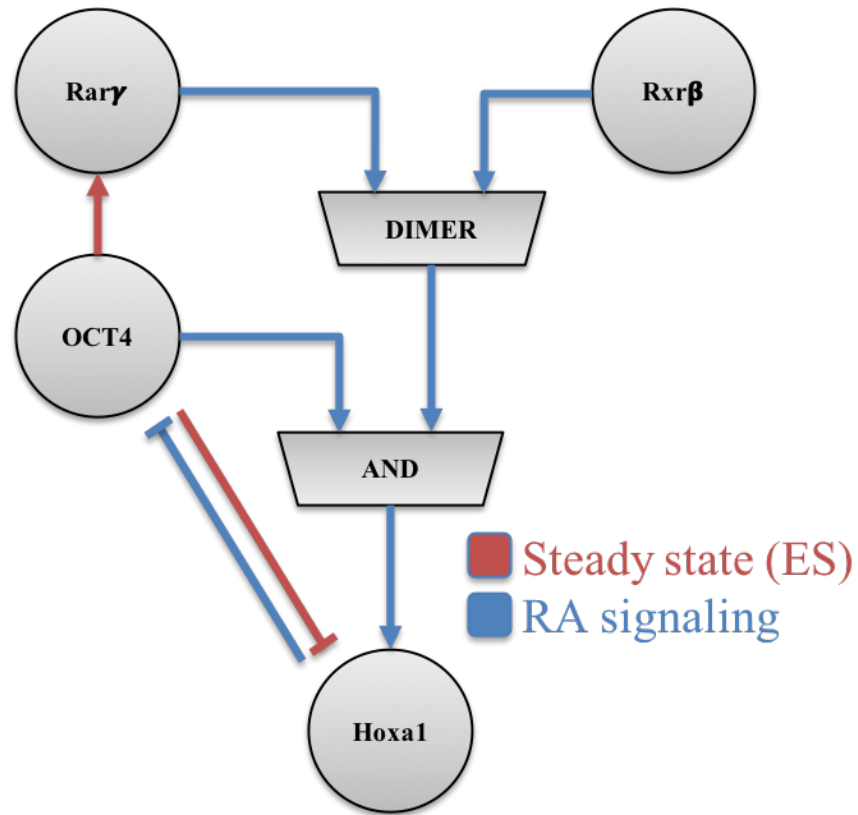
**Figure 47. Hoxa1 locus with three RA-sensitive enhancers**

(1) Integrative Genomics Viewer (IGV) snapshot of Hoxa1 locus. Distances of the enhancers from the TSS are shown. (2) Bar plot showing the enhancer RNA level of one of its putative enhancer (+51 kb) region measured by qPCR 24 hr after RNAi transfection.

Hoxa1<sup>-/-</sup> ES cells showed a higher level of OCT4 compared to wild-type cells after RA treatment for 48 to 96 hrs. This finding implies that the expression level of OCT4 and Hoxa1 genes are mutually exclusive: either OCT4 is highly expressed in the unstimulated ESCs preventing the expression of Hoxa1 or Hoxa1 is turned on suppressing the OCT4 at the mRNA level (Martinez-Ceballos, Chambon, & Gudas, 2005).

This accumulated data led us to conceptualize our knowledge on the role of OCT4 in the early steps of RA-induced neurogenesis. We built a composite motif network consisting of two fundamental modules: (1) A C1-FFL in which OCT4 directly up-regulates the expression level of Rary in the ESC state and indirectly up-regulates Hoxa1 up RA treatment via collaborative binding of the RXR:RAR heterodimer (2) OCT4 and Hoxa1 implement a mutual repression motif where either OCT4 is turned on and Hoxa1 is turned off (ESC state) or vice versa (long term RA signaling) (Figure 48).





**Figure 48. OCT4-related composite network motifs in RA-induced ESCs**

Composite network was recreated based on (Simandi, Horvath, et al., 2016).

## 7. DISCUSSION

In the post-genomic era, one of the biggest challenge is to integrate, interpret, and ultimately create meaningful models based on accumulating Next Generation Sequencing data (Koboldt, Steinberg, Larson, Wilson, & Mardis, 2013), which require an increasing intellectual contribution from the computer science community. Therefore interdisciplinary teams need to be set up to understand complex molecular biology processes by building predictive models, which seem to be essential for deciphering the processes of cell functions, replication, transcription, and translation.

Machine learning methods such as Random Forest and Support Vector Regression serve as novel tools to validate hypotheses by building classifiers and regressive models that can not only make predictions, but also determine the relative contribution of the input variables. We have demonstrated that chromatin openness profiled by ATAC-seq can be predicted from the binding pattern of key TFs using Random Forest classifier and Support Vector Regressor models. These computational approaches revealed that (1) despite the fact that PU.1 is the most prominent LDTF of macrophages, neither its solo binding nor the solo binding of other TFs are not sufficient to predict OCRs, and (2) the whole TF panel we used in the machine learning approach predicted open chromatin with a very high (82%) accuracy, suggesting that the most relevant TFs were included in the model.

This finding led us to the observation that more than half of the PU.1 binding sites do not overlap with OCRs. Therefore we examined the potential function of this type of PU.1 binding sites in classical and alternative polarization programs induced by LPS and IL-4, respectively. Our study revealed that a certain fraction of the ‘labelled’ sites is opened and activated by the binding of STAT6 (induced by IL-4 – alternative polarization) or p65 (induced by LPS -

classical polarization) in a signal-specific manner. A likely scenario is that in resting macrophages, the LDTF(s) attempt to establish open chromatin but no proper TF module is available to maintain open chromatin structure permanently (Voss & Hager, 2014). Further extension of our model will be possible when binding events can be examined at the single cell level (Brazda et al., 2011, 2014) and within the time-frame of milliseconds rather than within several minutes and by this way, the biases caused by population-based (bulk) techniques such as ChIP-seq can be avoided. Our results add another layer to the working model of enhancer activation and underline the importance of chromatin openness in shaping cell-type-specific enhancer repertoires and gene regulation. Collectively, our results lend support for a spectrum model according to which the transcriptional responses are tailored by the trade-off of the SDTFs and the available enhancer repertoire determined by the LDTFs.

First, on one end of the spectrum there are *de novo* enhancers strongly dependent on the corresponding SDTF in a sequence-specific manner and the LDTFs can neither open nor bind these sites in the resting state. Therefore, at these sites certain SDTFs seem to be mandatory factors to initiate and maintain any regulatory events. Second, the newly identified, labelled enhancers are bound by LDTF(s) in the unstimulated state, but the LDTF(s) do not have the ability to maintain open chromatin permanently. The third class of enhancers is the group of poised enhancers, which are already open before the stimuli meaning that they are strongly supported by the available TFs in the unstimulated state, but they need the SDTF(s) to become activated. In former studies, the binding of PU.1 and H3K4me1/2 was used to define enhancers and H3K27ac (active histone marks) to discriminate between poised and active enhancers (Creyghton et al., 2010; Ostuni et al., 2013). Fourth, the up-regulated, constitutively active enhancers can be driven without any stimuli and in some cases the binding of the SDTF(s) can

induce further up-regulation. Our results show that this group of enhancers requires a lower level of sequence-specificity compared to the *de novo* or labelled enhancers, and this raises the possibility that open chromatin has a distinct role in recruiting SDTFs via less-specific DNA binding and/or protein-protein interactions.

The need for formal models capturing the essence of a complex biological process appeared in parallel with the emergence of experimental molecular biology, even before the NGS revolution. During the modelling process, the most critical step is abstraction, by which we eliminate the unnecessary components and highlight the critical features of the phenomena to be modeled. The Operon-model was the first well-characterized regulatory process representing a very effective mechanism of prokaryotic gene regulation by which the organism can rapidly adapt to environmental stimuli (Jacob & Monod, 1961). Another example, which applied Automata Theory approach, is the Chemoton model (Gánti, 1975) proposing an automaton to formalize the essential properties of life (self-replication, metabolism and a bilayer membrane). In this study, we constructed a Nondeterministic Finite Automaton termed Regulation Automaton capturing the essential steps of enhancer formation using only the attributes Binding, Openness and Activity.

Interestingly, we found a similar mechanism in embryonic stem cells. The pluripotency factor OCT4 occupies genomic regions associated with the genes of signaling pathways such as RA signaling that are inactive in the ESC state. These regions are typically low accessible and often positive for the repressive histone mark H3K27me3. Upon certain stimuli, however, these regions will be bound by SDTFs which subsequently activate them by recruiting co-activators, similarly as co-LREs collaborate with STAT6 or p65 upon certain polarization signals in macrophages (Simandi, Nagy, et al., 2016). Notably, OCT4 maintains the gene

expression of Rarg in the unstimulated state while at low accessible regions it serves as a repressor of RA-target genes such as Hoxa1. In this regard, OCT4 is similar to IRF8, which also seems to have such a function. However, delineation of the repressive role of IRF8 at low accessible regions needs further studies. To formally describe this complex OCT4-related regulatory circuit, we built a composite network from the combination of two characteristic network motifs; (1) an C1-FFL depicting the direct and positive regulation of Rarg in the unstimulated state and the indirect up-regulation of Hoxa1 through by RXR:RAR heterodimer, (2) the mutual repression of OCT4 and Hoxa1 genes providing a regulatory switch between one of the most prominent pluripotency factor and the key regulator of early steps of neurogenesis essential to properly pattern the early mouse hindbrain and the associated neural crest. By using such modelling approaches, these studies can serve as building blocks for creating a firmly established theory of enhancer formation.

## 8. SUMMARY

**Summary 1. We have determined the contribution of key macrophage TFs to chromatin openness and enhancer activation in steady state and polarized mouse macrophages.**

Our findings include

- Chromatin openness can be accurately predicted from the binding pattern of key TFs using machine learning methods such Random Forest and Support Vector Regressor both qualitatively and quantitatively.
- The machine learning-based classification has also revealed that more than half of the PU.1 cistrome is associated to low accessible chromatin regions termed PU-labelled regulatory elements (LREs). Moreover our results show that having a remarkable fraction of LREs is a general phenomenon among the studied TFs (IRF8, CEBPA and RUNX1).
- Loss/gain of function experiments for PU.1 (PUER system) and IRF8 (*Irf8*<sup>-/-</sup>) has shown that PU.1 and IRF8 have an indispensable role in regulating gene expression in the steady state and/or in response to IL-4.
- There are indeed distinct TF modules collaboratively binding labelled regulatory elements (co-LREs) that regulate specific gene expression programs with different dynamics initiated by various macrophage polarizing stimuli including LPS (classical activation) and IL-4 (alternative activation).
- Identification of ~2,300 IRF8-LREs that gained openness in *Irf8*<sup>-/-</sup> cells suggesting a repressive role for IRF8 at these LREs.

- A formal model termed Regulation Automaton describing the possible states of enhancer formation and transitions among them.

**Summary 2. We have characterized a novel role of OCT4 in the early steps of RA-induced neurogenesis.**

Our findings include:

- Characterization of the OCT4 cistrome and its relation to chromatin openness has revealed that there is a remarkably fraction of OCT4 binding that are not associated to open chromatin (OCT4-labelled regulatory elements – OCT4-LREs) both in naïve and ground state.
- Examination of the interaction between OCT4-LREs and retinoic acid signaling pathway has uncovered that OCT4 plays a critical role in not only maintaining pluripotency state but also in the early steps of neuronal differentiation.
- siRNA mediated knockdown of OCT4 has shown that OCT4 is essential to activate differentiation-related enhancers of retinoic acid target genes such as Hoxa1.
- Construction of a composite network built from stereotypical network motifs describing the dual role of OCT4; maintaining the expression of Rxry and repressing Hoxa1 in pluripotency state and being essential role in mediating the up-regulation of Hoxa1 via RXR:RAR response elements upon 24h RA treatment.

## 9. ÖSSZEFOGLALÁS

**Összegzés 1. Meghatároztuk a kulcs makrofág transzkripciós faktorok hozzájárulását a kromatin nyitottsághoz és enhanszer aktivációhoz mind nyugvó mind polarizált egér makrofágokban.**

Eredményeink alapján a következő megállapításokat tettük:

- A Random Forest és Support Vector Regressor gépi tanuló eljárások alkalmazásával a kromatin nyitottsága nagy pontossággal megjósolható a kulcs TF-ok kötési mintázatából függetlenül attól, hogy “magas nyitottságú” és “alacsony nyitottságú” csoportokba osztjuk a kötőhelyeket (osztályozási feladat) vagy folytonos értéként kezeljük a nyitottságot (regressziós feladat).
- Az osztályozó gépi tanuló eljárás alkalmazása arra is rámutatott, hogy a PU.1 kötőhelyek több mint fele alacsony nyitottságú genomi régiót köt, amelyeket *jelölt szabályozó elemek*nek neveztünk el. Eredményeink alapján azt is elmondhatjuk, hogy ez általános jelenség a többi kulcs TF (IRF8, CEBPA és RUNX1) esetében is.
- A PUER, valamint az *Irf8*<sup>-/-</sup> sejtekben elvégzett kísérletek bebizonyították, hogy mind a PU.1-nak, mind az IRF8-nak elengedhetetlen szerepe van a nyugvó sejtekben zajló és az IL-4-függő génexpressziós szabályzásban is.
- Számos olyan TF modul létezik, amelyek tagjai együtt kötnek *jelölt szabályozó elemeket*. Ezek a modulok különböző módon válaszolnak a klasszikus (LPS) és alternatív (IL-4) polarizációs stimulusokra és specifikus géncsoportokat szabályoznak eltérő kinetikával.
- Az IRF8 hiányos sejtekben nyitottá váló mintegy 2 300 IRF8-Jelölt Szabályozó Elem az IRF8 represszor szerepére utal.



- Az általunk megkonstruált Szabályozó Automatának elnevezett formális modell segítségével leírhatóak a már ismert és az újonnan azonosított enhanszertípusok és a köztük előforduló állapotátmenetek.

## **Összegzés 2. Feltártuk az OCT4 transzkripciós faktor eddig nem ismert szerepét a neuronális differenciáció korai szakaszában**

Eredményeink alapján a következő megállapításokat tettük:

- Genomskálájú vizsgálataink megmutatták, hogy az OCT4 kötőhelyeinek egy jelentős része alacsony nyitottságú genomi régiót köt (OCT4-jelölt szabályozó elemek) mind az ún. “naiv” mind az ún. “alapállapotú” embrionális őssejtekben.
- Az OCT4-jelölt szabályozó elemek és a retinsav útvonal összefüggéseink vizsgálata rámutatott, hogy az OCT4 nemcsak a pluripotens állapot fenntartásában, hanem neuronális differenciáció korai szakaszában is jelentős szerepet játszik.
- Az OCT4 RNS szintű csendesítése megmutatta, hogy az OCT4 hozzájárul fontos gének (pl. Hoxa1) differenciációval összefüggő szabályzó elemeinek aktiválásához.
- Egy összetett motívumhálózat megalkotásával szemléltettük az OCT4 kettős szerepét: egyrészt a pluripotens állapotban fenntartja az Rxry szintjét és represszálja a Hoxa1 gént, másrészt 24 órás retinsav stimulus esetén elengethetetlen a Hoxa1 gén expressziós szintjének növeléséhez.

## 10. TABLE OF FIGURES

Figure 1. Various type of promoter elements.....	11
Figure 2. Classification of enhancer states in macrophages.....	15
Figure 3. Basic types of network motifs .....	19
Figure 4. An artistic Turing machine.....	20
Figure 5. An example FSA modeling an ATM machine .....	22
Figure 6. Early stage of RA-induced neurogenesis .....	29
Figure 7. BMDM model system .....	40
Figure 8. De novo motif analysis of distal OCRs in BMDMs .....	41
Figure 9. Comparison of methods profiling open chromatin regions.....	42
Figure 10. Flow chart of the machine learning approach.....	43
Figure 11. Prediction accuracies of the tested models .....	44
Figure 12. The results of Random Forest approach on the ‘full model’ .....	45
Figure 13. The results of Support Vector Regressor analysis .....	46
Figure 14. Binding hierarchy of PU.1 and other key TFs in macrophages.....	47
Figure 15. Interrelationship of chromatin openness and PU.1 binding in BMDMs .....	48
Figure 16. Various epigenomics marks across the three groups .....	49
Figure 17. Comparison of the four categories with the number of co-bound TFs .....	50
Figure 18. ChromHMM analysis of the key macrophage TFs and various histone marks.....	51
Figure 19. Characterization of LREs on the cistromes of key TFs.....	52
Figure 20. The key regulators form Co-LREs are widespread in macrophages.....	53
Figure 21. Examples for different co-LREs .....	54
Figure 22. IGV snapshots of alternative macrophage polarization marker genes.....	56

Figure 23. Loss/gain of function for PU.1 and IRF8 .....	57
Figure 24. Up- and Down-regulated sites ATAC signals .....	58
Figure 25. De novo motif analyses of up- and down-regulated sites in <i>Irf8</i> <sup>-/-</sup> cells.....	59
Figure 26. Distribution of up-regulated ATAC signals in <i>Irf8</i> <sup>-/-</sup> cells .....	60
Figure 27. Overlap between STAT6 (1h IL-4) and p65 (1h LPS) binding sites .....	61
Figure 28. Characterization of STAT6 and p65 binding sites.....	62
Figure 29. Distribution of co-LREs on STAT6 and p65 binding sites .....	63
Figure 30. IGV snapshot of <i>Ccl12</i> locus in the context of 1h IL-4 or LPS treatment .....	64
Figure 31. Clustering of enhancers overlapping with LREs up-regulated by IL-4 or LPS.....	65
Figure 32. Motif enrichment of the regulated enhancer clusters .....	66
Figure 33. Association of regulated enhancers to regulated gene bodies .....	67
Figure 34. GO analysis of the regulated gene bodies.....	68
Figure 35. Annotation of the regulated enhancers with co-LREs.....	69
Figure 36. Regulation Automaton.....	70
Figure 37. Classification of OCT4 binding sites based on chromatin openness (“2i”) .....	72
Figure 38. Classification of OCT4 binding sites based on chromatin openness (“serum”)....	73
Figure 39. IGV snapshots of putative OCT4 target genes .....	74
Figure 40. ChIP-qPCR measurements of OCT4 target genes .....	75
Figure 41. Western blot validation of efficiency of OCT4 siRNA experiments .....	75
Figure 42. The effect of NANOG or OCT4 depletion on RA target genes .....	76
Figure 43. Regulation of the mRNA level of <i>Hoxa1</i> gene and its putative enhancer region..	77
Figure 44. RA signaling pathway components in WT and OCT4i cells.....	78
Figure 45. Regulation of the mRNA level of <i>Rarg</i> gene and its putative enhancer region.....	78

Figure 46. mRNA level of Rxr and Rar isotypes in ESCs .....	79
Figure 47. Hoxa1 locus with three RA-sensitive enhancers.....	80
Figure 48. OCT4-related composite network motifs in RA-induced ESCs.....	81

## 11. REFERENCES

- Aderem, A., & Underhill, D. M. (1999). Mechanisms of phagocytosis in macrophages. *Annu Rev Immunol*. <https://doi.org/10.1146/annurev.immunol.17.1.593>
- Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6), 450–461. <https://doi.org/10.1038/nrg2102>
- Barish, G. D., Yu, R. T., Karunasiri, M., Ocampo, C. B., Dixon, J., Benner, C., ... Evans, R. M. (2010). Bcl-6 and NF-kappaB cistromes mediate opposing regulation of the innate immune response. *Genes Dev*, 24(24), 2760–2765. <https://doi.org/10.1101/gad.1998010>
- Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., & Natoli, G. (2014). Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers. *Molecular Cell*, 54(5), 844–857. <https://doi.org/10.1016/j.molcel.2014.04.006>
- Barta, E. (2011). Command line analysis of ChIP-seq results. *EMBnet. Journal*, 17(1), 13–17. <https://doi.org/10.14806/ej.17.1.209>
- Brazda, P., Krieger, J., Daniel, B., Jonas, D., Szekeres, T., Langowski, J., ... Vamosi, G. (2014). Ligand binding shifts highly mobile retinoid X receptor to the chromatin-bound state in a coactivator-dependent manner, as revealed by single-cell imaging. *Mol Cell Biol*, 34(7), 1234–1245. <https://doi.org/10.1128/MCB.01097-13>
- Brazda, P., Szekeres, T., Bravics, B., Toth, K., Vamosi, G., & Nagy, L. (2011). Live-cell fluorescence correlation spectroscopy dissects the role of coregulator exchange and chromatin binding in retinoic acid receptor mobility. *J Cell Sci*, 124(Pt 21), 3631–3642. <https://doi.org/10.1242/jcs.086082>
- Butler, J. E. F., & Kadonaga, J. T. (2002). The RNA polymerase II core promoter: A key

- component in the regulation of gene expression. *Genes and Development*.  
<https://doi.org/10.1101/gad.1026202>
- Calo, E., & Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Mol Cell*, 49(5), 825–837. <https://doi.org/10.1016/j.molcel.2013.01.038>
- Chen, L., Xuan, J., Riggins, R. B., Wang, Y., Hoffman, E. P., & Clarke, R. (2010). Multilevel support vector regression analysis to identify condition-specific regulatory networks. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btq144>
- Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74. <https://doi.org/10.1038/nature11247>
- Cooper, G. M. (2000). *The Cell: A Molecular Approach*. 2nd edition. Sinauer Associates.  
<https://doi.org/10.1016/B978-0-12-387738-3.00003-2>
- Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909), 1845–1848.  
<https://doi.org/10.1126/science.1162228>
- Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., ... Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*, 107(50), 21931–21936.  
<https://doi.org/10.1073/pnas.1016071107>
- Czimmerer, Z., Daniel, B., Horvath, A., R  ckerl, D., Nagy, G., Kiss, M., ... Nagy, L. (2018a). The Transcription Factor STAT6 Mediates Direct Repression of Inflammatory Enhancers and Limits Activation of Alternatively Polarized Macrophages. *Immunity*.  
<https://doi.org/10.1016/j.immuni.2017.12.010>
- Czimmerer, Z., Daniel, B., Horvath, A., R  ckerl, D., Nagy, G., Kiss, M., ... Nagy, L. (2018b).

- The Transcription Factor STAT6 Mediates Direct Repression of Inflammatory Enhancers and Limits Activation of Alternatively Polarized Macrophages. *Immunity*, 48(1), 75–90.e6. <https://doi.org/10.1016/j.immuni.2017.12.010>
- Daniel, B., Nagy, G., Hah, N., Horvath, A., Czimmerer, Z., Poliska, S., ... Nagy, L. (2014a). The active enhancer network operated by liganded RXR supports angiogenic activity in macrophages. *Genes and Development*. <https://doi.org/10.1101/gad.242685.114>
- Daniel, B., Nagy, G., Hah, N., Horvath, A., Czimmerer, Z., Poliska, S., ... Nagy, L. (2014b). The active enhancer network operated by liganded RXR supports angiogenic activity in macrophages. *Genes and Development*, 28(14), 1562–1577. <https://doi.org/10.1101/gad.242685>
- Daniel, B., Nagy, G., Horvath, A., Czimmerer, Z., Cuaranta-Monroy, I., Poliska, S., ... Nagy, L. (2018). The IL-4/STAT6/PPAR $\gamma$  signaling axis is driving the expansion of the RXR heterodimer cistrome, providing complex ligand responsiveness in macrophages. *Nucleic Acids Research*, 46(February), 4425–4439. <https://doi.org/10.1093/nar/gky157>
- Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*. <https://doi.org/10.1038/nmeth.1906>
- Feng, R., Desbordes, S. C., Xie, H., Tillo, E. S., Pixley, F., Stanley, E. R., & Graf, T. (2008). PU.1 and C/EBP $\alpha$ /beta convert fibroblasts into macrophage-like cells. *Proceedings of the National Academy of Sciences of the United States of America*, 105(16), 6057–6062. <https://doi.org/10.1073/pnas.0711961105>
- Fontana, M. F., Baccarella, A., Pancholi, N., Pufall, M. A., Herbert, D. R., & Kim, C. C. (2015). JUNB Is a Key Transcriptional Modulator of Macrophage Activation. *The Journal of Immunology*, 194(1), 177–186. <https://doi.org/10.4049/jimmunol.1401595>

- Frum, T., Halbisen, M. A., Wang, C., Amiri, H., Robson, P., & Ralston, A. (2013). Oct4 Cell-autonomously promotes primitive endoderm development in the mouse blastocyst. *Developmental Cell*. <https://doi.org/10.1016/j.devcel.2013.05.004>
- Gal, A., Balicza, P., Weaver, D., Naghdi, S., Joseph, S. K., Várnai, P., ... Hajnóczky, G. (2017). MSTO1 is a cytoplasmic pro-mitochondrial fusion protein, whose mutation induces myopathy and ataxia in humans. *EMBO Molecular Medicine*, 9(7), 1–18. <https://doi.org/10.15252/emmm.201607058>
- Gánti, T. (1975). Organization of chemical reactions into dividing and metabolizing units: The chemotons. *BioSystems*. [https://doi.org/10.1016/0303-2647\(75\)90038-6](https://doi.org/10.1016/0303-2647(75)90038-6)
- Garces de Los Fayos Alonso, I., Liang, H. C., Turner, S. D., Lager, S., Merkel, O., & Kenner, L. (2018). The role of activator protein-1 (AP-1) family members in CD30-positive lymphomas. *Cancers*. <https://doi.org/10.3390/cancers10040093>
- Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., ... Natoli, G. (2010). Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity*, 32(3), 317–328. <https://doi.org/10.1016/j.immuni.2010.02.008>
- Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., ... Natoli, G. (2010). Identification and Characterization of Enhancers Controlling the Inflammatory Gene Expression Program in Macrophages. *Immunity*. <https://doi.org/10.1016/j.immuni.2010.02.008>
- Gordon, S., & Martinez, F. O. (2010). Alternative activation of macrophages: mechanism and functions. *Immunity*, 32(5), 593–604. <https://doi.org/10.1016/j.immuni.2010.05.007>
- Guo, G., Pinello, L., Han, X., Lai, S., Shen, L., Lin, T. W., ... Orkin, S. H. (2016). Serum-Based



- Culture Conditions Provoke Gene Expression Variability in Mouse Embryonic Stem Cells as Revealed by Single-Cell Analysis. *Cell Reports*. <https://doi.org/10.1016/j.celrep.2015.12.089>
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., ... 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(4), 576–589. <https://doi.org/10.1016/j.molcel.2010.05.004>
- Heinz, S., Romanoski, C. E., Benner, C., Allison, K. A., Kaikkonen, M. U., Orozco, L. D., & Glass, C. K. (2013). Effect of natural genetic variation on enhancer selection and function. *Nature*, 503(7477), 487–492. <https://doi.org/10.1038/nature12615>
- Heinz, S., Romanoski, C. E., Benner, C., & Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*. <https://doi.org/10.1038/nrm3949>
- Hopcroft, J. E., Motwani, R., & Ullman, J. D. (2007). *Introduction to Automata Theory, Languages, and Computation*. Pearson/Addison Wesley. Retrieved from <https://books.google.hu/books?id=avUYAQAIAAJ>
- Iwafuchi-Doi, M., & Zaret, K. S. (2014). Pioneer transcription factors in cell reprogramming. *Genes & Development*. <https://doi.org/10.1101/gad.253443.114>
- Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3, 318–356. [https://doi.org/10.1016/S0022-2836\(61\)80072-7](https://doi.org/10.1016/S0022-2836(61)80072-7)
- Jacobs, S., Lie, D. C., Evans, R. M., DeCicco, K. L., Gage, F. H., DeLuca, L. M., & Shi, Y. (2006). Retinoic acid is required early during adult neurogenesis in the dentate gyrus. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.0511294103>
- Kaikkonen, M. U., Spann, N. J., Heinz, S., Romanoski, C. E., Allison, K. a., Stender, J. D., ...

- Glass, C. K. (2013). Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Molecular Cell*, 51(3), 310–325.  
<https://doi.org/10.1016/j.molcel.2013.07.010>
- Kang, K. K., Park, S. H., Chen, J., Qiao, Y., Giannopoulou, E., Berg, K., ... Ivashkiv, L. B. (2017). Interferon- $\gamma$  Represses M2 Gene Expression in Human Macrophages by Disassembling Enhancers Bound by the Transcription Factor MAF. *Immunity*.  
<https://doi.org/10.1016/j.immuni.2017.07.017>
- Kang, K., Park, S. H., Chen, J., Qiao, Y., Giannopoulou, E., Berg, K., ... Ivashkiv, L. B. (2017). Interferon-gamma Represses M2 Gene Expression in Human Macrophages by Disassembling Enhancers Bound by the Transcription Factor MAF. *Immunity*, 47(2), 235–250 e4.  
<https://doi.org/10.1016/j.immuni.2017.07.017>
- Kimura, H. (2013). Histone modifications for human epigenome analysis. *Journal of Human Genetics*. <https://doi.org/10.1038/jhg.2013.66>
- Kleene, S. C. (1956). Representation of Events in Nerve Nets and Finite Automata. *Automata Studies* %U [Http://Www.Diku.Dk/Hjemmesider/Ansatte/Henglein/Papers/Kleene1956.Pdf](http://www.Diku.Dk/Hjemmesider/Ansatte/Henglein/Papers/Kleene1956.Pdf).
- Koboldt, D. C. C., Steinberg, K. M. M., Larson, D. E. E., Wilson, R. K. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1), 27–38.  
<https://doi.org/10.1016/j.cell.2013.09.006>
- Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C. H., Ilicic, T., Henriksson, J., Natarajan, K. N., ... Teichmann, S. A. (2015). Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell*.  
<https://doi.org/10.1016/j.stem.2015.09.011>
- Krishnamurthy, S., & Hampsey, M. (2009). Eukaryotic transcription initiation. *Current Biology*.

<https://doi.org/10.1016/j.cub.2008.11.052>

- Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretzky, I., Jaitin, D. A., David, E., ... Amit, I. (2014). Chromatin state dynamics during blood formation. *Science*. <https://doi.org/10.1126/science.1256271>
- Lavin, Y., Winter, D., Blecher-Gonen, R., David, E., Keren-Shaul, H., Merad, M., ... Amit, I. (2014). Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. *Cell*. <https://doi.org/10.1016/j.cell.2014.11.018>
- Leddin, M., Perrod, C., Hoogenkamp, M., Ghani, S., Assi, S., Heinz, S., ... Rosenbauer, F. (2011). Two distinct auto-regulatory loops operate at the PU.1 locus in B cells and myeloid cells. *Blood*, 117(10), 2827–2838. <https://doi.org/10.1182/blood-2010-08-302976>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, W., Notani, D., & Rosenfeld, M. G. (2016). Enhancers as non-coding RNA transcription units: Recent insights and future perspectives. *Nature Reviews Genetics*. <https://doi.org/10.1038/nrg.2016.4>
- Link, V. M., Duttke, S. H., Chun, H. B., Holtman, I. R., Westin, E., Hoeksema, M. A., ... Glass, C. K. (2018). Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. *Cell*. <https://doi.org/10.1016/j.cell.2018.04.018>
- Liu, L., Jin, G., & Zhou, X. (2015). Modeling the relationship of epigenetic modifications to transcription factor binding. *Nucleic Acids Res*, 43(8), 3873–3885. <https://doi.org/10.1093/nar/gkv255>

- Mallo, M., Wellik, D. M., & Deschamps, J. (2010). Hox genes and regional patterning of the vertebrate body plan. *Developmental Biology*. <https://doi.org/10.1016/j.ydbio.2010.04.024>
- Mancino, A., Termanini, A., Barozzi, I., Ghisletti, S., Ostuni, R., Prosperini, E., ... Natoli, G. (2015). A dual cis -regulatory code links IRF8 to constitutive and inducible gene expression in macrophages, 1–16. <https://doi.org/10.1101/gad.257592.114>. GENES
- Mangan, S., & Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.2133841100>
- Martinez-Ceballos, E., Chambon, P., & Gudas, L. J. (2005). Differences in gene expression between wild type and Hoxa1 knockout embryonic stem cells after retinoic acid treatment or Leukemia Inhibitory Factor (LIF) removal. *Journal of Biological Chemistry*. <https://doi.org/10.1074/jbc.M414397200>
- McKercher, S. R., Torbett, B. E., Anderson, K. L., Henkel, G. W., Vestal, D. J., Baribault, H., ... Maki, R. A. (1996). Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *The EMBO Journal*. <https://doi.org/10.1002/j.1460-2075.1996.tb00949.x>
- Mosser, D. M., & Edwards, J. P. (2008). Exploring the full spectrum of macrophage activation. *Nature Reviews Immunology*. <https://doi.org/10.1038/nri2448>
- Müller, E., & Corthay, A. (2017). Toll-Like Receptor Ligands and Interferon-  $\gamma$  Synergize for Induction of Antitumor M1 Macrophages, 8(October). <https://doi.org/10.3389/fimmu.2017.01383>
- Murray, P. J., & Wynn, T. A. (2011). Protective and pathogenic functions of macrophage subsets. *Nature Reviews Immunology*. <https://doi.org/10.1038/nri3073>
- Murray, R. Z., & Stow, J. L. (2014). Cytokine secretion in macrophages: SNAREs, Rabs, and membrane trafficking. *Frontiers in Immunology*. <https://doi.org/10.3389/fimmu.2014.00538>

- Nerlov, C., & Graf, T. (1998). PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. *Genes and Development*, 12(15), 2403–2412. <https://doi.org/10.1101/gad.12.15.2403>
- Nichols, J., & Smith, A. (2009). Naive and Primed Pluripotent States. *Cell Stem Cell*. <https://doi.org/10.1016/j.stem.2009.05.015>
- Niwa, H., Burdon, T., Chambers, I., & Smith, A. (1998). Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes and Development*. <https://doi.org/10.1101/gad.12.13.2048>
- Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., ... Natoli, G. (2013). Latent enhancers activated by stimulation in differentiated cells. *Cell*, 152(1–2), 157–171. <https://doi.org/10.1016/j.cell.2012.12.018>
- Pál Dömösi, János Falucska, Géza Horváth, Zoltán Mecsei, B. N. (2002). Formális Nyelvek és Automaták, 1–235. Retrieved from <http://brahms.emu.edu.tr/benedeknagy/FormalisNyelvekAutomatak.pdf>
- Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: Five essential questions. *Nature Reviews Genetics*. <https://doi.org/10.1038/nrg3458>
- Piccolo, V., Curina, A., Genua, M., Ghisletti, S., Simonatto, M., Sabò, A., ... Natoli, G. (2017). Opposing macrophage polarization programs show extensive epigenomic and transcriptional cross-talk. *Nature Immunology*. <https://doi.org/10.1038/ni.3710>
- Rabin, M. O., & Scott, D. (1959). Finite automata and their decision problems. *IBM J. Res. Dev.*, 3(2), 114–125. <https://doi.org/10.1147/rd.32.0114>
- Radzisheuskaya, A., Le Bin Chia, G., Dos Santos, R. L., Theunissen, T. W., Castro, L. F. C., Nichols, J., & Silva, J. C. R. (2013). A defined Oct4 level governs cell state transitions of

- pluripotency entry and differentiation into all embryonic lineages. *Nature Cell Biology*.  
<https://doi.org/10.1038/ncb2742>
- Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., ... Carroll, J. S. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, *481*(7381), 389–393. <https://doi.org/10.1038/nature10730>
- Rozenberg, G. (1980). *The mathematical theory of L systems / Grzegorz Rozenberg, Arto Salomaa*. (A. Salomaa, Ed.), *Pure and applied mathematics (Academic Press)*; 90. New York: Academic Press.
- Samokhvalov, I. M., Samokhvalova, N. I., & Nishikawa, S. I. (2007). Cell tracing shows the contribution of the yolk sac to adult haematopoiesis. *Nature*, *446*(7139), 1056–1061.  
<https://doi.org/10.1038/nature05725>
- Scott, E. W., Simon, M. C., Anastasi, J., & Singh, H. (1994). Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science*.  
<https://doi.org/10.1126/science.8079170>
- Siersbaek, M. S., Loft, A., Aagaard, M. M., Nielsen, R., Schmidt, S. F., Petrovic, N., ... Mandrup, S. (2012). Genome-wide profiling of peroxisome proliferator-activated receptor gamma in primary epididymal, inguinal, and brown adipocytes reveals depot-selective binding correlated with gene expression. *Mol Cell Biol*, *32*(17), 3452–3463.  
<https://doi.org/10.1128/MCB.00526-12>
- Simandi, Z., Horvath, A., Wright, L. C., Cuaranta-Monroy, I., De Luca, I., Karolyi, K., ... Nagy, L. (2016). OCT4 Acts as an Integrator of Pluripotency and Signal-Induced Differentiation. *Molecular Cell*, *63*(4), 647–661. <https://doi.org/10.1016/j.molcel.2016.06.039>
- Simandi, Z., Nagy, L., Gudas, L. J., De Luca, I., Cowley, S. M., Deleuze, J.-F., ... Cuaranta-

- Monroy, I. (2016). OCT4 Acts as an Integrator of Pluripotency and Signal-Induced Differentiation. *Molecular Cell*, 63(4), 647–661. <https://doi.org/10.1016/j.molcel.2016.06.039>
- Thomas, K. E., Galligan, C. L., Newman, R. D., Fish, E. N., & Vogel, S. N. (2006). Contribution of interferon-beta to the murine macrophage response to the toll-like receptor 4 agonist, lipopolysaccharide. *The Journal of Biological Chemistry*, 281(41), 31119–30. <https://doi.org/10.1074/jbc.M604958200>
- Thomson, J. A. (1998). Embryonic stem cell lines derived from human blastocysts. *Science*. <https://doi.org/10.1126/science.282.5391.1145>
- Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2012). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, 14(2), 178–192. <https://doi.org/bbs017> [pii]10.1093/bib/bbs017
- Tsai, Z. T., Shiu, S. H., & Tsai, H. K. (2015). Contribution of Sequence Motif, Chromatin State, and DNA Structure Features to Predictive Models of Transcription Factor Binding in Yeast. *PLoS Comput Biol*, 11(8), e1004418. <https://doi.org/10.1371/journal.pcbi.1004418>
- Turing, A. M. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem %U <http://www.cs.helsinki.fi/u/gionis/cc05/OnComputableNumbers.pdf>. *Proceedings of the London Mathematical Society*, 2 %Z Turin(42), 230–265.
- Voss, T. C., & Hager, G. L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics*. <https://doi.org/10.1038/nrg3623>
- Wong, E., Baur, B., Quader, S., & Huang, C. H. (2012). Biological network motif detection: Principles and practice. *Briefings in Bioinformatics*. <https://doi.org/10.1093/bib/bbr033>
- Young, R. A. (2011). Control of the embryonic stem cell state. *Cell*.

<https://doi.org/10.1016/j.cell.2011.01.032>

Zaret, K. S., & Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes Dev*, 25(21), 2227–2241. <https://doi.org/10.1101/gad.176826.111>

Zaret, K. S., & Mango, S. E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. *Current Opinion in Genetics & Development*. <https://doi.org/10.1016/j.gde.2015.12.003>

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., ... Shirley, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*. <https://doi.org/10.1186/gb-2008-9-9-r137>

Zhu, Y., Sun, L., Chen, Z., Whitaker, J. W., Wang, T., & Wang, W. (2013). Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res*, 41(22), 10032–10043. <https://doi.org/10.1093/nar/gkt826>



## **12. LIST OF KEYWORDS**

enhancer, epigenomics, transcriptional regulation, macrophages, embryonic stem cells, quantitative and qualitative modelling, supervised learning models

### **13. KULCSSZAVAK LISTÁJA**

enhanszer, epigenomika, transzkripció szabályozás, makrofágok, embrionális őssejtek,  
kvantitatív és kvalitatív modellezés, felügyelt tanulási modellek

## 14. ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, Prof. Dr. László Nagy, for his excellent mentorship and guidance throughout my Ph.D. training.

I gratefully acknowledge Prof. Dr. Fésüs László and Prof. Dr. József Tózsér the former and recent head of the Department of Biochemistry and Molecular Biology for the opportunity to work in a professional, internationally recognized work environment.

I would also like to thank my co-supervisor Dr. Benedek Nagy and my advisors Prof. Fuxreiter Mónika and Dr. Lóránt Székvölgyi for their expert advice and words of encouragement.

I am also very thankful to Dr. Bálint L. Bálint, László Steiner, Dr. Szilárd Póliska, Dr. Lajos Széles, Ixchelt Cuaranta-Monroy, Dr. Bence Dániel, Dr. Zsolt Czimmerer, and Dr. Zoltán Simándi, Dr. Emanuele Raineri, Prof. Dr. Ivo G. Gut and all my co-authors for their contributions to my PhD programme.

I would like to express my deepest gratitude to my beloved family; my mother, brother, my grandparents and Lilla Ozgyin for their love and infinite patience.

## 15. APPENDIX



UNIVERSITY of  
DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY  
UNIVERSITY OF DEBRECEN

H-4002 Egyetem tér 1, Debrecen

Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number:  
Subject

DEENK//2019.PL  
PhD Publikációs Lista

Candidate: Attila Horváth

Neptun ID: BC1RA5

Doctoral School: Doctoral School of Molecular Cellular and Immune Biology

### List of publications related to the dissertation

1. **Horváth, A.**, Dániel, B., Széles, L., Cuaranta-Monroy, I., Czimmerer, Z., Ozgyin, L., Steiner, L., Kiss, M., Simándi, Z., Pólska, S., Giannakis, N., Raineri, E., Gut, I. G., Nagy, B., Nagy, L.: Labelled regulatory elements are pervasive features of the macrophage genome and are dynamically utilized by classical and alternative polarization signals. *Nucleic Acids Res.* 47 (6), 2778-2792, 2019.  
DOI: <http://dx.doi.org/10.1093/nar/gkz118>  
IF: 11.561 (2017)
2. Simándi, Z., **Horváth, A.**, Wright, L. C., Cuaranta-Monroy, I., De, L. I., Károlyi, K., Sauer, S., Deleuze, J. F., Gudas, L. J., Cowley, S. M., Nagy, L.: OCT4 Acts as an Integrator of Pluripotency and Signal-Induced Differentiation. *Mol. Cell.* 63 (4), 647-661, 2016.  
DOI: <http://dx.doi.org/10.1016/j.molcel.2016.06.039>  
IF: 14.714

### List of other publications

3. Ozgyin, L., **Horváth, A.**, Hevessy, Z., Bálint, B. L.: Extensive epigenetic and transcriptomic variability between genetically identical human B-lymphoblastoid cells with implications in pharmacogenomics research. *Sci Rep.* 9, 1-16, 2019.  
DOI: <http://dx.doi.org/10.1038/s41598-019-40897-9>  
IF: 4.122 (2017)
4. Ivády, G., Madar, L., Dzsudzsák, E., Koczok, K., Kappelmayer, J., Krulisova, V., Macek, J. M., **Horváth, A.**, Balogh, I.: Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. *BMC Genomics.* 19, 1-8, 2018.  
DOI: <http://dx.doi.org/10.1186/s12864-018-4544-x>  
IF: 3.73 (2017)





5. Simándi, Z., Pájer, K., Károlyi, K., Sieler, T., Jiang, L. L., Kolostyák, Z., Sári, Z., Fekecs, Z., Pap, A., Patsalos, A., Contreras, G. A., Rehó, B., Papp, Z., Guo, X., **Horváth, A.**, Kiss, G., Keresztessy, Z., Vámosi, G., Hickman, J., Xu, H., Dormann, D., Hortobágyi, T., Antal, M., Nógrádi, A., Nagy, L.: Arginine Methyltransferase PRMT8 Provides Cellular Stress Tolerance in Aging Motoneurons.  
*J. Neurosci.* 38 (35), 7683-7700, 2018.  
DOI: <http://dx.doi.org/10.1523/JNEUROSCI.3389-17.2018>  
IF: 5.97 (2017)
6. Czimmerer, Z., **Horváth, A.**, Dániel, B., Nagy, G., Cuaranta-Monroy, I., Kiss, M., Kolostyák, Z., Pólska, S., Steiner, L., Giannakis, N., Varga, T., Nagy, L.: Dynamic transcriptional control of macrophage miRNA signature via inflammation responsive enhancers revealed using a combination of next generation sequencing-based approaches.  
*Biochim. Biophys. Acta. Gene Regul. Mech.* 1861 (1), 14-28, 2018.  
DOI: <http://dx.doi.org/10.1016/j.bbagr.2017.11.003>  
IF: 5.179 (2017)
7. Fejes, Z., Czimmerer, Z., Szűk, T., Pólska, S., **Horváth, A.**, Balogh, E., Jeney, V., Váradi, J., Fenyvesi, F., Balla, G., Édes, I., Balla, J., Kappelmayer, J., Nagy, B. J.: Endothelial cell activation is attenuated by everolimus via transcriptional and post-transcriptional regulatory mechanisms after drug-eluting coronary stenting.  
*PLoS One.* 13 (6), 1-20, 2018.  
DOI: <http://dx.doi.org/10.1371/journal.pone.0197890>  
IF: 2.766 (2017)
8. Czimmerer, Z., Nagy, Z. S., Nagy, G., **Horváth, A.**, Silye-Cseh, T., Kriston, Á., Jonás, D., Sauer, S., Steiner, L., Dániel, B., Deleuze, J. F., Nagy, L.: Extensive and functional overlap of the STAT6 and RXR cisomes in the active enhancer repertoire of human CD14+ monocyte derived differentiating macrophages.  
*Mol. Cell. Endocrinol.* 471, 63-74, 2018.  
DOI: <http://dx.doi.org/10.1016/j.mce.2017.07.034>  
IF: 3.563 (2017)
9. Ozgyin, L., **Horváth, A.**, Bálint, B. L.: Lyophilized human cells stored at room temperature preserve multiple RNA species at excellent quality for RNA sequencing.  
*Oncotarget.* 9 (59), 31312-31329, 2018.  
DOI: <http://dx.doi.org/10.18632/oncotarget.25764>
10. Simándi, Z., **Horváth, A.**, Cuaranta-Monroy, I., Sauer, S., Deleuze, J. F., Nagy, L., RXR heterodimers orchestrate transcriptional control of neurogenesis and cell fate specification.  
*Mol. Cell. Endocrinol.* 471, 51-62, 2018.  
DOI: <http://dx.doi.org/10.1016/j.mce.2017.07.033>  
IF: 3.563 (2017)





11. Dániel, B., Nagy, G., **Horváth, A.**, Czimmerer, Z., Cuaranta-Monroy, I., Póliska, S., Hays, T. T., Sauer, S., Francois-Deleuze, J., Nagy, L.: The IL-4/STAT6/PPAR[gamma] signaling axis is driving the expansion of the RXR heterodimer cistrome, providing complex ligand responsiveness in macrophages.  
*Nucleic Acids Res.* 46 (9), 4425-4439, 2018.  
DOI: <http://dx.doi.org/10.1093/nar/gky157>  
IF: 11.561 (2017)
12. Dániel, B., Nagy, G., Czimmerer, Z., **Horváth, A.**, Hammers, D. W., Cuaranta-Monroy, I., Póliska, S., Tzerpos, P., Kolostyák, Z., Hays, T. T., Patsalos, A., Houtman, R., Sauer, S., Francois-Deleuze, J., Rastinejad, F., Bálint, B. L., Sweeney, H. L., Nagy, L.: The Nuclear Receptor PPAR[gamma] Controls Progressive Macrophage Polarization as a Ligand-Insensitive Epigenomic Ratchet of Transcriptional Memory.  
*Immunity.* 49 (4), 615-626, 2018.  
DOI: <http://dx.doi.org/10.1016/j.immuni.2018.09.005>  
IF: 19.734 (2017)
13. Czimmerer, Z., Dániel, B., **Horváth, A.**, Rückerl, D., Nagy, G., Kiss, M., Peloquin, M., Budai, M., Cuaranta-Monroy, I., Simándi, Z., Steiner, L., Nagy, B. J., Póliska, S., Bankó, C., Bacsó, Z., Schulman, I. G., Sauer, S., Deleuze, J. F., Allen, J. E., Benkő, S., Nagy, L.: The Transcription Factor STAT6 Mediates Direct Repression of Inflammatory Enhancers and Limits Activation of Alternatively Polarized Macrophages.  
*Immunity.* 48 (1), 75-90, 2018.  
DOI: <http://dx.doi.org/10.1016/j.immuni.2017.12.010>  
IF: 19.734 (2017)
14. Gál, A., Balicza, P., Weaver, D. A., Naghdi, S., Joseph, S. K., Várnai, P., Gyuris, T., **Horváth, A.**, Nagy, L., Seifert, E. L., Molnár, M. J., Hajnóczky, G.: MSTO1 is a cytoplasmic pro-mitochondrial fusion protein, whose mutation induces myopathy and ataxia in humans.  
*EMBO Mol Med.* 9 (7), 967-984, 2017.  
DOI: <http://dx.doi.org/10.15252/emmm.201607058>  
IF: 10.293
15. Imre, L., Simándi, Z., **Horváth, A.**, Fenyőfalvi, G., Ifj., N. P. P., Niaki, E. F., Hegedűs, É., Bacsó, Z., Weyemi, U., Mauser, R., Ausio, J., Jeltsch, A., Bonner, W., Nagy, L., Kimura, H., Szabó, G.: Nucleosome stability measured in situ by automated quantitative imaging.  
*Sci Rep.* 7 (1), 1-15, 2017.  
DOI: <http://dx.doi.org/10.1038/s41598-017-12608-9>  
IF: 4.122



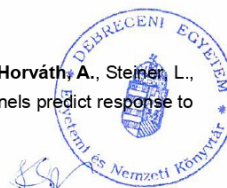


16. Kiss, M., Czimmerer, Z., Nagy, G., Bieniasz-Krzywiec, P., Ehling, M., Pap, A., Póliska, S., Botó, P., Tzerpos, P., **Horváth, A.**, Kolostyák, Z., Dániel, B., Szatmári, I., Mazzone, M., Nagy, L.: Retinoid X receptor suppresses a metastasis-promoting transcriptional program in myeloid cells via a ligand-insensitive mechanism.  
*Proc. Natl. Acad. Sci. U. S. A.* 114 (40), 10725-10730, 2017.  
DOI: <http://dx.doi.org/10.1073/pnas.1700785114>  
IF: 9.504
17. Varga, T., Mounier, R., **Horváth, A.**, Cuvellier, S., Dumont, F., Póliska, S., Ardjoune, H., Juban, G., Nagy, L., Chazaud, B.: Highly Dynamic Transcriptional Signature of Distinct Macrophage Subsets during Sterile Inflammation, Resolution, and Tissue Repair.  
*J. Immunol.* 196 (11), 4771-4782, 2016.  
DOI: <http://dx.doi.org/10.4049/jimmunol.1502490>  
IF: 4.856
18. Varga, T., Mounier, R., Patsalos, A., Gogolák, P., Peloquin, M., **Horváth, A.**, Pap, A., Dániel, B., Nagy, G., Pintye, É., Póliska, S., Cuvellier, S., Ben Larbi, S., Sansbury, B. E., Spite, M., Brown, C. W., Chazaud, B., Nagy, L.: Macrophage PPAR[gamma], a Lipid Activated Transcription Factor Controls the Growth Factor GDF3 and Skeletal Muscle Regeneration.  
*Immunity.* 45 (5), 1038-1051, 2016.  
DOI: <http://dx.doi.org/10.1016/j.immuni.2016.10.016>  
IF: 22.845
19. Simándi, Z., **Horváth, A.**, Nagy, P., Nagy, L.: Prediction and Validation of Gene Regulatory Elements Activated During Retinoic Acid Induced Embryonic Stem Cell Differentiation.  
*JoVE.* 2016 (112), e53978, 2016.  
DOI: <http://dx.doi.org/10.3791/53978>  
IF: 1.232
20. Czimmerer, Z., Varga, T., Kiss, M., Vázquez, C. O., Doan-Xuan, Q. M., Rückerl, D., Tattikota, S. G., Yan, X., Nagy, Z. S., Dániel, B., Póliska, S., **Horváth, A.**, Nagy, G., Varallyay, É., Poy, M. N., Allen, J. E., Bacsó, Z., Abreu-Goodger, C., Nagy, L.: The IL-4/STAT6 signaling axis establishes a conserved microRNA signature in human and mouse macrophages regulating cell survival via miR-342-3p.  
*Genome Med.* 8 (1), 1-22, 2016.  
DOI: <http://dx.doi.org/10.1186/s13073-016-0315-y>  
IF: 7.071
21. Rácz, R., Bereczki, J., Sramkó, G., Kosztolányi, A., Tóth, J. P., Póliska, S., **Horváth, A.**, **Barta, E.**, Barta, Z.: Isolation and Characterisation of 15 Microsatellite Loci from *Lethrus apterus* (Coleoptera: Geotrupidae).  
*Ann. Zool. Fenn.* 52 (1-2), 45-50, 2015.  
DOI: <http://dx.doi.org/10.5735/086.052.0204>  
IF: 0.753





22. Simándi, Z., Czipa, E., **Horváth, A.**, Kőszeghy, Á., Bordás, C., Pólska, S., Juhász, I., Imre, L., Szabó, G., Dezső, B., Barta, E., Sauer, S., Károlyi, K., Kovács, I., Hutóczki, G., Bognár, L., Klekner, Á., Szűcs, P., Bálint, B. L., Nagy, L.: PRMT1 and PRMT8 regulate retinoic acid-dependent neuronal differentiation with implications to neuropathology. *Stem Cells*. 33 (3), 726-741, 2015.  
DOI: <http://dx.doi.org/10.1002/stem.1894>  
IF: 5.902
23. Cuaranta-Monroy, I., Simándi, Z., Kolostyák, Z., Doan-Xuan, Q. M., Pólska, S., **Horváth, A.**, Nagy, G., Bacsó, Z., Nagy, L.: Highly efficient differentiation of embryonic stem cells into adipocytes by ascorbic acid. *Stem Cell Res.* 13 (1), 88-97, 2014.  
DOI: <http://dx.doi.org/10.1016/j.scr.2014.04.015>  
IF: 3.693
24. Gyöngyösi, A., Dócs, O., Czimmerer, Z., Orosz, L., **Horváth, A.**, Török, O., Méhes, G., Nagy, L., Bálint, B. L.: Measuring expression levels of small regulatory RNA molecules from body fluids and formalin-fixed, paraffin-embedded samples. *Methods Mol. Biol.* 1182, 105-119, 2014.  
DOI: [http://dx.doi.org/10.1007/978-1-4939-1062-5\\_10](http://dx.doi.org/10.1007/978-1-4939-1062-5_10)
25. Dániel, B., Nagy, G., Hah, N., **Horváth, A.**, Czimmerer, Z., Pólska, S., Gyuris, T., Keirsse, J., Gysemans, C., Van Ginderachter, J. A., Bálint, B. L., Evans, R. M., Barta, E., Nagy, L.: The active enhancer network operated by liganded RXR supports angiogenic activity in macrophages. *Genes Dev.* 28 (14), 1562-1577, 2014.  
DOI: <http://dx.doi.org/10.1101/gad.242685.114>  
IF: 10.798
26. Czimmerer, Z., Hulvely, J., Simándi, Z., Varallyay, É., Havelda, Z., Szabó, E., Varga, A., Dezső, B., Balogh, M., **Horváth, A.**, Domokos, B., Török, Z., Nagy, L., Bálint, B. L.: A versatile method to design stem-loop primer-based quantitative PCR assays for detecting small regulatory RNA molecules. *PLoS One*. 8 (1), 1-10, 2013.  
DOI: <http://dx.doi.org/10.1371/journal.pone.0055168>  
IF: 3.534
27. Meskó, B., Pólska, S., Váncsa, A., Szekanecz, Z., Palatka, K., Holló, Z., **Horváth, A.**, Steiner, L., Zahuczky, G., Podani, J., Nagy, L.: Peripheral blood derived gene panels predict response to infliximab in rheumatoid arthritis and Crohn's disease. *Genome Med.* 5 (6), 59-69, 2013.  
DOI: <http://dx.doi.org/10.1186/gm463>  
IF: 4.942







**UNIVERSITY of  
DEBRECEN**

**UNIVERSITY AND NATIONAL LIBRARY  
UNIVERSITY OF DEBRECEN**

H-4002 Egyetem tér 1, Debrecen  
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

28. Laczik, M., Tukacs, E., Uzonyi, B., Domokos, B., Doma, Z., Kiss, M., **Horváth, A.**, Batta, Z.,  
Maros-Szabó, Z., Török, Z.: Geno viewer, a SAM/BAM viewer tool.  
*Bioinformation*. 8 (2), 107-109, 2012.

**Total IF of journals (all publications): 195,742**

**Total IF of journals (publications related to the dissertation): 26,275**

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

03 May, 2019

