THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

Examination of the transcription factors acting in bone marrow derived macrophages

by Gergely Nagy

Supervisor: Dr. Endre Barta

UNIVERSITY OF DEBRECEN
DOCTORAL SCHOOL OF MOLECULAR CELL AND IMMUNE BIOLOGY

DEBRECEN, 2016

Table of contents

1. Introduction

1.1. Transcriptional regulation

1.1.1. Transcriptional initiation

Transcriptional initiation, elongation and termination are main steps of nascent RNA synthesis. As once RNA polymerase has been launched, it runs along the gene, of these steps – in our present knowledge – initiation may provide the most extended regulatory possibilities. In eukaryotes, the pre-initiation complex (PIC) assembled on the transcription factor binding sites (TFBSs) of the promoter of protein coding genes is composed of transcription factors (TFs), their co-regulators, as well as RNA polymerase II (Pol II) (**Figure 1**) (Green, Trends Biochem Sci., 2000; Grünberg and Hahn, Trends Biochem Sci., 2013). Pol II also is a complex built up from RNA polymerase II B (RPB1-12) subunits coded by Polr2a-k genes (Ruprich-Robert and Thuriaux, Nucleic Acids Res., 2010). To launch Pol II, not only the presence of a large number of regulatory proteins (Tsai and Nussinov, Biochem J., 2011), but also their enzymatic activity is needed: beside the multiple serine phosphorylation on the carboxy-terminal domain (CTD) of RPB1, helicase and topoisomerase activities are also indispensable to start PolII (Heidemann et al., Biochim Biophys Acta., 2013; Baranello et al., Transcription, 2013).

Once PIC has been formed, PolII synthesizes truncated and elongated transcripts in an abortive and productive manner, respectively (Dvir, Biochim Biophys Acta., 2002; Mandal et al., PNAS, 2004; Core et al., Science, 2008; Gaertner and Zeitlinger, Development, 2014). This former property may be needed for further regulatory functions, e.g. the operation of short transcripts in some comparable ways as happens in prokaryotes. In *E. coli*, the abortive transcript of bacteriophage T7 gene 10 has an antiterminator effect thus facilitating the expression of the further promoterless genes of the polycistron located directly downstream (Lee et al., Nucleic Acids Res., 2010). In eukaryotes, the ratio of the truncated and elongated

transcripts correlates with gene expression (Core et al., Science, 2008), but the possible roles of the short products still remain mostly unknown (Seila et al., Cell Cycle, 2009).

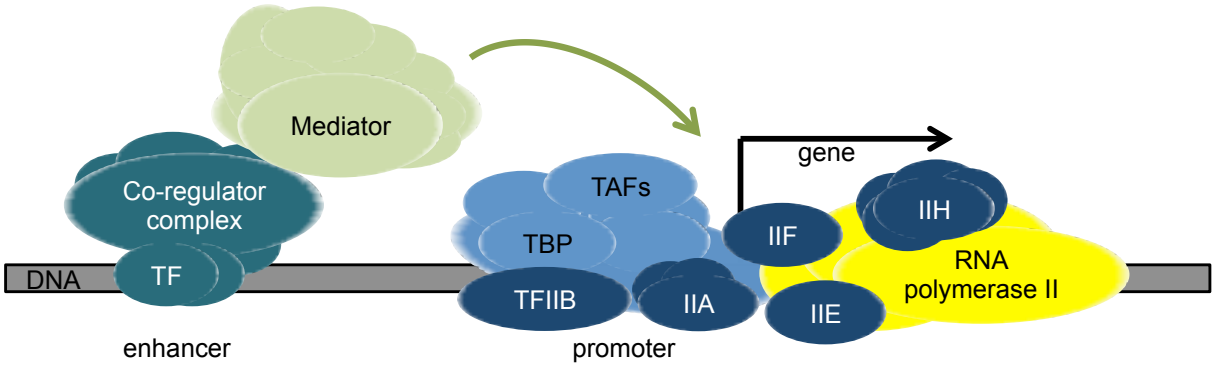


Figure 1. The model of the main components of PIC and the associating complexes driven by an enhancer

For transcription and DNA binding by most TFs, DNA has to be unwrapped from its multiple compacted state. DNA has a double helical structure wound on the nucleosome core complex, and nucleosomes can be further structured in fibers, and eventually chromatids. Nucleosome core is built up from the octamer of two of the core histone proteins H2A, H2B, H3 and H4. Histone H1/H5 is involved in chromatin compaction by holding nucleosomes together (Luger et al., Nature, 1997). As it is true for other proteins, histones also may carry several, notably close to 200 posttranslational modifications (**Table 1**) (Tan et al., Cell, 2011), among which acetylation (Allfrey et al., Proc Natl Acad Sci U S A., 1964) and methylation (Murray, Biochemistry, 1964) are the most studied and probably the most important ones (Kouzarides, Cell, 2007; Hildmann et al., Appl Microbiol Biotechnol., 2007). There are numerous enzymes conducting – creating and eliminating – these modifications, which eventually build up the epigenetic patterns of the different cell types.

Histone modification	Reference
Acetylation	Allfrey et al., Proc Natl Acad Sci U S A., 1964
Methylation	Murray, Biochemistry, 1964
ADP ribosylation	Nishizuka et al., J Biol Chem., 1968
Carbonylation	Wondrak et al., Biochem J., 2000
Biotinylation	Stanley et al., Eur J Biochem., 2001
Deimination / citrullination	Hagiwara et al., Biochem Biophys Res Commun., 2002; Wang et al., Science, 2004
SUMOylation	Shiio and Eisenman, Proc Natl Acad Sci U S A., 2003
Ubiquitination	Zhang, Genes Dev., 2003
Phosphorylation	Nowak and Corces, Trends Genet., 2004
Proline isomerization	Nelson et al., Cell, 2006
Propionylation / butyrylation	Chen et al., Mol Cell Proteomics, 2007
Crotonylation	Tan et al., Cell, 2011
O-palmitoylation	Zou et al., J Biol Chem., 2011

Table 1. The list of the main histone modifications

If it is accessible, TFs recognize and bind specific DNA elements (motifs) in the core promoter and in farther regions called enhancers (or silencers) (**Figure 1**) (e.g. Bohmann et al., Science, 1987). Promoters assign the beginning of the genes and are able to collect the minimal complex to promote gene expression, and enhancers can further augment this transcriptional activity. Proximal enhancers could be found before by enhancer trap experiments using e.g. 1 to 10 kb regions upstream to the transcription start site (TSS) (e.g. Fujisawa et al., J. Biochem, 2000; Szántó et al., Mol Cell Biol., 2004; Loudig et al., Biochem J., 2005), but distal enhancers seem that can be anywhere in the genome – even on other chromosomes –, thus these are still hard to be found (Banerji et al., Cell, 1981). The operation of these cis- and trans-regulatory elements can be explained by the looping of the DNA chain (Razin et al., FEBS Lett., 2013; Levine et al., Cell, 2014); during which the binding sites of the promoter and enhancer(s) – counting linearly in base pairs – can be very far from each other, but are getting close in the 3-dimension chromatin structure.

Insulator elements occupied by CCCTC-binding factor (CTCF) and Cohesin have been shown to be the focal points of chromatin interactions (Merkenschlager and Odom, Cell,

2013; Sofueva et al., EMBO J., 2013). Cohesin has been described as a ring-like complex holding sister chromatids together during the metaphase of cell division (Gruber et al., Cell, 2003) and fixing chromatin loops needed for transcriptional regulation during interphase (Hadjur et al., Nature, 2009; Merkenschlager and Odom, Cell, 2013). In this, SMC1 and SMC3 proteins form a long, open ellipse, which is closed by RAD21. The complex is so huge that can include two chromatid fibers, which are 30 nm in diameter (Gruber et al., Cell, 2003); but it is yet unknown how it fixes the ends of the chromatin loops and how many DNA double helices can be included in the relaxed structure. For loop formation the number of encompassed double helices is minimum two: the distribution and frequency of insulator binding have been found not to be random, but these anchor points were rather definitely paired (Sofueva et al., EMBO J., 2013).

1.1.2. Co-regulators and histone modifications

Co-factors, or more correctly co-regulators (rather co-activators or rather co-repressors), which interact with TFs but unable to bind DNA directly, join to the associated complex and modulate its structure, regulate the modifications and the release of nucleosomes and ultimately for the gene expression (**Figure 1**) (Xu et al., Curr Opin Genet Dev., 1999). Co-activators such as (E)P300/CBP proteins may possess intrinsic histone acetyltransferase (HAT) activity, while co-repressors typically have histone deacetylase function (HDAC superfamily). The acetylation of histones H3 and H4 at multiple lysine residues (H3K9 and 27, and H4K5, 8, 12 and 16) is generally specific for active enhancers (and even promoters) (Kouzarides, Cell, 2007). Histone methyltransferases (HMTs) (Trievel et al., Cell, 2002; Nguyen and Zhang, Genes Dev., 2011; Kuhn and Xu, Prog Mol Biol Transl Sci., 2009) and histone demethylases (HDMs) (Natoli et al., Curr Opin Drug Discov Devel., 2009) are the other better-known groups of histone modifying co-regulators (Allis et al., Cell, 2007).

Histone methylation – beside that of the arginine residues (Kuhn and Xu, Prog Mol Biol Transl Sci., 2009) – is also specific for lysine residues. As methylation concerns to such lysines, which can be also acetylated, these modifications are usually mutually exclusive. Unsurprisingly, the effect of acetylation and methylation of these residues is opposing: e.g. H3K9ac and H3K27ac are active, while H3K9me and H3K27me are repressive histone marks. Although H3K4 mono- and dimethylation (H3K4me1-2) are global enhancer (and promoter) marks and H3K4me3 is specific for the 5' end of the expressed genes (Kouzarides, Cell, 2007), H3K4ac is also registered as an active histone mark, however its ratio is usually less than the ratio of the methylated sites (Guillemette et al., PLoS Genet., 2011).

(E)P300 and CBP has been described as binding partners of the 300 kD adenovirus early-region 1A (E1A) protein (Whyte et al., Cell, 1989; Eckner et al., Genes Dev., 1994) and the cAMP-response element (CRE) binding (CREB) protein (Chrivia et al., Nature, 1993), respectively, but by now, it became clear that these proteins are closely related – thus have been classed in the lysine acetyltransferase 3 (KAT3) family – bearing the same domains and similar and overlapping functions. Both proteins carry a bromodomain (BRD) (Haynes et al., Nucleic Acids Res., 1992), which is specific for protein-protein interactions through acetyl-lysine binding. In human, there are 44 further members of the BRD superfamily, which predominantly act in the chromatin by binding acetylated histone residues (Filippakopoulos et al., Cell, 2012). Both P300 and CBP have HAT domain, thus can catalyze the lysine acetylation of several proteins e.g. those of the H3 and H4 histones. There is high overlap between their specificity, but their affinity to the different residues is not identical (Henry et al., Biochemistry, 2013). As co-regulators, primarily co-activators, they have several further interaction domains: C- and N-terminal "transcription adaptor putative zinc finger" (TAZ) domains, further zinc-fingers – such as a special plant homeodomain (PHD) (Park et al., FEBS Lett., 2013) and a ZZ-type domain –, "kinase-inducible domain interacting" (KIX)

domain, and the nuclear receptor co-activator interlocking (NRC) or CREB binding domain (Hay et al., J Am Chem Soc., 2014; Filippakopoulos et al., Cell, 2012). These domains let KAT3 proteins to interact with at least 400 proteins e.g. several TFs, thus became two of the main active enhancer marks. There are 18 further HAT proteins with different catalytic domains from different protein families, so there is an even higher redundancy between these chromatin modifier enzymes (Bedford et al., Epigenetics, 2010).

Enzymes that reduce and thus contribute to keep balance in the level of chromatin acetylation typically associate with co-repressor complexes. There are 19 known HDAC proteins grouped in four classes (Hildmann et al., Appl Microbiol Biotechnol., 2007). Class I includes HDAC1-3 and 8 proteins, class IIa includes HDAC4-5, 7 and 9 proteins, class IIb consists of HDAC6 and 10, class III collects sirtuins (SIRT1-7), and HDAC11 is the only, less known class IV protein. The activity of the mammalian homologues of yeast silent information regulator (SIR) family is NAD^+-dependent (Finnin et al., Nat Struct Biol., 2001), while the activity of the further HDAC proteins requires Zn^{2+} ion (Vannini et al., Proc Natl Acad Sci U S A., 2004). Class I HDAC proteins – except for HDAC8 which is functional as a single protein (Hu et al., J Biol Chem., 2000) – form the catalytic parts of different complexes: HDAC1 and 2 can join to the yeast switch independent 3 analogue A (SIN3A) (Hassig et al., Cell, 1997), nucleosome remodeling deacetylase (NURD) (Xue et al., Mol Cell, 1998), RE1-silencing transcription factor (REST) co-repressor (CoREST) (You et al., Proc Natl Acad Sci U S A., 2001) and MIDEAS complexes (Itoh et al., Nucleic Acids Res., 2015); while HDAC3 interacts with the deacetylase activation domain (DAD) of the nuclear receptor co-repressor (NCoR) and "silencing mediator for retinoid or thyroid-hormone receptors" (SMRT) complexes (Zhang et al., Mol Cell, 2002) through an inositol tetraphosphate molecule (Watson et al., Nature, 2012). Class IIa HDACs show low enzymatic activity but can interact with SMRT/NCoR proteins and by these or by their other partners recognized by

the long N-terminal tail, they function also as co-repressors (Hudson et al., J Biol Chem., 2015).

The multisubunit Mediator, composed of 25-30 distinct proteins (together having 1 MDa molecular weight), as co-regulator is indispensable part of PIC (**Figure 1**) (Myers and Kornberg, Annu Rev Biochem., 2000). None of the 31 mediator proteins have known enzymatic activity but have interaction surfaces with TFs, co-factors and the Pol II complex thus have communicational roles between these proteins (Tsai et al., Cell, 2014). Mediator binds also "super-enhancers" – outstandingly active enhancer domains having major effects on the determination of cell identity (Whyte et al., Cell, 2013). It has been also shown that "regulatory hotspots", where tens of TFs bind together to a relatively short DNA region (directly or indirectly), are in tight connection with Mediator (more precisely with MED1) (Siersbaek et al., Cell Rep., 2014). These recent findings also proved that transcriptional initiation depends on the co-operation of close to hundred or even more proteins. In the knowledge that it is not easy to find two genes with the same regulatory background, we can see how complex these regulatory networks are.

1.2. Promoter and enhancer sequences guiding transcription factors

1.2.1. General transcription factors

TFs that transduce signals into the nucleus, inside are directed also by specific DNA patterns (TFBSs) usually of the non-coding regions. If such a region is directly upstream to a TSS then can be called promoter. But promoters cannot be described just by their genomic location; they show specific motifs. TATA-box, with TATAAA consensus sequence – called Goldberg-Hogness box –, was the first regulatory element that has been described in the eukaryotic promoter (Lifton et al., Cold Spring Harb Symp Quant Biol., 1978). It is bound by TATA-binding protein (TBP), which is part of the TFIID complex also composed of TATA

associated factor (TAF1-13) proteins (**Figure 1**). TFIID is bound directly to the TFIIA trimer and the TFIIB protein (Thomas and Chiang, Crit Rev Biochem Mol Biol., 2006; Deng and Roberts, Chromosoma, 2007), which latter can occupy B recognition elements (BREs) (Lagrange et al., Genes Dev., 1998). During PIC assembling, further Pol II binding units join to promote gene expression: TFIIE and F proteins and the TFIIH complex consisted of 10 proteins. TFIIH has a core complex possessing helicase and ATPase activity and a cyclin-dependent kinase (CDK)-activating kinase (CAK) subcomplex (built up from 3 protein units) (Levine and Tjian, Nature, 2003; Thomas and Chiang, Crit Rev Biochem Mol Biol., 2006). Of these "general transcription factors", TBP and TFIIB can be considered as "real" TFs with specific binding sites, but some of the TAFs have also been shown to have sequence preference.

The TBP family in mammals consists of three members having similar DNA-binding domain (DBD). This is an about 180 amino acid long, quasi-symmetric core domain showing a saddle-shaped structure with a convex and a concave surface, which latter fits into the minor groove of DNA (Nikolov et al., Proc Natl Acad Sci U S A., 1996). While TBP has a general role in PIC assembly, TBP-like factors 1 and 2 (TBPL1-2) are specific for sperm and oocyte development, respectively (Akhtar and Veenstra, Cell Biosci., 2011). TBP and TBPL2 proteins have similar N-terminal domains, which are missing from TBPL1. TBPL1 cannot bind TATA-box but can likewise interact with TFIIA and B.

TFIIB is a unique TF coded by the Gtf2b gene. It contains a C-terminal quasi-duplicated domain containing 2x5 alpha helices. Helices 4-5 of the C-terminal part form a helix-turn-helix (HTH) domain recognizing the upstream BRE (SSRCGCC) in the major groove. The downstream BRE (RTDKKKK) is bound in the minor groove by helices 2-4 of the N-terminal part. Between the two halves of this domain there is a further, positively charged alpha helix, which is responsible for the interaction with TBP; and the N-terminal

helix 5 binds TAF9. The most highly conserved N-terminal regions, the Zn-ribbon and the B-finger (built up from three beta-sheets) that are stabilized by a central Zn^{2+} ion are responsible for the recruitment and binding of Pol II (Deng and Roberts, Chromosoma, 2007).

Most TAFs (TAF3-4, 6 and 8-13) contain histone fold motif (HFM) protein interaction domains built up from three short alpha helices connected with random coil loops. TAF4/4B, TAF12, TAF9 and TAF6 show sequence similarity to core histones H2A, H2B, H3 and H4, respectively. These TAFs bind DNA by different domains but together – similarly to the nucleosome proteins – have higher affinity to the core promoter by forming an octamer of the TAF6/9 heterotetramer and the TAF4B/TAF12 heterodimers (Shao et al., Mol Cell Biol., 2005). TAF1 is the largest component of TFIID having several domains. It has kinase domains at both ends responsible for e.g. histone H2B phosphorylation (Maile et al., Science, 2004). The middle DUF3591 domain bears HAT activity, but it is unknown which subdomain(s) is/are responsible for this activity, whether its triple barrel, the DNA-binding winged helix and/or the alpha helical subdomain. The C-terminal part is the RAPiD domain, which interacts with the CTD of TAF7. TAF7 has likewise a triple barrel domain in the N-terminus and a coiled region in the middle, which both perfectly fit into the cavity of TAF1 protein. It is clear now that the triple barrels complement each other and this interaction can be the basis of the repression of HAT activity by TAF7 (Wang et al., Cell Res., 2014). TAF1 has two C-terminal tandem BRDs also, which recognize and bind the acetylated H3K14 residue but prefer the acetylated K5/K12 and K8/K16 pairs of histone H4 (Jacobson et al., Science, 2000; Thomas and Chiang, Crit Rev Biochem Mol Biol., 2006).

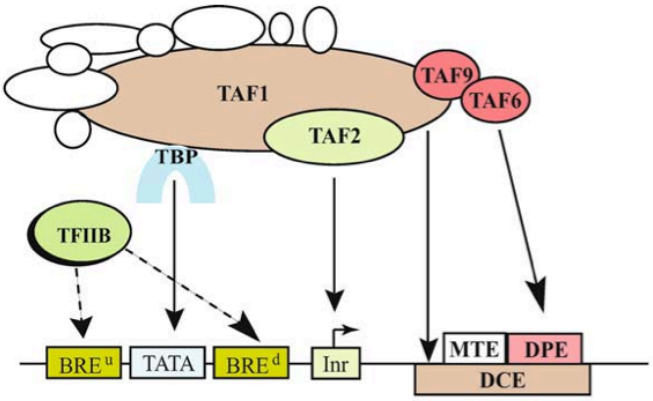


Figure 2. The response elements specific for TFIIB and D proteins (Thomas and Chiang, Crit Rev Biochem Mol Biol., 2006)

TATA-box is located upstream to the TSS, usually between positions -30 and -25. BREs can be both upstream and downstream relative to the TATA-box in positions -38 to -32 and -23 to -17, respectively (**Figure 2**). The next PolII positioning element "Initiator" (Inr) that has YYANWYY consensus sequence in position -2 to +5, had been shown to be occupied by TAF1/TAF2 proteins; however this "dimerization" has not been confirmed. The downstream core element (DCE) among positions +6 and +34 is similarly covered by TAF1. DCE is built up from three sub-elements (CTTC, CTGT and AGC) with 4 and 8 spacers between, respectively. Its alternative is the downstream promoter element (DPE) located between +28 and +34, which consensus sequence (RGWCGTG) is bound by the TAF6/TAF9 heterotetramer. DPE does not exclude the presence of "motif ten element" (MTE), which has been described in *Drosophila* in position +18 to +29 (Thomas and Chiang, Crit Rev Biochem Mol Biol., 2006).

Before the next-generation sequencing (NGS) era, it has been described that most of the promoters (80-90%) are TATA-less, half of them contain Inr, and half have DPE or BRE, covering together close to 100% of the TSSs (e.g. Gershenzon and Ioshikhes, Bioinformatics, 2005). These core TFBSs are bound by TFIIB and D proteins and thus initiate PIC assembling on the TSS. Nowadays, it seems that these aforementioned elements are even less abundant than have been thought to be. NGS made possible to determine every TSS, which

can be re-screened together. DeepCAGE (Valen et al., Genome Res., 2009) and massive RACE (Tsuchihara et al., Nucleic Acids Res., 2009) are the genome-wide versions of "cap analysis of gene expression" (CAGE) (Shiraki et al., PNAS, 2003) and "rapid amplification of 5' cDNA ends" (5' RACE) (Ambion, 2008), respectively. These methods showed that two kinds of transcription initiation sites exist: narrow ones with less than 10 bp width (TSSs) and broader ones with 25-250 bp widths. These latter can be called transcription start regions or TSRs. TSSs adding up to 22% tend to carry the classical promoter elements, while TSRs have rather CpG island motifs (Balwierz et al., Genome Biology, 2009). CpG islands, which were known to be relatively rare methylation sites, had been thought first not to be enriched in promoters (Gardiner-Garden and Frommer, J Mol. Biol., 1987). Then, their GC-box elements were shown in more and more promoters including some of indispensable genes such as the one of beta-globin (HBB) (Giglioni, Biochem Biophys Res Commun., 1989) and cytochrome P450, family 1, subfamily A, polypeptide 1 (CYP1A1) (Imataka et al., EMBO J., 1992).

The first identified CpG-binding TF was specificity protein 1 (SP1) (Dynan and Tjian, Cell, 1983a; Dynan and Tjian, Cell, 1983b; Kadonaga et al., Cell 1987), which is the founding member of the SP1-like family including SP1-9 proteins. The structure and function of SP1 and 4 are very similar: both contain all SP1 family-specific domains and both were shown to be activators. In the N-terminal part, SP3 lacks the serine/threonine-rich region and SP2 lacks the glutamine-rich region too. These both can act as repressors e.g. by competing with the activators (SP1/4) or by recruiting co-repressors. SP1-3 are ubiquitous proteins, while the others are developmental state and cell type specific ones (Zhao and Meng, Develop. Growth Differ., 2005). SP5-9 proteins have much shorter N-terminal part, but their C-terminal "Buttonhead" and DNA-binding domains are very similar to those of the longer family members. The three C2H2-type zinc-fingers (DBD) can bind both GC- (GGGGCGGGG) and GT-boxes (GGTGTGGGG). SP1-like family is closely related to the Krüppel-like factor

(KLF) family, which include further 17 TFs (KLF1-17) having the same type of DBD (Vliet et al., Genomics, 2006). KLFs are likewise able to bind GC-box, such as the Yamanaka-factor KLF4 (Takahashi and Yamanaka, Cell, 2006), which is able to repress the expression of the histidine decarboxylase (HDC) gene by competing with SP1 on the promoter of HDC (Ai et al., J Biol Chem., 2004).

During searching for further promoter specific motif enrichments, similar motifs appeared with different approaches. Chromatin immunoprecipitation (ChIP) coupled with NGS (ChIP-seq) is a method to fish out and sequence all DNA fragments that are bound by and thus possible to be cross-linked with a given protein that can be caught by a specific antibody (Robertson et al., Nat Methods, 2007). Since 2007, several other methods – also that were targeting regulatory sites – were transposed to NGS platforms instead of single molecule/sequence detection or microarray. Applying the new methods, beside GC-box, familiar and less familiar motifs were determined. From the ChIP-seq derived data of B-cell and macrophage promoters, the following motifs showed enrichment: an ETS binding site (EBS), the motif of NRF1 and GFY, the CCAAT-box (bound by NFY) and the CRE (Heinz et al., Mol. Cell, 2010). In a previous screen of promoters, beside these, the motif of MYC, AP-1 and YY1 appeared (Xie et al., Nature, 2005), and in a more recent paper, a further, HTH domain protein bound element has been published based on the combination of "formaldehyde-assisted isolation of regulatory elements" (FAIRE-) (Gaulton et al., Nat. Genet., 2010) and ChIP-seq data (Hong et al., Genomics Inform., 2012). Most of these TFs, which elements were detected in promoters, first time had been identified as viral oncoproteins forming diverse superfamilies that have indispensable functions.

1.2.2. The ETS superfamily

The first “erythroblast transformation-specific” (ETS or E26) protein has been identified as a fusion oncogene of the E26 retrovirus leading to leukemia in chicken (Leprince et al., Nature, 1983). There are 27/28 members of the superfamily in mouse and human, respectively, which are separated to 12 phylogenetic groups having a highly conserved DBD (**Table 2**). This ETS domain of about 85 amino acid residues (a winged HTH) binds to a purine-rich GGAA/T core motif, but the neighboring nucleotides also have some significance (Oikawa and Yamada, Gene, 2003; Gutierrez-Hartmann et al., Trends Endocrinol Metab., 2007). From both the classical RE cloning examinations (e.g. Landry et al., Blood, 2005; Zhu et al., Cell Res., 2006; Nowling et al., Mol Immunol., 2008; Okada et al., PLoS One, 2011; Oh et al., Biochim Biophys Acta., 2012) and the motif enrichments of ChIP experiments, e.g. the motif collection of HOMER (Heinz et al., Mol. Cell, 2010), it seems that there are four main types of EBSs: Most ETS proteins were shown to bind ccGGAAgt or caGGAAgt sequence, SPDEF binds mainly caGGATga sequence, while the SPI family tends to bind rather gaGGAAgt sequence; however the flanking nucleotides (shown in lower case) usually seem to modulate the protein affinity to the binding site.

Another approach was also carried out to determine ETS classes based on their DNA binding (Wei et al., EMBO J., 2010): with high-throughput microwell-based TF DNA-binding specificity assay and protein-binding microarrays, similar four classes could be divided. The families of class I with accGGAAgt consensus motif (PEA3, TCF, ETS, ER71, ERF, ERG and ELG), class II with cccGGAAgt consensus motif (TEL, ESE and ELF), class III with agaGGAAgt consensus motif (SPI) and class IV with ccGGAT consensus motif (SPDEF) were broadly validated by the previously mentioned methods. Although the caGGAAgt motif is missing from the results of this screen, it seems to be a functional binding site as Liu et al., by applying ChIP-seq, described that ETV2 (ER71) binds this motif, which

is specific for mainly enhancers in differentiating embryonic bodies (Liu et al., EMBO Rep., 2015). According to the HOMER motif matrix database, FLI1, ERG and ETS1 also bind this kind of EBS, but based on this study, these cannot replace the function of ETV2 in hemangioblast induction, thus it seems that the binding and function of ETS proteins is not as redundant as it was supposed before.

Family	Full name	Protein	Full name
SPI	Spleen focus forming virus (SFFV) proviral integration oncogene	SPI1 (PU.1)	Purine-rich nucleic acid binding protein 1
		SPIB	Spleen focus forming virus (SFFV) proviral integration oncogene B-C
		SPIC	
TEL	Translocation E26 transforming-specific leukaemia gene	ETV6	ETS variant 6-7
		ETV7	
ESE	Epithelium specific ETS	ELF3	ETS-like factor 3,5
		ELF5	
		EHF	ETS homologous factor
ELF	ETS-like factor	ELF1	ETS-like factor 1-2,4
		ELF2	
		ELF4	
SPDEF	SAM pointed domain containing ETS transcription factor	PDEF	(SAM) pointed domain containing ETS transcription factor
PEA3	Polyomavirus enhancer activator 3	ETV1	ETS variant 1,4-5
		ETV4	
		ETV5	
TCF	Ternary complex factor	ELK1	ETS-like gene 1,3-4
		ELK3	
		ELK4	
ETS	Erythroblast transformation-specific protein	ETS1	Erythroblast transformation-specific protein 1-2
		ETS2	
ER71	ETS-related protein 71	ETV2	ETS variant 2
ERF	ETS2 repressor factor	ERF	ETS2 repressor factor
		ETV3	ETS variant 3
		ETV3L	ETS variant 3-like
ERG	ETS related gene	ERG	ETS-related gene
		FEV	Fifth Ewing variant
		FLI1	Friend leukemia integration 1
ELG	ETS-like gene	GABPA	GA repeat binding protein alpha

Table 2. ETS protein families

In promoters, the first two kinds of ETS motif (rather ccGGAAgt) show enrichment in both classifications (Xie et al., Nature, 2005, Heinz et al., Mol. Cell, 2010, Hong et al., Genomics Inform., 2012), which, based on ChIP-seq results, are generally bound by the ubiquitous GABP dimer, ETS2, ETV6 and TCF family proteins (Oikawa and Yamada, Gene, 2003), and can be bound probably cell type specifically by the members of the PEA3 and ELF families. GABPA is a special ETS protein as it heterodimerizes with a non-ETS co-factor (GABPB1) that possess the trans-activation activity (Thompson et al., Science, 1991). It is the only member of the ELG family, but there are three other families (ER71, TEL and PDEF), which are represented by one protein in mouse (ETV2, ETV6 and SPDEF, respectively). From mouse, ETV7 (TEL) is missing: this gives the difference between the ETS gene number compared to human. Next to the ETS domain, GABPA has two C-terminal helices, which are responsible for heterodimerization. These helices are similarly present in ETV6 and the ETS family proteins but have autoinhibitory roles in these proteins (Hollenhorst et al., Annu. Rev. Biochem., 2011). “Sterile alpha motif” pointed (SAM/PNT) domain is an about 80 amino acid long helix-loop-helix (HLH) interaction domain connecting to non-HLH domains. Beside “SAM pointed domain containing ETS transcription factor” (SPDEF), TEL, ESE, ELG, ERG (except for FEV) and the ETS families bear this HLH domain, which have divergent functions (Hollenhorst et al., Annu. Rev. Biochem., 2011). ELF, ERF, PEA3, ER71, SPI and TCF families do not have SAM domain but have other activator/repressor domains (Oikawa and Yamada, Gene, 2003; Gutierrez-Hartmann et al., Trends Endocrinol Metab., 2007). ETS2 repressor factor (ERF), TEL and TCF families have different N- or C-terminal repressor domains and they are really known to have repressor functions (e.g. Tanaka et al., Cell Struct Funct., 2013).

1.2.3. The AP-1 and CREB proteins

Activator protein 1 (AP-1) has been described as a phorbol 12-O-tetradecanoate 13-acetate (TPA) inducible TF, which binds a promoter specific TPA response element (TRE) having a TGARTCA consensus motif (Comb et al., Nature, 1986; Lee et al., Cell, 1987). Shortly later, AP-1 proteins were identified as JUN (Bohmann et al., Science, 1987) and its interacting partners, FOS and FOS-related antigens (FRAs, called FOS-like proteins now) (Rauscher et al., Science. 1988; Curran and Franza, Cell, 1988). FOS oncogene was first described as a 55 kDa phosphoprotein (p55) of the Finkel-Biskis-Jinkins (FBJ) murine osteosarcoma virus (Curran and Teich, J Virol., 1982). JUN has got its name from the avian sarcoma virus 17 (ASV 17) protooncogene as the Japanese "ju-nana" means "17" (Maki et al., Proc Natl Acad Sci U S A., 1987). This time, another TF – the CCAAT-box and enhancer-binding protein (C/EBP) – that bears similar leucine-dominated domain as FOS, JUN and MYC, was assumed to form double helical protein dimers called leucine zipper (Landschulz et al., Science, 1988). By now, probably all basic leucine zipper (bZIP) proteins are known in mammals, thus it became clear that FOS and JUN proteins form distinct families and form heterodimers with ATF/CREB, BATF and MAF family proteins (**Figure 3**) (Newman and Keating, Science, 2003). Thus based on the definition that the AP-1 complex includes FOS/JUN but not necessarily only these proteins, there are eight bZIP families, which members can form "AP-1" heterodimers (**Table 3**).

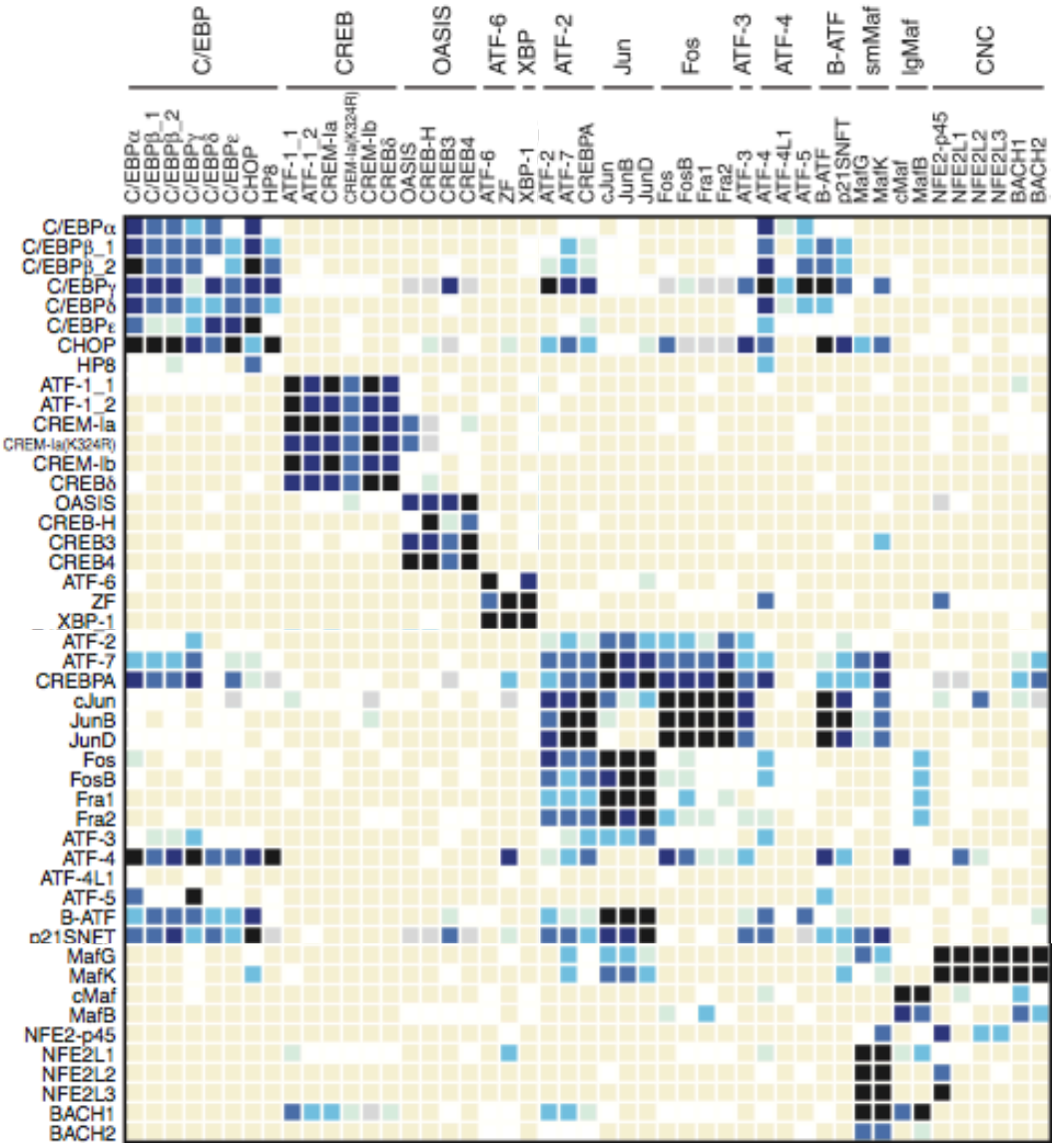


Figure 3. Consensus interaction matrix for 47 human bZIP peptides (Newman and Keating, Science, 2003)

Fluorescent proteins (probes) are listed at the top, and surface-phase proteins are listed at the left. Control peptides spotted in duplicate are indicated name_1, name_2. Peptide pairs were assigned a Z-score that corresponds to the highest value for which the probability of seeing the observed number of occurrences by chance is less than 10-4. Z > 20 (black); Z > 10 (dark blue); Z > 5 (medium blue); Z > 2.5 (light blue); Z > 1.5 (light green); at least 75% observations with Z > 1.0 (yellow); no assignment with confidence meeting P-value test (white); signal observed was not reciprocal (i.e., for the heterodimer XY, ZXY > 2.5, ZYX < 1) (gray).

Group	Full name	Protein	Full name
ATF2	Activating transcription factor 2	ATF2	Activating transcription factor 2, 7
		ATF7	
		CREB5	cAMP responsive element binding protein 5
ATF3	Activating transcription factor 3	ATF3	Activating transcription factor 3
		JDP2	JUN dimerization protein 2
ATF4	Activating transcription factor 4	ATF4	Activating transcription factor 4-5
		ATF5	
BATF	bZIP transcription factor, ATF-like	BATF	bZIP transcription factor, ATF-like, 2-3
		BATF2	
		BATF3	
FOS	FBJ (Finkel-Biskis-Jinkins) murine osteosarcoma viral oncogene homolog protein	FOS	FBJ murine osteosarcoma viral oncogene homolog protein, B
		FOSB	
		FOSL1	FOS-like 1-2
		FOSL2	
JUN	Avian sarcoma virus 17 (ASV 17) ju-nana (=17) protooncogene	JUN	JUN protooncogene, B, D
		JUNB	
		JUND	
lgMAF	Large musculoaponeurotic fibrosarcoma protein	MAF	Musculoaponeurotic fibrosarcoma protein, A, B
		MAFA	
		MAFB	
		NRL	Neural retina leucine zipper protein
smMAF	Small musculoaponeurotic fibrosarcoma protein	MAFF	Musculoaponeurotic fibrosarcoma protein F, G, K
		MAFG	
		MAFK	

Table 3. The (rather) AP-1 related bZIP protein families

Many of these heterodimers, typically those that includes ATF/CREB factors, can also be considered as CREB proteins as these usually bind the CRE recognition site with different affinities (**Table 4**) (e.g. Yamamoto et al., Nature, 1988). Similarity between AP-1 and CREB proteins and their binding sites soon became clear (e.g. Hai et al., Genes Dev., 1988; Hai and Curran, Proc. Nati. Acad. Sci. U S A., 1991). The first activating (or adenovirus) transcription factor (ATF) was shown to bind the TGACGTCA sequence (CRE) in the E3 promoter of adenovirus type 5 (Hurst and Jones, Genes Dev., 1987). Shortly later, it was shown that this complex contains an AP-1 protein that is fishable with anti-JUN antibody (Hai et al., Genes Dev., 1988). After the identification of CRE-BP1 (called ATF2 now) that has been found highly expressed in the brain (Maekawa et al., EMBO J., 1989), further CREB/ATF proteins

and their interaction affinities were examined and described (e.g. Hai and Curran, Proc. Nati. Acad. Sci. U S A., 1991; Newman and Keating, Science, 2003). Now, respectively 7+8 CREB/ATF proteins are known in mammals, which are classified into 6 families (listed in **Tables 3-4**).

Group	Full name	Protein	Full name
ATF6	Activating transcription factor 6	ATF6	Activating transcription factor 6, 6B
		ATF6B	
CNC	Cap'n'collar-type bZIP proteins	BACH1	BTB and CNC homology protein 1-2
		BACH2	
		NFE2	Nuclear factor erythroid 2
		NFE2L1	NFE2-like 1-3
		NFE2L2	
		NFE2L3	
CREB1	cAMP responsive element binding protein 1	CREB1	cAMP responsive element binding protein 1
		ATF1	Activating transcription factor 1
		CREM	CRE modulator
CREB3	cAMP responsive element binding protein 3	CREB3	cAMP responsive element binding protein 3
		CREB3L1	cAMP responsive element binding protein 3-like 1-4
		CREB3L2	
		CREB3L3	
		CREB3L4	
CREBL2	CRE-binding protein-like 2	CREBL2	CRE-binding protein-like 2
CREBRF	CREB3 regulatory factor	CREBRF	CREB3 regulatory factor
CREBZF	CREB/ATF bZIP transcription factor	CREBZF	CREB/ATF bZIP transcription factor
XBP1	X-box binding protein 1	XBP1	X-box binding protein 1

Table 4. The other, rather CREB related bZIP protein families

bZIP transcription factor, ATF-like (BATF) protein was described as an AP-1 protein of hematopoietic cells, which competes with FOS proteins for the dimerization with JUNs and thus for TRE binding (Dorsey et al., Oncogene, 1995; Echlin et al., Oncogene, 2000). It lacks any trans-activation domains (TADs) thus it really forms repressor heterodimers, which inhibit cellular growth. BATF3 that has been called first JUN dimerization protein 1 (JDP1)

was described together with the similarly JUN repressor JDP2 (Aronheim et al., Mol Cell Biol., 1997).

MAF had been identified as an oncogene of avian sarcoma 42 (AS42) transforming retrovirus isolated from a chicken musculoaponeurotic fibrosarcoma (Kawai et al., Virology, 1992), and its paralogues (MAFK, F, B, G and "neural retina leucine zipper", NRL proteins) were also identified soon (Kataoka et al., Mol Cell Biol., 1994). These cannot bind TRE/CRE but the extended form called MAF-recognition element (MARE) as homodimers or together with FOS/JUN proteins as the 3(-3) bp extension(s) does not affect the basic TGA half sites.

Nuclear factor erythroid 2 (NFE2) (Mignotte et al., Nucleic Acids Res., 1989), the ubiquitous NFE2-related factors (NRFs, called now nuclear factor erythroid 2-like, NFE2L proteins) (Chan et al., Proc. Natl. Acad. Sci. U S A., 1993) and the "BTB and CNC homology" (BACH) proteins (Oyake et al., Mol Cell Biol., 1996) were shown to bind MARE, as well. These are all members of the Cap'n'collar (CNC)-type bZIP protein family having the CNC domain. BACH genes have a further "broad complex-tramtrack-bric-a-brac" (BTB) domain that is responsible for the interaction with such proteins that have also this kind of domain. BACH1-2 proteins were both shown to heterodimerize with the MAFK, which, as a small MAF, lacks any TADs. As all CNC proteins have high affinity to small MAFs (Newman and Keating, Science, 2003) and bind TRE half sites (TGA) (Oyake et al., Mol Cell Biol., 1996), this suggests a similar role and probable competition with JUN proteins. NFE2L proteins are still called NRFs, but these are not related to the EWG paralogue nuclear respiratory factor 1 (NRF1) that will be discussed below.

All of the previously listed AP-1 and/or CREB (-related) proteins share a very similar DBD and dimerization mechanism. This highly conserved, usually C-terminal bZIP domain continues with the DNA-interacting region in the N-terminal direction. Leucine zipper contains generally 4-5 leucines separated by 6-6, mainly hydrophobic amino acids, and the

slightly double helical backbones contact through the hydrophobic Y-shaped residues. The N-terminal arms of this parallel coiled coil structure contact with the major groove of the DNA, and unsurprisingly, this dimer bind an inverted repeat element (Lee, J Cell Sci., 1992). AP-1 and CREB proteins share in a palindrome sequence built up from two TGA half sites with one and two spacers, respectively. These are the previously mentioned TRE and CRE, which latter has been shown several times to be a promoter proximal element (e.g. Xie et al., Nature, 2005; Heinz et al., Mol. Cell, 2010; Hong et al., Genomics Inform., 2012). There are "AP-1 proteins", which prefer the extended version of these elements: MAF proteins recognize the upstreamly extended TGCTGA half sites, and their inverted repeats with one spacer are the MAREs.

Together there are 40 better-known members of the AP-1 and CREB related protein group, sometimes without consistent nomenclature as ATF proteins often mingle with the CREB proteins (**Tables 3-4**): CREB5 together with ATF7 is part of the ATF2 family; ATF3 show high similarity to JDP2; ATF1 is a CREB1 family protein together with the CREB1 and CRE modulator (CREM) protein; and ATF6 proteins have similar structure and function as of the CREB3 proteins. There are four further, highly conserved, unique bZIP proteins in connection with the CREB group (**Table 4**): CREBL2, CREBRF, CREBZF and XBP1.

CREB-like 2 (CREBL2) is a special bZIP protein, which probably forms heterodimer with CREB1, driving together the preadipocyte – adipocyte transition in 3T3-L1 cells (Ma et al., Biochem. J., 2011). CREB3 regulatory factor (CREBRF) or Luman recruitment factor (LRF) has been described as a regulator of CREB3 (Luman) (Audas et al., Mol Cell Biol., 2008). CREB3-like and ATF6 proteins have a transmembrane domain anchoring them into the endoplasmic reticulum (ER). Under ER stress, intramembranous proteolysis activates these proteins that then can be translocated into the nucleus. In there, ATF6 induces the generation of a potent splice variant of the X-box binding protein 1 (XBP1) to expand the ER

stress response (Yoshida, Cell, 2001). CREBRF is a short half-life nuclear protein that seems to form heterodimers with the inflowing CREB3 leading to its accelerated proteasomal degradation (Audas et al., Mol Cell Biol., 2008). CREB/ATF bZIP transcription factor (CREBZF) also called "Zhangfei" (ZF) has been likewise shown to be a CREB3 repressor (Misra et al., J. Biol. Chem., 2005), but in the same time it augments ATF4 activity on CRE elements (Hogan et al., FEBS Letters, 2006).

Several studies have been examined the pairing of bZIP proteins and found different tendencies in their homo- and heterodimerization affinities (Newman and Keating, Science, 2003): ATF6, CREB, CREBZF, large MAF and XBP1 proteins have been found to have high affinity to form homodimers, while ATF2-4, BATF, CNC, JUN, FOS and small MAF proteins tend to rather form heterodimers. Unsurprisingly, JUN is the most "permissive" AP-1/CREB protein, as it can form homodimer and can interact with several other group members including ATF2, ATF3, BATF, FOS and small MAF proteins. FOS proteins have the highest affinity to JUNs, but they are able to bind ATF2 and ATF4, but also FOS and MAFB proteins with lower affinities. ATF3, ATF4 and CNC proteins seem to form only heterodimers based on these results. If we insist on the definition that AP-1 binds TRE, only FOS/JUN heterodimers will fit because although BATF/JUN binds TRE, it is rather a repressor dimer, MAF(B)/JUN needs a longer motif, and the dimers with ATF/CREB proteins bind CRE with much higher affinity. Ultimately there are tens of bZIP proteins, which can bind the promoter specific CREs, but these show cell type and activation specificity to make the regulatory system less redundant.

1.2.4. Other promoter specific transcription factor families

MYC had been identified as an oncogene of avian myelocytomatosis virus MC29 (Duesberg et al., Proc Natl Acad Sci U S A., 1977) and the cellular form was first isolated

from chicken embryos (Vennström et al., J Virol., 1982). In Burkitt's lymphoma, the translocated copies of MYC loose the regulatory control, show constitutively elevated gene expressional level and as TFs eventually reprogram the B cells (Croce and Nowell, Blood, 1985). There are two further proteins identified in the MYC family: MYCN has been described in human neuroblastoma cell lines (Kohl et al., Cell, 1983) and MYCL(1) has been described in human small cell lung cancer (Nau et al., Nature, 1985). MYC-associated factor X (MAX), which has a similar bHLH-ZIP domain as MYC proteins, has been described as heterodimerizing partner of all three MYC family members binding together the enhancer-box (E-box: CACGTG) (Blackwood and Eisenman, Science, 1991). Basic HLH-ZIP domain is the N-terminally extended version of the bZIP domain having similar dimerization and DNA binding features. MAX is smaller than MYC proteins: it is lacking any TADs but has a C-terminal nuclear localitation signal (NLS), which can be needed for the nuclear translocation of its dimeric form (Kato et al., Genes Dev., 1992). This homodimer was found to be a repressor by occupying E-boxes without activation capability (Blackwood and Eisenman, Science, 1991).

MAX dimerization protein 1 (MXD1 or MAD) (Ayer et al., Cell, 1993) and MAX interactor 1 (MXI1) (Zervos et al., Cell, 1993) were the founding members of a new bHLH-ZIP family called MAD (including also MXD3-4). MAD proteins are repressors, which dimerize with MAX thus supersede MYC from the heterodimer and the E-boxes. Similarly to the MAD proteins, MAX network transcriptional (MNT) repressor forms dimer with MAX and similarly has a SIN3 interaction domain (SID) (Hurlin et al., Genes Dev., 1997), which binds the SIN3 co-repressor (Harper et al., Proc Natl Acad Sci U S A., 1996). The next identified bHLH-ZIP protein, MAX-like protein X (MLX) has a similar function as of MAX, but its affinity is restricted to MXD1, MXD4 (Billin et al., J Biol Chem., 1999) and the MLX interacting proteins (MLXIP and MLXIP-like) (Billin et al., Mol Cell Biol., 2000). MLXIP

(MondoA) is a large, generally cytosolic protein having more cytoplasmic localization domains (CLDs), and in the nucleus it acts as an activator. “MAX gene associated” (MGA) protein is a special TF having two DBDs (Hurlin et al., EMBO J., 1999). In the N-terminus it has a T-domain, which dimeric form has been shown to bind a long palindrome sequence (aattTcaCACcTAgGTGtgAaatt) called Brachyury (T) binding site (Müller and Herrmann, Nature, 1997). Close to the C-terminus it has the bHLH-ZIP domain, which provides dimerization and E-box binding together with MAX. This heterodimerization turns MGA to a transcriptional activator on both T and E-boxes (Hurlin et al., EMBO J., 1999).

Nuclear respiratory factor 1 (NRF1) is a unique TF having paralogues in insects (*Drosophila* “erect wing”, EWG), sea urchin (P3A2) and chicken (IBR/F), as well (Evans and Scarpulla, J Biol Chem., 1989; Virbasius et al., Genes Dev., 1993). It forms homodimers and binds the palindrome RCGCRYGCGY consensus sequence by its N-terminal part, which contains the NLS and is responsible for (PPARGC1A/PPRC1) co-activator binding, too. It has a further conserved CTD responsible for activation (Fazio et al., J Biol Chem., 2001; Andersson and Scarpulla, Mol Cell Biol., 2001). NRF1 has been first described as the activator of cytochrome C and other members of the mitochondrial respiratory apparatus – this is where its name came from –, but it regulates several other genes with diverse functions. Recently, it seems that has a global role in transcriptional regulation on TATA-less promoters (e.g. Baar et al., FASEB J., 2003; Chang et al., J Biol Chem., 2004; Hossain et al., J Biol Chem., 2009).

It took some time to identify and differentiate the proteins that bind the CCAAT-box (e.g. Bucher, J. Mol. Biol., 1990) that had been recognized as a conserved upstream element (Efstratiadis et al., Cell, 1980). It is bound by NFY (Dorn et al., Proc. Nati. Acad. Sci. U S A., 1987), but C/EBP also has a similar half site (gcAA); it has a TTgcgcAA consensus sequence (Ryden and Beemon, Mol Cell Biol., 1989). Nuclear transcription factor Y (NFY) also termed

CCAAT-binding factor (CBF) is a heterotrimeric complex with a short consensus sequence (Kahle et al., Mol Cell Biol., 2005). CCAAT-box is more abundant than the TATA-box, but NFY has only a synergistic "enhancer" role by binding in the proximity of the promoters, between positions -50 and -100, usually at about position -80. It is composed of three uniquely conserved subunits: NFYB and C contain an HFM domain responsible for dimerization, and their heterodimer provides the surface for NFYA binding. The heterodimer is carried into the nucleus via importin 13, but NFYB and C do not bear independent NLS and due to their smaller size, as monomers can diffuse through the nuclear pores. NFYA has a non-classical, positively charged C-terminal NLS recognized by importin beta. This NLS overlaps with the DBD and the subunit interaction domain (Kahle et al., Mol Cell Biol., 2005). NFYB/C has been also shown to associate with other complexes such as with the H3/4 histone dimer and thus bind DNA independently of NFYA.

"Yin-yang 1" (YY1) has got its name from its dual nature as it has been described as a TF with both activator and repressor interaction domains (Shi et al., Cell, 1991). Its DBD contains four C2H2-type Krüppel-like zinc-finger motifs, which bind dominantly CCAT consensus sequence but also able to bind the ACAT sequence (Yant et al., Nucleic Acids Res., 1995).

The detection of the recognition sequence (TCTCGCGAGA) of general factor X (GFX) also preceded the identification of the binding protein. It was not easy to "deorphanize" this motif, as it is only bound in the methylated state of the middle cytosines (Raghav et al., Mol. Cell, 2012). In this study, GFX motif was highly enriched in the ChIP regions of the SMRT co-repressor in preadipocytes. Based on a ChIP-seq motif enrichment derived from the Encyclopedia of DNA Elements (ENCODE), this motif seemed to be bound by the "zinc-finger and BTB domain containing 33" (ZBTB33) or KAISO, which was then confirmed experimentally. KAISO had been first described as an interacting TF of catenin

delta 1 (CTNND1) having a BTB interaction domain and a DBD with three C2H2-type zinc-finger motifs (Daniel and Reynolds, Mol Cell Biol., 1999). Then it has been shown to work as a DNA methylation dependent repressor (Prokhortchouk et al., Genes Dev., 2001). There are three paralogues of "GFX": ZBTB4, 21 and 38, which all are longer, having several additional C2H2-type zinc-finger motifs (Ensembl database), thus these may bind different elements, but until now, their role remained unknown.

It has been likewise unknown which gene coded general factor Y (GFY), but by now, ChIP-seq results suggest solutions for this issue as well. The motif that was enriched in promoter proximal regions (Xie et al., Nature, 2005; Heinz et al., Mol. Cell, 2010; Hong et al., Genomics Inform., 2012) seems partially identical (TCCCA) with the one of "selenocysteine tRNA gene transcription activating factor" (STAF) also known as zinc-finger protein 143 (ZFP143) (Schaub et al., EMBO J., 1997), however the other part (ACTACA) seems to be bound by another domain or protein that can be searched in the interaction partners of STAF. STAF may be able to bind this "half site" too, because it includes seven tandemly repeated C2H2-type zinc-fingers and as an activator, it is able to bind other DNA sequences (Schuster et al., EMBO J., 1995).

There is an even better candidate to be GFY called Ronin or THAP11. THAP name is derived from "Thanatos (Greek death god) associated protein" as THAP1 has been described as a proapoptotic factor (Roussigne et al., Oncogene, 2003). THAP has a typically N-terminal C2CH-type zinc-finger DBD that is known from the *Drosophila* P element transposase (Roussigne et al., Trends Biochem Sci., 2003). In human, there has been 12 THAP proteins described that were numbered form 0 to 11, but it became clear that these are coded by less genes. In mouse, there are only 8 THAP genes (1-4, 6-8 and 11). The name "Ronin" means a masterless Japanese samurai; as this protein has been seen independent from any pluripotency master regulators, however it is related to pluripotency via GFY motif binding (Dejosez et al.,

Cell, 2008; Dejosez et al., Genes Dev., 2010). More recently, it has been shown that host-cell factor C1 (HCFC1), which is a co-regulator with both activator and repressor functions, together with Ronin associates with other CpG-binding TFs as GABP, YY1 and STAF (Michaud et al., Genome Res., 2013). Thus it is really imaginable that there is a cooperation or competition between Ronin and STAF for the binding of GFY elements.

Beside the most common promoter proximal regulatory regions, there are several further elements, enhancers or silencers, which are bound by similar or the same domains as listed above. These are the different HTH (such as winged helix), bZIP and zinc-finger domains, but HLH – in addition to the bHLH-ZIP domain – is similarly capable of DNA binding. There are hundreds of zinc-finger proteins (ZFPs and ZNFs), which can recognize numerous sequences, but most of them are unknown. One of the best-known zinc-finger superfamilies is the group of the nuclear receptors.

1.3. Nuclear receptors

Nuclear receptor (NR) is a talkative name, as covers such TFs recognizing signal molecules. Upon ligand binding, their conformation changes and this modulates their affinity to the corresponding response elements. In the same time, NRs may join to or accumulate the members of PIC (Mangelsdorf et al., Cell, 1995; Nuclear Receptor Nomenclature Committee, Cell, 1999; Evans and Mangeldorf, Cell, 2014). This happens in the appearance (influx or production) of membrane-diffusible lipids, which can penetrate through the cell and nuclear membranes. These lipids are generally hormones including steroids, retinoids, thyroid hormone and vitamin D, but there are several members of the superfamily without identified ligand called orphan receptors. DBD, between the variable N-terminal A/B and the middle hinge domains, is composed of two conserved zinc-fingers, which bind typically to the AGGTCA sequence. The C-terminal ligand-binding domain (LBD) provides ligand

specificity and selectivity, and also dimerization and other interaction surfaces (Mangelsdorf et al., Cell, 1995; Grosdidier et al., Mol. Endocrinol., 2012); in the case of orphan receptors, this is responsible for the constitutive or posttranslational modification dependent regulation.

NRs can be discriminated into four classes (Mangelsdorf et al., Cell, 1995; Evans and Mangeldorf, Cell, 2014) (**Table 5**). The first cloned NRs – GR (Hollenberg et al., Nature, 1985) and ER (Green et al., Nature, 1986) – were the founding members of the class of steroid hormone receptors (class I). Glucocorticoid, mineralcorticoid, progesterone, androgen and estrogen receptors (GR, MR, PR, AR and ER alpha and beta, respectively) form homodimers, which bind inverted repeats of the AGAACA half site, or in the case of ERs, the RGGTCA half site. These elements are called IR3 because there are 3 spacer nucleotides in them between the two NR half sites. Class II receptors form heterodimer with retinoid X receptor (RXR) (Mangelsdorf et al., Nature, 1990) and usually bind direct repeats (DRs) of the RGKTCA half site. RXR, however its ligands, 9-cis retinoic acid and 9-cis-13,14-dihydroretinoic acid are described (Rühl et al., PLoS Genet., 2015), belongs to class III, which collects the dimeric orphan receptors that also bind dominantly DRs. Monomeric orphan receptors of class IV bind the consensus hexamer.

The heterodimerizing partners of RXR (class II) bind a large variety of lipids (Dawson and Xia, Biochim Biophys Acta., 2012). Peroxisome proliferator-activated receptors (PPAR alpha, delta and gamma) bind polyunsaturated fatty acids (PUFAs) with high affinity but can bind also monounsaturated (MUFA) and saturated fatty acids (SFA). These PUFA ligands are typically eicosanoids such as leukotrienes and prostaglandins. Knowing the physiological significance of PPARs e.g. in adipocytes and obesity, a high number of synthetic agonists, thiazolidenediones or glitazones were developed for medical use (Clarke et al., Am J Clin Nutr., 1999). Retinoic acid receptors (RAR alpha, beta and gamma) bind all-trans and 9-cis retinoic acid (ATRA and 9cRA) hormones, derivatives of vitamin A (retinol), but numerous

agonists and antagonists were developed for them, as well. Thyroid hormone receptors (TR alpha and beta) are also endocrine receptors, which bind triiodothyronine (T3).

Class	Short name	Trivial name	Regular name	Binding site
Class I	ERa / ESR1	Estrogen receptor alpha / 1	NR3A1	IR3 (AGGTCA)
	ERb / ESR2	Estrogen receptor beta / 2	NR3A2	
	GR	Glucocorticoid receptor	NR3C1	IR3 (AGAACA)
	MR	Mineralcorticoid receptor	NR3C2	
	PR / PGR	Progesterone receptor	NR3C3	
	AR	Androgen receptor	NR3C4	
Class II	TRa / THRa	Thyroid hormone receptor alpha	NR1A1	DR4
	TRb / THRb	Thyroid hormone receptor beta	NR1A2	
	RARa	Retinoic acid receptor alpha	NR1B1	DR5, DR2, (DR1)
	RARb	Retinoic acid receptor beta	NR1B2	
	RARg	Retinoic acid receptor gamma	NR1B3	
	PPARa	Peroxisome proliferator-activated receptor alpha	NR1C1	DR1
	PPARd / PPARb	Peroxisome proliferator-activated receptor delta / beta	NR1C2	
	PPARg	Peroxisome proliferator-activated receptor gamma	NR1C3	
	EcR	*Ecdysone receptor*	*NR1H1*	*IR1*
	LXRb	Liver X receptor beta	NR1H2	DR4
	LXRa	Liver X receptor alpha	NR1H3	
	FXRa	Farnesoid X receptor alpha	NR1H4	DR5, IR1
	FXRb	Farnesoid X receptor beta	NR1H5	
	VDR	Vitamin D receptor	NR1I1	DR3
	PXR	Pregnane X receptor	NR1I2	DRs
	CAR	Constitutive androstane receptor	NR1I3	DR5
Class III	REV-ERBa	Reverse-ERB alpha	NR1D1	DR2
	REV-ERBb	Reverse-ERB beta	NR1D2	
	HNF4a	Hepatocyte nuclear factor 4 alpha	NR2A1	DR1
	HNF4g	Hepatocyte nuclear factor 4 gamma	NR2A2	
	RXRa	Retinoid X receptor alpha	NR2B1	DRs (RGKTCA)
	RXRb	Retinoid X receptor beta	NR2B2	
	RXRg	Retinoid X receptor gamma	NR2B3	
	TR2	Testicular orphan receptor 2	NR2C1	DR1
	TR4	Testicular orphan receptor 4	NR2C2	
	COUP-TFa	Chicken ovalbumin upstream promoter-transcription factor alpha	NR2F1	DRs, IRs
	COUP-TFb	Chicken ovalbumin upstream promoter-transcription factor beta	NR2F2	
	COUP-TFg / EAR2	Chicken ovalbumin upstream promoter-transcription factor gamma / V-ErbA-Related Protein 2	NR2F6	
	GCNF	Germ cell nuclear factor	NR6A1	DR0

Table 5. The classification of nuclear receptors (part 1)

Class	Short name	Trivial name	Regular name	Binding site
Class IV	*KNIRPS*	*KNIRPS*	*NR0A1*	Half site
	KNRL	*KNIRPS-like*	*NR0A2*	
	DAX1	Dosage-sensitive sex reversal-adrenal hypoplasia congenital critical region on the X chromosome, gene 1	NR0B1	
	SHP	Short heterodimeric partner	NR0B2	
	RORa	RAR-related orphan receptor alpha	NR1F1	
	RORb	RAR-related orphan receptor beta	NR1F2	
	RORg	RAR-related orphan receptor gamma	NR1F3	
	TLX	Tailless homolog orphan receptor	NR2E1	
	PNR	Photoreceptor-cell-specific nuclear receptor	NR2E3	
	ERRa / ESRRa	Estrogen related receptor alpha	NR3B1	
	ERRb / ESRRb	Estrogen related receptor beta	NR3B2	
	ERRg / ESRRg	Estrogen related receptor gamma	NR3B3	
	NGFI-B / NUR77	Nerve-growth-factor-induced gene / Nuclear hormone receptor 77	NR4A1	
	NURR1	Nur-related factor 1	NR4A2	
	NOR1	Neuron-derived orphan receptor 1	NR4A3	
	SF-1	Steroidogenic factor 1	NR5A1	
	LRH1	Liver receptor homolog-1	NR5A2	

Table 5. The classification of nuclear receptors (part 2)

The following NRs of class II all have steroid derivative ligands (Dawson and Xia, Biochim Biophys Acta., 2012): Vitamin D receptor (VDR) binds the active form of vitamin D, the 1,25a-(OH_2)-vitamin-D_3 hormone; and liver X receptors (LXR alpha and beta) bind oxysterols. Although farnesoid X receptors (FXR alpha and beta) had been named from the isoprenoid farnezol, farnezol binding by the LBD of FXR has never been demonstrated; its natural ligands are the primary bile acids (e.g. chenodeoxycholic acid) (Edwards et al., J Lipid Res., 2002). The xenobiotic-sensing pregnane X receptor (PXR) (Kliewer et al., Cell, 1998) or steroid X receptor (SXR) (Blumberg et al., Genes Dev., 1998) was first described in *Xenopus laevis* as an orphan receptor (xONR1) (Smith et al., Nucleic Acids Res., 1994), but it revealed that by having a large, flexible ligand-binding pocket it can accommodate to various

ligands such as pregnane derivatives (e.g. pregnenolone) and secondary bile acids (e.g. lithocolic acid) (Wu et al., Drug Discov Today, 2013). Constitutive androstane receptor (CAR) that had been called first MB67 (Baes et al., Mol Cell Biol., 1994) is an unusual NR having constitutive activity, which can be blocked by its natural ligands (androstane metabolites) (Choi et al., J Biol Chem., 1997). There were two potent activator molecules (TCPOBOP and CITCO) found for CAR thus it indeed became an adopted orphan receptor (Tzameli et al., Mol Cell Biol., 2000; Maglich et al., J Biol Chem., 2003).

Beside LXRs and FXRs, there is a further NR1H protein, which has been put to the beginning of the NR1 family (**Table 5**): Ecdysone receptor (EcR) is an arthropod specific receptor, however its orthologues have been found in nematodes, *Annelida* and *Echinodermata*, as well (Ensembl Genomes database). EcR has been described in *Drosophila* as a new member of the NR superfamily (class II) responsible for ecdysteroid binding thus ultimately for the metamorphosis (Koelle et al., Cell, 1991). Connection of this gene with RXR, itself is a good indicative that NRs are very ancient proteins and this class II NR – RXR interaction is specific probably for the whole *Animalia* kingdom.

Kazuhiko Umesono found a regularity in the DNA binding of these heterodimers, namely that – similarly to the relation of TRE/CRE – the spacer length between the half sites of the DR elements is specific for the heterodimers, so there is a receptor selectivity of the regulatory elements (Umesono et al., Cell, 1991): they described that VDR binds DR3, TRs bind DR4, and RARs prefer DR5 compared to the other elements (**Figure 4**). Then the list started to become complete: by now we know that PPARs bind DR1, RARs bind DR2 too, LXRs also bind DR4, FXR prefers IR1 over DR5, CAR binds DR5 as well, and EcR binds IR1 (Mangelsdorf et al., Cell, 1995). PXR is similarly flexible in this property by the capability of binding multiple DRs (Frank et al., J Mol Biol., 2005).

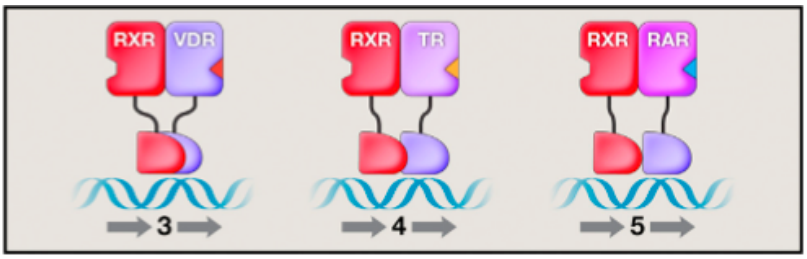


Figure 4. The motif specificity of class II nuclear receptor heterodimers (Evans and Mangeldorf, Cell, 2014)

In addition to RXR, which itself can form also homodimers (and bind DR1) *in vitro* (Mangelsdorf et al., Cell, 1995), the other class III NRs function variously, as well. REV-ERB alpha has been described as the reversely overlapping gene of TR alpha (Lazar et al., Mol Cell Biol., 1989), which had been called first "cellular avian erythroblastosis virus A" (C-ERBA) (Vennström and Bishop, Cell, 1982). REV-ERB alpha and beta typically bind DR2 elements and are responsible for setting the circadian rhythm. REV-ERB is likewise an adopted orphan receptor subfamily as these proteins have a tight LBD specialized for heme binding (Woo et al., J. Mol. Biol., 2007; Raghuram et al., Nat Struct Mol Biol., 2007). Hepatocyte nuclear factor 4 (HNF4 alpha and gamma) constitutively binds fatty acids (Wisely et al., Structure, 2002) and by DR1 binding maintains normal hepatic functions.

Testicular orphan receptors (TR2 and 4) bind DR1 elements too and seem really orphan receptors. TR2 has been shown to be a NR turned to an activator through the phosphorylation of its DBD (Khan et al., Proteomics, 2006). Chicken ovalbumin upstream promoter-transcription factor (COUP-TF) had been described also as a DR1 binder (Wang et al., J Biol Chem., 1987), however later it was observed on other DR and IR elements. COUP-TF beta represses several TFs including NRs: HNF4 is one of its targets (Achatz et al., Mol Cell Biol., 1997), but it interacts with AR by competing with its co-activators in prostate cancer cells (Song et al., PLoS One, 2012). This may be a regular heterodimerization as COUP-TF beta binds both the DBD and LBD of AR. The last member of class III is the germ

cell nuclear factor (GCNF), which binds DR0 elements in the ovary, testis and embryonic stem cells thus controlling gametogenesis and early embryogeneis, respectively (Chen et al., Mol Endocrinol., 1994; Greschik et al., Mol Cell Biol., 1999). This is likewise an ancient and indispensable protein as it is present in worms and insects, as well (Ensembl database).

1.4. Transcription factors in macrophage development

Macrophages (big eaters, Greek) (Metchnikoff, 1882) are professional phagocytes that are descendant of the myeloid lineage (e.g. Geissmann et al., Science, 2010). Their circulating precursors are the monocytes, which can differentiate into macrophages and also dendritic cells after exiting from the blood vessels. These two cell types are closely related having similar morphology and similar, e.g. phagocytic functions, thus both has been called histiocytes (tissue cells) in connective tissue. There are several kinds of special macrophages determined by the tissue environment: Kupffer cells are present in the liver (Kupffer, 1876), Langerhans cells can be found in the skin (Langerhans, 1868), microglia is specific for the central nervous system (Hortega, 1920), osteoclasts are responsible for bone resorption (Kolliker, 1873) and alveolar macrophages keep the lungs clean (Cowdry, J Exp Med., 1925). By now it became clear that a significant amount of macrophages show self-renewal (Sieweke and Allen, Science, 2013). All microglia and the majority of Langerhans cells and alveolar macrophages are derived from the multipotent precursors developed from the early embryonic progenitors of the yolk sac or even the hematopoietic cells of the fetal liver. The other extreme is the skin macrophages, and most of the mucosal and peritoneal macrophages, which are mainly resupplied by bone marrow derived monocytes. The macrophages of other organs are derived approximately equally from stem cell-like precursors and monocytes.

Although their origin and tissue milieu modulate macrophage function and metabolism – e.g. Kupffer cells similarly to hepatocytes produce bilirubin (Gottlieb, Can Med

Assoc J., 1934) –, it seems that macrophages always share in the same TFs. With the help of these TFs, more cell types could be transformed into macrophages: the retroviral expression of PU.1 (ETS) and C/EBP alpha (bZIP) could reprogram different kinds of fibroblasts into macrophages (Feng et al., Proc Natl Acad Sci U S A., 2008), while C/EBP alpha alone could trans-differentiate pre-B cells into the macrophage state by repressing the B-cell specific signals (Di Tullio et al., Proc Natl Acad Sci U S A., 2011). This is not that surprising as lymphoid cells also express PU.1, which together with the C/EBP alpha can reach the same effect that was observed at fibroblast cells. This means that the high amount of PU.1 and C/EBP alpha can reprogram probably any cell types similarly to other lineage determining factors, such as MYOD in the muscle (Cooper et al., Neuromuscul Disord., 2007) and NEUROG2 in neurons (Aravantinou-Fatorou et al., Stem Cell Reports, 2015) or the Yamanaka factors in stem cells (Takahashi and Yamanaka, Cell, 2006). During macrophage differentiation, "Runt-related transcription factor 1" (RUNX1) is up regulated by FLI1 (ETS) and together with "T-cell acute lymphocytic leukemia protein 1" (TAL1, bHLH) and C/EBP beta, it induces PU.1 expression, which together with C/EBP alpha ultimately determines the macrophage lineage (Lichtinger et al., EMBO J., 2012). C/EBP beta, FLI1 and later PU.1 co-localize with AP-1, which also seems to be an important factor in the macrophage differentiation process. TAL1, FLI1 and RUNX1 are essential for haematopoiesis and C/EBP alpha and beta proteins are critical to maintain normal macrophage function.

There is a further group of proteins essential for macrophage function, the interferon (IFN) regulatory factor (IRF) family, which has 9 members (IRF1-9) (Miyamoto et al., Cell, 1988; Taniguchi et al., Annu Rev Immunol., 2001). IRF proteins were described as effectors of pattern-recognition receptors (PRRs), which are activated by pathogen-associated molecular patterns (PAMPs) (Medzhitov and Janeway, Science, 2002; Honda and Taniguchi, Nat Rev Immunol., 2006): Toll-like receptors (TLRs) (Rock et al., Proc. Natl. Acad. Sci. U S

A., 1998) recognize several kinds of PAMPs (e.g. Takeuchi and Akira, Cell, 2010), retinoic acid inducible gene-I (RIG-I)-like receptors (RLRs) are specialized dominantly to dsRNA viruses (Yoneyama and Fujita, Immunol Rev., 2009), nucleotide-binding oligomerization domain (NOD)-like receptors (NLRs) are able to recognize several types of bacterial peptidoglycans (Inohara et al., Annu. Rev. Biochem., 2005) and C-type lectin receptors (CLRs) bind typically carbohydrates on microorganisms such as fungi (Willment and Brown, Trends Microbiol., 2008).

The discovery of IRF element, the IFN-stimulated response element (ISRE, GAAANNGAAA) also preceded the identification of its binding proteins (Levy et al., Genes Dev., 1988). ISRE is a direct repeat element, which suggests that it is bound by the homo- or heterodimer of two IRFs, and indeed, most PRR pathways had been shown to effect through IRF3/IRF7 heterodimers (Honda and Taniguchi, Nat Rev Immunol., 2006). Beside the GAAA sequence, IRFs have further nucleotide preference: the sequence of the extended half site is AANNGAAA, where the upstream AA (and TT on the other strand) is bound in the minor groove, thus the upstream co-binding IRF partner is not superseded in the major groove (Panne et al., Cell, 2007). Beside the about 120 amino-acid long N-terminal DBD, which also shows IRF-IRF interaction, there is a C-terminal IRF-associated domain (IAD) in all family members: IAD1 in IRF3-9 and IAD2 in IRF1-2 proteins (Meraro et al., J Immunol., 1999). The phosphorylation of the C-terminal region promotes dimerization and beside ISRE occupation, the binding of e.g. CBP/P300 co-activators (Chen et al., Nat Struct Mol Biol., 2008).

IRF proteins form heterodimers also with PU.1. Their common binding site (GGAAgtGAAA) – called ETS-IRF composite element (EICE) (Brass et al., Genes Dev., 1996) – has been shown to be bound by PU.1/ IRF1, PU.1/IRF8 (Eklund et al., J Biol Chem., 1998) and PU.1/IRF4 heterodimers (Meraro et al., J Immunol., 1999). This type of

interactions seems to be less dependent on PRR pathways because IRFs in these complexes rather work as lineage determining factors. In unstimulated bone marrow derived macrophages (BMDMs), IRF8 typically occupies EICEs together with PU.1, while lipopolysaccharide (LPS) stimulation directs IRF8 to "triple" IRF sites (GAAAnnGAAAnnGAAA), which DNA-protein interaction seems only rarely PU.1-dependent (Mancino et al., Genes Dev., 2014). At these sites, IRF8 interacts with e.g. IRF1, but at other sites, the AP-1/IRF (AICE) and IRF/PU.1 composite elements are enriched likewise. Previously in B cells, it has been shown that the BATF/JUN-IRF4 complex bound AICEs (Li et al., Nature, 2012; Glasmacher et al., Science, 2012). The lineage-determining role of IRF4 and IRF8 had been already confirmed earlier in mouse knock out (KO) experiments, where the lack of IRF4 caused impaired lymphocyte function (Mittrücker et al., Science, 1997), while the lack of IRF8 caused chronic myelogenous leukemia-like syndrome (Holtschke et al., Cell, 1996). But of course there are several types of lymphoid and myeloid cells with different IRF preferences. IRF7 has been shown to be essential for the monocyte-macrophage differentiation (Lu and Pitha, J Biol Chem., 2001), but this effect could not be attributed to a direct interaction with PU.1.

1.5. NGS methods in functional genomics

1.5.1. Getting and aligning raw NGS data

The structural and functional features of the chromatin are examined nowadays with the opportunities of NGS. Several methods were expanded to genome-wide scale by developing the specific DNA library preparation protocols. The appropriately sized DNA fragments e.g. from ChIP experiments just need adaptor ligation (Robertson et al., Nat Methods. 2007), while RNA has to be reverse transcribed before and then get the adaptor sequences (Nagalakshmi et al., Science, 2008; Wilhelm et al., Nature, 2008). Raw sequencing

data looks very similarly in all cases: one get tens of millions of relatively short (usually 50 or 100 nucleotide long) sequence reads, which are the single or paired ends of the random fragments generated during library preparation. In the case of whole genome, DNA input or control IgG ChIP sequencing, sequences cover the given genome approximately equally and show the repetitive/lacking regions of the given organism, thus these latter two are often used as control for ChIP-seq samples. The other methods give different types of read enrichments: RNA-seq derived fragments – because of splicing – usually cover only the expressed exons thus reads draw column-like shapes, while fragments derived from global run-on sequencing (GRO-seq) (Core et al., Science, 2008), nuclear RNA-seq, and phospho-PolII ChIP-seq draw the whole primary transcripts (Mitchell et al., PLoS One, 2012). The nature of ChIP-seq coverage depends on the used antibody. In the case of antibodies for TFs, fragments contain the distinct binding sites showing together a quasi-normal distribution and the position of the highest coverage (summit) of the peaks will appoint to the location of the occupied TFBSs (Wang et al., Genome Res., 2012). Fragments fished out with antibodies for histone modifications usually draw hill and valley-like landscapes.

The first step during processing raw NGS data is mapping the sequence reads onto the given reference genome. For this, one can apply the "basic local alignment search tool" called BLAST (Altschul et al., J Mol Biol., 1990), but it would take even months depending on the read number, the genome size and the performance of the used computer. This was the reason to develop faster and faster aligner tools. BLAST-like alignment tool (BLAT) was 500 times faster than the popular aligners in the beginning of the century (Kent, Genome Res., 2002). After 2007, more and more sophisticated programs suddenly followed BLAT. The initial NGS studies used the "short oligonucleotide alignment program" (SOAP) (Li et al., Bioinformatics, 2008), MAQ (Li et al., Genome Res., 2008) and Bowtie (Langmead et al., Genome biology, 2009), but by now, there are several newer and updated read mappers, as

well. The use of Burrows-Wheeler transformation (BWT) (Burrows and Wheeler, Digital Equipment Corporation, 1994) to create the so-called prefix tries – a kind of "index" – for each read was a big step to make alignment even faster. After its integration into SOAPv2 and Bowtie, the creators of MAQ utilized also BWT for their new program, the Burrows-Wheeler alignment (BWA) tool (Li and Durbin, Bioinformatics, 2009). The use of paired-end reads makes fragment mapping more sure and reliable and can extend the covered regions if the tags do not overlap. This feature is now available for the most software.

The alignment of spliced reads for RNA-seq was a unique challenge, which had been overcome by BLAT (Kent, Genome Res., 2002) and the nowadays TopHat software (developed from Bowtie) (Trapnell et al., Bioinformatics, 2009). TopHat was complemented to a widely used RNA-seq analysis pipeline including the Cufflinks (Trapnell et al., Nat Biotechnol., 2010) and CummeRbund software (Trapnell et al., Nat Protoc., 2012). Cufflinks is a package, which builds up each detectable, even unknown transcript (and also splice variants) and counts their expression values. The gene expressional levels of different cell types or conditions that are usually determined in FPKM (fragments per kb of exon per million mapped fragments) values can be compared by Cuffdiff; and CummeRbund was developed for the further analysis, exploration and visualization of the data. RPM (reads per region per million mapped reads) and RPKM (reads per kb of a region per million mapped reads) are similar units used for measuring enrichments of (other) NGS methods generating single end reads.

Raw sequence data are typically stored in FASTA (Pearson and Lipman, Proc. Natl. Acad. Sci. U S A., 1988) or FASTQ format, which latter was first used by the Open Bioinformatics Foundation (OBF) (http://www.open-bio.org; Cock et al., Nucleic Acids Research, 2010). Both format let the storage of several sequences, but FASTQ was developed right for the NGS reads having quality score for each nucleotides. "Quality" is symbolized as

“Q” in the “FASTQ” format name, while “A” means “all”, both nucleotide and peptide sequences in the “FASTA” term. FASTA format contains a header line starting with “>” character followed by a short description – even with some detailed information – of the sequence, which is broken into the next lines usually per 60 nucleotides (or amino acids). FASTQ format is built up from four-line units: the sequence identifier is started by “@” symbol, the second line contains the sequence, the third line is for an optional description following a “+” symbol (which is obligatory) and the fourth line contains the quality scores.

1.5.2. The determination of read and motif enrichments

By aligning sequences to a reference genome, one gets genomic location information, as well, which is stored in “sequence alignment/map” (SAM) or in its compressed version, the “binary alignment/map” (BAM) format, which can be created by SAMtools (Li et al., Bioinformatics, 2009). BAM has a belonging index file format (BAI), which is needed for visualization in genome browsers such as the UCSC genome browser (Kent et al., Genome Res., 2002), the Ensembl browser (Hubbard et al., Nucleic Acids Res., 2002) or the Integrative Genomics Viewer (IGV) (Thorvaldsdottir et al., Brief Bioinform., 2012). BAM files contain both the sequences and the genomic coordinates, thus allelic variants are traceable if the (local or whole genome) coverage is sufficiently high. “Browser extensible data” or BED is a simpler “tab separated value” (TSV) text format containing no sequence information, only genomic coordinates in field 1-3; and the identifier, score and strand information are optional in field 4-6, respectively. There are further possibilities in this format: it may contain further data thus look like “general feature format” (GFF) or “general transfer format” (GTF) in genome browsers. Regions/boxes can be colored (e.g. strand) specifically and there is also a genomic junction feature built in, which all need a specific header line and the filling of further fields. By using BEDTools, one can compare genomic

coordinate files, e.g. get intersections, difference sets, distances and merges of the regions, even strand specifically (Quinlan and Hall, Bioinformatics, 2010).

Several programs that generate primarily BED format can predict regions with statistically significant read enrichment. For TF ChIP-seq, there are several peak caller programs that have been developed using different algorithms (Bailey et al., PLoS Comput Biol., 2013): ChIP-seq analysis in R (CSAR) searches for read enrichments following Poisson distribution (Muiño et al., Plant Methods, 2011); the model-based analysis of ChIP-seq (MACS) (Zhang et al., Genome Biol., 2008) and the hypergeometric optimization of motif enrichment (HOMER) (Heinz et al., Mol. Cell, 2010) works based on the local density of the reads also with Poisson distribution; CisGenome uses negative binomial (Ji et al., Nat Biotechnol., 2008), and ZINBA uses zero inflated negative binomial distribution (Rashid et al., Genome Biol., 2011). There are further, even more sophisticated methods for peak calling, such as the BayesPeak, using the hidden Markov model (Spyrou et al., BMC Bioinformatics, 2009). Most of these methods deal with input and IgG controls, which provide relative local coverage thus smoothing the background and highlighting the enrichments. This list of approaches testifies that there is still a lack of consensus on what to call a peak and how to count their number. It appears though, that peak caller tools work relatively well depending on the quality of the sample libraries, and all give a quality/height score, the edges and the summit of the peaks.

Peak width is usually a technical issue depending on fragment length, but the summit has a biological relevance, as it is very close to the putative TFBS(s) (Wang et al., Genome Res., 2012). In several cases, TFBSs – thus also the peak summits – are closer to each other than the average fragment length. This needs a special approach to separate these double or even multiple peaks, which approach takes account into the reads of the immediate vicinity of the enrichment not only those of the broader background. For this, MACS integrated an

independent software named PeakSplitter (EMBL-EBI Bertone Group Software). It is a command line program for separating broader regions into subpeaks based on coverage information generated e.g. by MACS(2). The two main types of continuous-valued coverage files (unlike BED files) are BEDGRAPH and wiggle (WIG), which latter can be indexed and binary compressed to BIGWIG format (Rhead et al., Nucleic Acids Res., 2010). Similarly, IGVtools makes possible to create indexed and also binary compressed BEDGRAPH files called "tiled data files" or TDFs, which need much less memory to visualize and browse them, as well (Thorvaldsdottir et al., Brief Bioinform., 2012).

Peak summits are suitable to detect direct or indirect binding of the different TFs and co-regulators. By using co-activators such as P300, it is predictable which TFs are acting in a given cell at the different regulatory sites. The first widely used package, which was able to search for motif enrichments and to map the found matrices (back) to the sequences thus designating the TFBSs, was the "multiple expectation-maximization for motif elicitation" (MEME) and the "motif alignment search tool" (MAST), respectively (Bailey et al., J. Steroid Biochem. Mol. Biol., 1997). Based on the validated TFBSs, more motif matrix databases were established: TRANSFAC (Knuppel et al., J. Comput. Biol., 1994), MEME and JASPAR (Sandelin et al., Nucleic Acids Res., 2004) have databases with their own similarity weight matrix formats, which are now used in several other databases. HOMER also developed a system for motif enrichment and TFBS search, which include matrices of others as well as the ones with HOMER's own format enriched from numerous ChIP-seq samples (Heinz et al., Mol. Cell, 2010). There is a special ChIP-seq method called ChIP-exo, developed directly for the detection of TFBSs at a single nucleotide resolution, in which fragments are shortened by an exonuclease from the 5' up to the TFBS, thus resulting in a column-like shape after the alignment and marking the exact place of DNA binding (Rhee and Pugh, Cell, 2011). The application of this method may result a much better resolution in determining TFBSs

compared to a simple ChIP-seq, but DNase-seq (Boyle et al., Cell, 2008) is also suitable to predict the exact binding sites (Piper et al., Nucleic Acids Res., 2013).

Histone modifications usually cover broader regions of the genome, which calls for different kinds of algorithms. The spatial clustering for identification of ChIP-enriched regions (SICER) (Zang et al., Bioinformatics, 2009) and ZINBA (Rashid et al., Genome Biol., 2011) were developed for this purpose, but certain peak callers such as MACS2 and HOMER are also able to find these kinds of regions by using different parameters (Bailey et al., PLoS Comput Biol., 2013). The DNA binding of TFs disrupts histone continuity establishing the so-called nucleosome-free (or more precisely nucleosome-depleted) regions (NFRs) with valley-like shapes in the histone modification landscapes. There are several methods allowing the detection of nucleosome occupied and depleted regions. The first ones worked based on ChIP-chip data (Yuan et al., Science, 2005; Lee et al., Nat. Genet., 2007), while the newer ones are working based on MNase (Albert et al., Nature, 2007) or ChIP-seq data (Sun et al., PLoS One, 2009). HOMER also includes an NFR prediction function for ChIP-seq data searching regions with the greatest differential in ChIP signal.

Special NGS methods emerged in order to determine the regulatory regions and as Sono-seq (Auerbach et al., Proc. Natl. Acad. Sci. U S A., 2009), FAIRE- (Gaulton et al., Nat. Genet., 2010) and ATAC-seq (Buenrostro et al., Nat Methods, 2013), give peak-like enrichments. Peak callers are also suitable in these cases to determine NFRs. Sono-seq is a simple method to find accessible chromatin regions using a size selection following the sonication of the cross-linked chromatin. Assay for transposase accessible chromatin coupled with NGS (ATAC-seq) is a newer method with a very simple and quick protocol for the detection of the possible regulatory elements of even only a few thousand cells. Nowadays, the predictors of peaks, broader regions (SICER, ZINBA) and NFRs are becoming suitable for the processing of DNase I, MNase-seq or any kinds of NGS data.

2. Aims of the study

Macrophage is a well-characterized cell type. This is why mouse BMDM is a good model to investigate the epigenetic landscapes determined by TFs. Active regulatory regions occupied by TFs and their co-activators are typically surrounded by active histone marks having a valley-like shape also called nucleosome-free region or NFR. In histone modification ChIP-seq enrichments, these are the functionally most important regions, so we wanted to determine all putative NFRs. For this, we carried out ChIP-seq for H3K4me2, H3K4me3, H3K27ac and H4ac, and developed a novel NFR prediction method. For comparison, we also applied a hidden Markov model based approach integrated in the HOMER software package. The next goal was to validate our method with motif enrichment analyses and to determine the known and less known TFs that are able to bind the elements matching with the found motifs. But the mapped TFBSs might be occupied by several members of the different TF (super)families, so we needed gene expressional data for the identification of the active DNA-binding proteins. To find the acting TFs, we performed GRO-seq and RNA-seq to detect the nascent and matured mRNA level of the putative regulators. We also aimed to focus on the DNA-binding of PU.1, which is known as a pioneer and lineage determining TF in macrophages.

Nuclear receptors (NRs) enable the cells to sense and respond to intrinsic and environmental lipid signals. Class II NRs, the heterodimerizing partners of RXR, recognize a large variety of lipid molecules, and then – together with RXR – activate different pathways. Thus, RXR seems rather a passive, assistant molecule that is unable to act on its own. Upon the RXR ligandation of BMDM cells we would have been expected gene activation by LXRs, RARs and PPARs, but it was a question whether there were distinct pathways specific only for RXR. To answer this question, we carried out ChIP-seq for RXR, the co-activator P300 and the lineage determining TF PU.1; and to follow gene expressional changes upon

treatment, we applied GRO-seq and RNA-seq. GRO-seq lets detect the direct regulatory effects not only at the level of gene expression but also at the level of enhancer transcription, thus based on proximity, it can be used to assign the active enhancers to regulated genes. As there were no applicable tools to distinguish the transcribed regulatory regions from the expressed genes, we developed a tool that was able to predict and annotate each transcriptional event. To confirm the annotation of RXR mediated regulatory sites, we carried out ChIP-seq also for CTCF and RAD21 proteins to detect the insulator regions that assigned the borders of the regulatory units. For this, another prediction method was developed. Finally, at some selected genomic regions, chromosome conformation capture coupled with NGS (3C-seq) was performed to corroborate the found promoter–enhancer interactions.

3. Materials and Methods

3.1. The differentiation of bone marrow derived macrophage (BMDM) cells

Bone marrow was flushed from the femur of wild-type C57BI6/J male mice. 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 (including macrophage colony-stimulating factor, MCSF) for 5 days. For ChIP experiments, at least two biological replicates were used; and cells were treated for 1 hour with vehicle or 100 nM LG268 (LG100268) ligand, gift from M. Leibowitz (Ligand Pharmaceuticals). Experimental procedures were done by Bence Dániel according to the protocol of Barish et al. (Barish et al., Mol Endocrinol., 2005).

3.2. Chromatin immunoprecipitation coupled with next-generation sequencing

BMDM cells were cross-linked with di(N-succinimidyl) glutarate (DSG) (Sigma) for 30 minutes and then with formaldehyde (Sigma) for 10 minutes. After fixation, chromatin was sonicated with Diagenode Bioruptor to generate 200-1000 bp fragments. Chromatin was immuno-precipitated with pre-immune IgG (Millipore, 12-370) and antibodies against H4ac (Millipore, 06-866), H3K27ac (ab4729), H3K4me2 (Upstate, 07-030), H3K4me3 (ab8580), RXR (sc-774), P300 (sc-585), PU.1 (sc-352), CTCF (Millipore, 07-729) and RAD21 (ab992). Respectively about 2 and 10 million cells were used for histone and TF ChIP-seq. Chromatin-antibody complexes were precipitated with protein A coated paramagnetic beads (Life Technologies). After 6 washing steps complexes were eluted and reverse cross-linked. DNA fragments were column purified (Qiagen, MinElute). The amount of immuno-precipitated DNA was quantified with Qubit fluorometer (Invitrogen). DNA was applied for quantitative PCR (QPCR) analysis or library preparation. QPCR results were presented as means +/-SD of

technical triplicates from more biological replicates. Experimental procedures were done by Bence Dániel based on the protocol of Barish et al. (Barish et al., Genes Dev., 2010).

ChIP-seq libraries were prepared with Ovation Ultralow Library Systems (NuGen) according to the manufacturer's instructions. 1 ng immunoprecipitated DNA was submitted to end repair reaction. Adaptors were ligated to end repaired DNA fragments. Libraries were amplified with adaptor specific primers in 16 PCR cycles and then were gel-purified with E-Gel systems (Life Technologies) to remove primers. Libraries were quantified by Qubit fluorometer and the quality was assessed with Agilent 1000 DNA Chip. Sequencing was carried out with Illumina HiScanSQ sequencer. Experimental procedures were done by Bence Dániel and Tibor Gyuris.

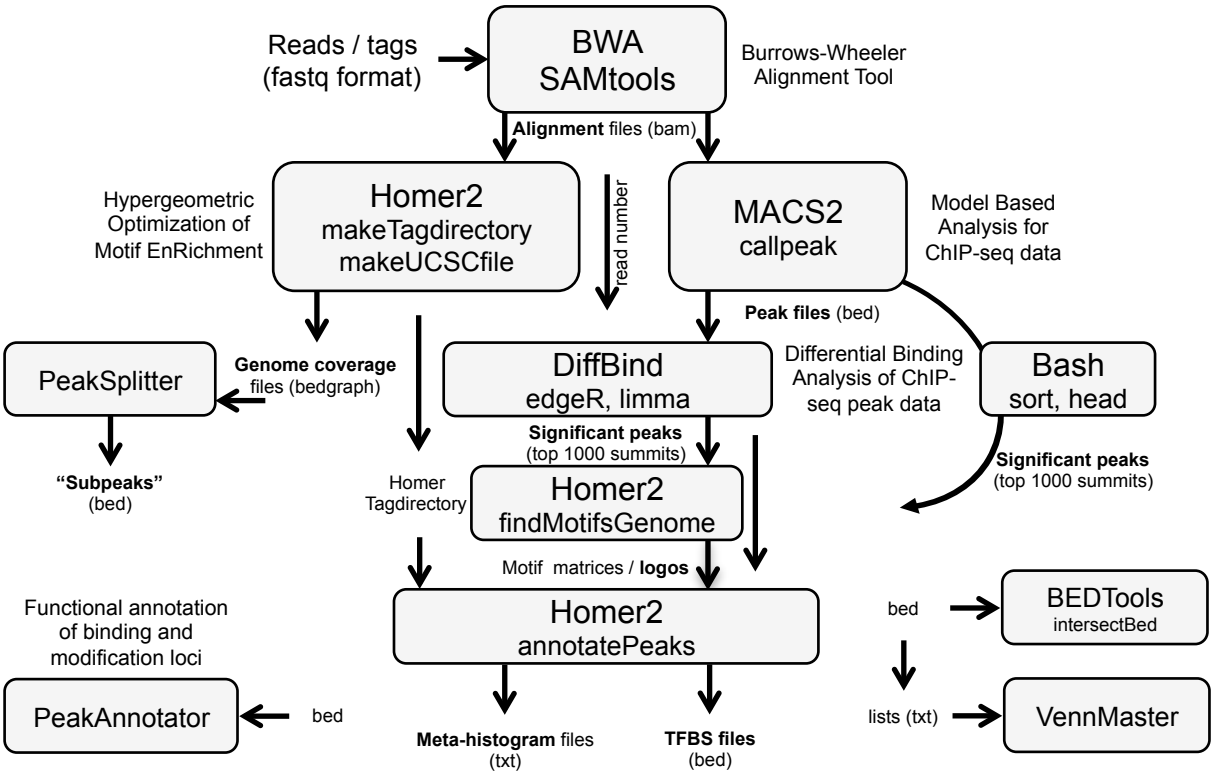


Figure 5. The bioinformatic pipeline used to analyze ChIP-seq data

3.3. ChIP-seq analysis

3.3.1. Primary analysis

The primary analysis of the ChIP-seq derived raw sequence reads has been carried out using our ChIP-seq analysis command line pipeline (Barta, EMBnet.Journal, 2011). Alignment to the mm9 mouse genome assembly was done by the BWA tool (Li and Durbin, Bioinformatics, 2009), and BAM files were created by SAMTools (Li et al., Bioinformatics, 2009) (**Figure 5**). Genome coverage (BEDGRAPH and TDF) files were generated by makeUCSCfile.pl (HOMER) (Heinz et al., Mol. Cell, 2010) and IGVtools, respectively, and used for visualization with IGV2 (Thorvaldsdottir et al., Brief Bioinform., 2012).

3.3.2. The determination of nucleosome-free regions (NFRs)

For the prediction of NFRs from histone modification data, PeakSplitter (EMBL-EBI Bertone Group Software) was applied to determine "subpeaks" – the nucleosome occupied regions (NORs) – by using the genome coverage information of BEDGRAPH files. NFRs received the sum of the "height" score of the surrounding two NORs (**Figure 6**). Score thresholds were set between 8 and 50 for the enhancer marks and between 25 and 500 for the H3K4me3 mark. The upper threshold was necessary to avoid the detection of artifacts as neither isotype nor any other controls were used for this analysis. Each NOR was limited in ±50 nucleotides compared to its summit and then further extended by 50 nucleotides towards the neighboring NOR(s). NFRs were defined as the regions between a pair of repositioned NORs if their distance was in the range of 50 and 1100 bp. The inner edges of the original NOR predictions were suitable for the better positioning of NFRs, therefore the regions between these edges (extended by 100 bp in both directions) were used as the second NFR prediction. Finally, the weighted average of the two kinds of NFR was calculated, where the summit-based prediction had two times the weight of the second NFR prediction.

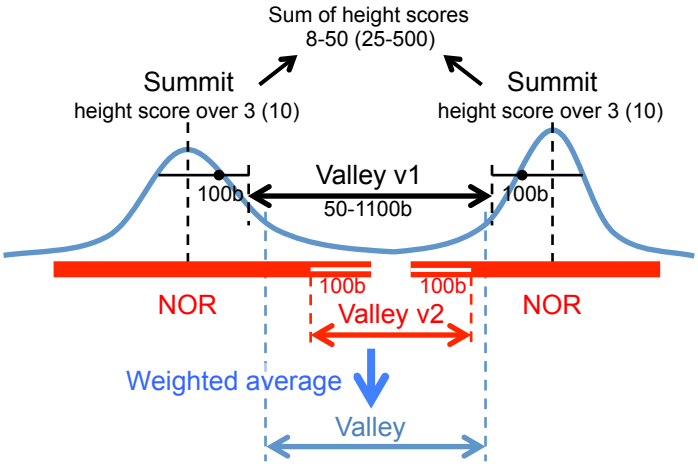


Figure 6. The scheme of NFR prediction (length and score thresholds are included; the H3K4me3 specific thresholds are in parentheses)

3.3.3. Peak prediction

The read enrichments (peaks) of transcription (co-)factors (and H3K4me3) were predicted by MACS2 (**Figure 5**). Artifacts were determined based on their presence in all, together 19 samples – including (co-)factor and also histone modification ChIP-seq results – in the same genomic location, and then eliminated from the peak sets. Two parallels of the control and LG268-treated RXR samples were analyzed by DiffBind v1.0.9 (Ross-Innes et al., Nature, 2012) with an input control: consensus peaks were determined from the peaks detected from at least two of the four samples; peaks with significantly changing "binding affinity" were defined using the "full library size" parameter. The peak score threshold of the further samples was selected manually.

3.3.4. Secondary analyses

The average read distribution histograms and heat maps centered to peak summits or the middle of the NFRs were made by annotatePeaks.pl (HOMER) (**Figure 5**). Overlaps were defined by BEDTools (Quinlan and Hall, Bioinformatics, 2010) and visualized by VennMaster-0.37.5 (Kestler et al., Bioinformatics, 2005). In the case of NFRs, the central 200 bp of the top 5000 hits (determined based on the scores derived from PeakSplitter) were used for the prediction of motif enrichments by findMotifsGenome.pl (HOMER). In the case of

peaks, summit +/-50 bp of the top 1000 peaks (determined based on the score given by PeakSplitter or MACS2) was used for the prediction of motif enrichments. On figures containing motif logo(s), target percent refers to the ratio of the peaks having the given motif, while background percent shows the hit ratio of an ~50,000 element set of random genomic sequences with 100 or 200 bp length in accordance with the target sequence length. P-values were calculated based on the comparison of these two ratios.

RXR is capable to bind multifarious DR sites (e.g. Umesono et al., Cell, 1991), but motif enrichment determination programs (including HOMER) with default settings are often unable to discriminate these DR elements from each other thus predicting the merge of the different DR motifs, which results the enrichment of NR half sites. For this reason, RXR motif enrichments have been predicted one-by-one based on their RGGTCANnRGGTCA consensus sequence with the optimization (-opt) function of findMotifsGenome.pl (HOMER). Under the annotated RXR peaks (described later at GRO-seq analysis), the extended sequences (11+6+11 bp) of the remaining putative composite elements without enrichment – detected as „half sites" – were obtained by homerTools extract, if the motif score exceeded 6. From these sequences, the DRn (RGKKSANnRGKKSA), ERn (TSMMCYNnRGKKSA) and IRn (RGKKSANnTSMMCY) elements were searched by fuzznuc (EMBOSS), where n was set between 0 and 5, and 1 mismatch was allowed.

3.4. Global run-on sequencing (GRO-seq)

Global run-on assay and library preparation was performed as described earlier with minor modifications (Core et al., Science, 2008; Hah et al., Cell, 2011). Shortly, for the run-on reactions, 8 million nuclei, 0.25 mM ATP, 1 uM alpha-32P-CTP, 0.25 mM GTP, 0.25 mM 5-Bromo-UTP and 0.5% Sarkosyl were used. After the isolation steps, RNA was hydrolyzed with 0.2 N NaOH. Upon purification, nascent RNA was enriched using anti-bromo-deoxy-U

antibody conjugated beads. Upon elution and precipitation, tobacco acid pyrophosphatase (TAP, Epicenter) was applied to remove 5' cap and then for end repair, T4 polynucleotide kinase (PNK, NEB) was used at low, then at high pH in the presence of ATP. 5' and 3' adaptors were ligated by T4 RNA ligase and both was followed by a bead-binding step. The affinity-enriched RNA was then reverse transcribed, amplified, and polyacrylamide gel electrophoresis (PAGE) purified. Libraries were generated from two biological replicates of the nuclei of BMDMs treated with 100 nM LG268 in a 0, 30, 60 and 120-minute time series. Libraries were sequenced with Illumina HiScanSQ sequencer. Experimental procedures were done by Nasun Hah and László Nagy.

3.5. GRO-seq analysis

3.5.1. Primary analysis

The primary analysis of the raw sequencing data has been carried out similarly as detailed for ChIP-seq. Alignment to the mm9 genome assembly was made by BWA tools after splitting the adaptors from the sequences by using FASTX toolkit. These steps were done by Endre Barta. The pool of sequence reads obtained from 2x4 macrophage samples was used for transcript prediction and annotation. Genome coverage (BEDGRAPH) files were generated by using makeUCSCfile.pl (HOMER) with the following parameters for both strands: -fsize 5e8 -fragLength 120 -noadj -style chipseq.

3.5.2. The determination of transcripts

Based on the pooled coverage information, all "subpeaks" – which had higher coverage compared to the neighboring regions and the background – were determined strand-specifically by PeakSplitter. The length of these "expressed units" was limited in +/-250 bp relative to their summit, and the number of unique reads was counted for each unit from the

pooled BAM file by intersectBed (BEDTools). Regulatory regions – irrespectively to their location relative to the genes – typically show divergent transcription, which are termed here as divergent sites. Divergent sites were determined based on those pairs of units, which were divergent, contained more than 15 reads, more than 33.33 reads per kb, and overlapped following a 150-150 bp shift towards the 5' direction. In the case of pairs with overlapping units (meeting with these criteria), units with the highest read number were selected.

BEDTools (mergeBed) and other command line programs were used to build up transcripts from the expressed units that were lying closer to each other than 600 bp on the same strand. As several "alignment gaps" could be found (mostly in introns, probably due to the genetic differences between the used strain and the reference genome), these gaps were filled in in the expressed transcripts according to the Ensembl reference genome annotation. The genomic location of transcripts was downloaded from Ensembl (GRCm37.p7) via BioMart. The short, typically non-coding transcripts overlapping in sense direction with protein coding transcripts were omitted from the further annotation steps.

As both promoters and enhancers show divergent transcription, the ChIP-seq read enrichment of H3K4me3 data was used to separate promoter regions from the proximal and intronic enhancers. H3K4me3 "subpeaks" were predicted in a very similar way as GRO-seq enrichments. Subpeaks that exceeded score 10 were limited in +/-1 kb compared to their summit (scores were determined by PeakSplitter). Of the overlapping ones, the most highly covered and the very upstream GRO-seq units were assigned to each H3K4me3 subpeaks, primarily from the divergent sites built up from at least 30 reads and then the other single initiation units above 50 reads. All possible transcripts were built up from the previous predictions based on these putative TSSs.

3.5.3. The annotation of transcripts

Longer transcriptional events were categorized by the following criteria: With regard to the TSRs and the alternative 5' UTR usage, a transcript was annotated as a known gene if their direction was identical and the predicted TSS marked by H3K4me3 was closer to the known TSS than 1.5 kb (group 1). When it was possible, 3' overhangs were collected separately from the transcript bodies. The remaining H3K4me3 marked intergenic transcripts with at least 1 kb length and – because of intronic enhancers that show also divergent transcription – the at least 3 kb long H3K4me3 marked antisense transcripts were collected in group 2. Putative transcripts (longer than 1 kb) with divergent or single TSS were collected from the remaining regions without H3K4me3 mark (group 3). The rest of the H3K4me3 marked or divergent sites overlapping with known TSS were collected as "full pausing" sites (group 4). Unknown transcripts were re-annotated, the potential genes overlapping with longer ones in the same direction were identified, and a known and unknown transcript category was formed.

3.5.4. The expressional analysis of nascent transcripts

All known transcript bodies were identified, and with respect to the ~45 nucleotide/s polymerase speed, the up to 50 kb 5' ends of these were used for gene expression analyses excluding the "alignment gaps", 3' overhangs and any divergent sites (TSSs and intronic enhancers) as these latter showed usually higher coverage than the other parts of the genes. The calculation of unique read number for each sample was done on those fragments, of which joint length was longer than 0.5 kb. Expressional values were fit to the input of the maSigPro program that had been initially developed for time-course microarray experiments (Conesa et al., Bioinformatics, 2006). Thus RPKM-like (~1/20 RPKM) gene expression values (reads per kb of transcript fragments per mapped read number of control) were

determined. maSigPro analysis was done by Attila Horváth. Java Treeview was used to visualize the expressional data of the different gene sets.

3.5.5. The annotation of RXR-bound regulatory regions

DiffBind v1.0.9 was applied to determine the significantly changing divergent sites upon 30-minute LG268 treatment by using two parallels with the same parameters as for the RXR peaks (described before). The RXR peaks that showed significantly regulated expression of divergent transcripts were annotated to the closest regulated nascent gene transcript within 0.5 Mb by PeakAnnotator (EMBL-EBI Bertone Group Software).

3.6. RNA-seq

Two biological replicates of BMDMs were treated with 100 nM LG268 in a 0, 30, 60 and 120-minute time series. RNA-seq libraries were prepared by using TruSeq RNA Sample Preparation Kit (Illumina) according to the manufacturer's protocol: 2.5 µg total RNA was used for the library preparation. Poly-A tailed RNA (mRNA) molecules were purified with poly-T oligo-attached magnetic beads. Following the purification, mRNA was fragmented using divalent cations at 85 ºC, and then first strand cDNA was generated using random primers and SuperScript II reverse transcriptase (Invitrogen, Life Technologies). This was followed by the second strand cDNA synthesis, then double stranded cDNA fragments went through an end repair process, the addition of a single 'A' nucleotide and then barcode indexed adapter ligation. Adapter-ligated products were enriched with adapter specific PCR to create the cDNA library. Agarose gel electrophoresis was performed on E-Gel EX 2% agarose gel (Invitrogen, Life Technologies) and the libraries were purified from the gel using QIAquick Gel Extraction Kit (Qiagen). Fragment size and molar concentration were checked on Agilent BioAnalyzer using DNA1000 chip (Agilent Technologies). Libraries were

sequenced with Illumina HiScanSQ sequencer. Experimental procedures were done by Bence Dániel, Szilárd Póliska and Tibor Gyuris.

3.7. RNA-seq analysis

TopHat (Trapnell et al., Bioinformatics, 2009) and Cufflinks (Trapnell et al., Nat. Biotechnol., 2010) toolkits were used for mapping spliced reads, making transcript assemblies and getting gene expression values in FPKM format (**Figure 7**). This analysis was done by Attila Horváth. Java Treeview was used to visualize the expressional data of the different gene sets.

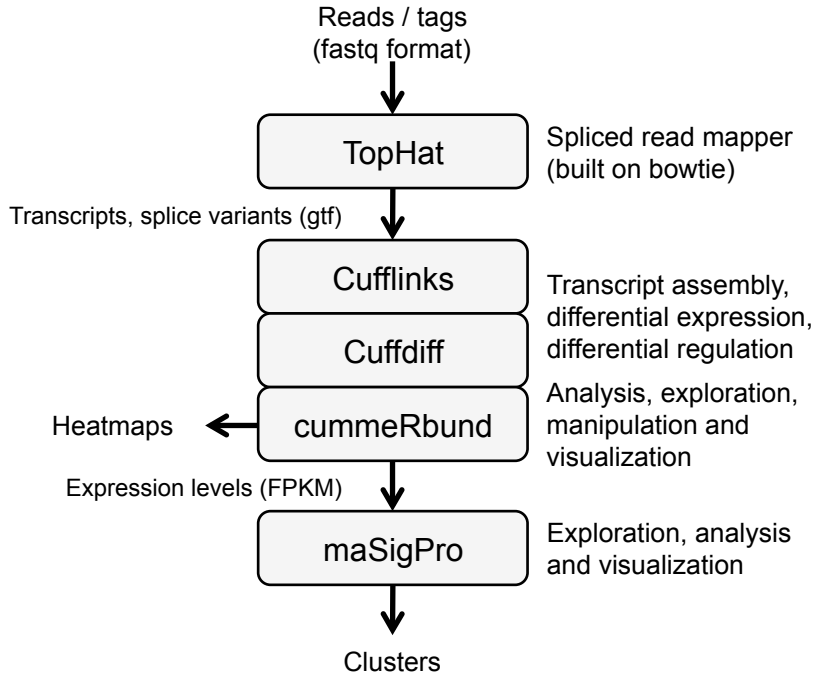


Figure 7. The bioinformatic pipeline used to analyze RNA-seq data

3.8. Domain predictions based on the CTCF and RAD21 “co-peaks”

Regions co-occupied both by CTCF and RAD21 proteins (co-peaks) having peak scores over 15 and the ratio of these less than 3 were considered as insulators. The closest insulators located within 1 Mb were assigned as putative borders of “functional domains” if the overall score of the individual co-peaks showed less than 5/3 fold difference between the pairs. The active domains located closer than 100 kb were united to major active topological domains, and the remaining regions between them were defined as inactive topological domains. The predicted domains were annotated to the regulated genes by using intersectBed (BEDTools). The differential binding analysis of RAD21 enrichments on the insulator (CTCF/RAD21 co-peaks) and the “active/passive” ($GRO^{+/-}$) RXR-bound regions was done similarly as described above.

3.9. Chromosome conformation capture (3C)

3C experiments were completed as described previously with minor modifications (Miele et al., Curr Protoc Mol Biol., 2006; Hagege et al., Nat Protoc., 2007). 1 million cells were fixed with 2% formaldehyde for 10 minutes. Nuclei were isolated in a buffer containing 10 mM Tris-HCl pH 7.5, 10 mM NaCl, 0.2% NP40 (Sigma) and protease inhibitor tablets (Roche). Chromatin was digested with 400U of HindIII (Fermentas) restriction enzyme at 37 ºC for 16 hours and for an additional 1 hour with 100U. Chromatin fragments were ligated with 100U of T4 DNA ligase (Fermentas) at 16 ºC for 4 hours. After ligation, chromatin was de-cross-linked overnight at 65 ºC. Ligation products were column purified (Roche, High Pure PCR Template Preparation Kit) and DNA concentration was determined by Nanodrop. Tandem primers were designed in the close proximity of the restriction enzyme cutting sites. Experimental procedures were done by Bence Dániel.

3.10. 3C-sequencing

Experiments were carried out as previously described (Stadhouders et al., Nat Protoc., 2013). After the first digestion and ligation, the 3C DNA pool was purified with phenol/chloroform/isoamyl alcohol (25:24:1) (Sigma). The second restriction digestion was performed by using DpnII (NEB) for 16 hours according to the manufacturer's instruction. The second ligation was performed at 16 ºC for 6 hours with 200U of T4 DNA ligase. DNA was then purified again with phenol/chloroform/isoamyl alcohol (25:24:1) followed by QIAquick gel purification column (Qiagen) purification. The bait specific inverse PCRs were performed using primers coupled to Universal Illumina adapters and barcode sequences. Reaction mixes were purified by QIAquick gel purification columns. Amplicon libraries were quantified and qualified by Agilent using DNA 7500 chip cartridge. Amplicon libraries were sequenced on Illumina MiSeq and HiSeq2000 sequencer. Two technical replicates of two biological replicates were sequenced. Experimental procedures were done by Bence Dániel.

3.11. 3C-seq analysis

Samples were de-multiplexed by FASTX tools based on their index sequence and then based on their bait sequence ending with (3') HindIII recognition site. BWA tools were used to align the remaining 68 to 82 nucleotide long fragments (starting with 5' HindIII site) onto the mm9 genome assembly. Mapped reads were counted in 500, 1000, 2000 and 4000 bp windows of the genome. Read numbers were normalized to 1000 reads, and finally visualized in BEDGRAPH format. For the frequency analysis of putative distal interactions, the target coverage of 1 Mb bins (covering the whole mouse genome) was determined as thousandths of all putative interactions. Bins of the inactive and active topological domains (predicted as described above) were discriminated, and then the RXR dependent and independent ones of

the active domains were determined. The distribution of interaction frequencies were plotted for Abcg1 and Vegfa enhancer (B1) baits.

3.12. The phylogenetic comparison of AP-1/CREB related bZIP proteins

44 AP-1/CREB related bZIP protein sequences were collected from the Ensembl database through BioMart. For phylogenetic analysis, the integrated tools of Molecular Evolutionary Genetics Analysis (MEGA 5.05) software were used (Tamura et al., Mol Biol Evol., 2011). The 56 amino acid long fragment of the basic domains was applied for multiple alignment by using "multiple sequence comparison by log- expectation" (MUSCLE) (Edgar, BMC Bioinformatics, 2004). Phylogenetic tree was created by neighbor-joining statistical method. Pairwise distances were estimated applying Poisson model and the resulting similarity matrix was sorted in command line according to the phylogenetic tree. Java Treeview was used to visualize the similarity matrix.

3.13. The description of the server used for the analyses

The ngsdeb.med.unideb.hu cluster is built up from 7 nodes each having 2x6 2.8GHz processors. The head node has 128GB memory and 8x600GB serial attached SCSI (SAS) disks, and the further nodes have 6x48GB memory and 6x48GB SAS disks. The server has further 12x12TB disks integrated by a redundant array of independent disks (RAID) system. Typically 6 processors and 30GB memory were sufficient for the analyses.

4. Results

4.1. Determining the putative regulatory regions of macrophages based on histone coverage information

Histone modification landscapes typically show low resolution, broad read enrichments, but in these regions, the narrower regulatory sites – promoters and enhancers – draw relatively sharp valleys called nucleosome free regions (NFRs). As we were interested in the regulatory network of BMDM cells, firstly, we gained information from this kind of data. As this approach can be considered as unusual, we developed a novel method for a broadly configurable NFR prediction from histone modification ChIP-seq data. TFs leading macrophage differentiation are quite well known, however the exact way and location of their binding and action still raise questions, such as 1) what size of the regulatory regions is nucleosome-depleted; 2) which co-binding TFs are included in a regulatory complex; and 3) which additional TFs play role in the steady-state macrophage function.

To address these questions, we performed ChIP-seq, two replicates for each histone modification: the best-known active enhancer marks (H4ac and H3K27ac), as well as the general enhancer and TSS marks H3K4me2 and me3, respectively (Kouzarides, Cell, 2007). We chose H3K4me2 instead of H3K4me1 because of its better cell-specificity and sharper appearance (Northrup and Zhao, Immunity, 2011). Firstly, we applied findPeaks, which includes the NFR predictor of HOMER (Heinz et al., Mol. Cell, 2010). Although the total number of mapped reads was significantly different between the H3K4me3 replicates, their IP efficiencies were sufficiently high to accurately predict NFRs. Differences in the number of predicted NFRs were mainly due to the variations in IP efficiencies. **Table 6** shows the total and effective read numbers of the samples determined by HOMER. Effective read numbers and IP efficiencies were calculated based on the enriched regions. We identified more than 22,000 active regulatory regions and a similar number of potential binding sites ("Total

regions" and "Total NFRs" in **Table 6**). However, these numbers are much lower than expected in macrophages, where approximately 45,000 binding sites were identified only for the lineage-determining PU.1 (Heinz et al., Mol. Cell, 2010). This implies that only a fraction of the enhancers could be found by applying this NFR prediction method, however, it worked well for the TSS specific histone mark. The union of the predicted regions from H3K4me3 replicates (14,199) overlaps with the TSSs of 11,889 genes, which matches well with the previous expectations for any cell types.

Sample / parallel		HOMER					PeakSplitter based prediction			
		Total reads (million)	Effective reads (million)	IP efficiency (%)	Total regions	Total NFRs	Total NORs	Total NFRs	Consensus NFRs	Union of NFRs
H4ac	1	39.662	6.800	17.14	26,237	23,641	148,725	83,981	49,048	121,330
	2	20.399	3.151	15.44	22,193	19,644	239,103	102,022		
H3K27ac	1	29.011	2.167	7.47	13,425	8,957	131,380	50,154	23,733	74,478
	2	44.698	5.845	13.08	24,223	21,610	115,701	56,432		
H3K4me2	1	12.684	1.609	12.69	22,368	19,581	275,138	112,148	49,754	155,605
	2	22.345	5.354	23.96	31,889	41,426	195,051	110,943		
H3K4me3	1	5.543	0.889	16.05	11,553	17,165	25,939	15,989	11,872	19,651
	2	9.109	2.148	23.59	13,858	18,945	27,066	16,818		

Table 6. The statistics of the histone modification ChIP-seq results

From left to right: General numbers of the region (findPeaks parameters: -size 1000 and -minDist 2500) and NFR prediction of HOMER (switch: -nfr) and our method.

To find as much NFRs as possible, PeakSplitter (EMBL-EBI Bertone Group Software) was used to determine the nucleosome occupied regions (NORs) for each modification. The thus identified NORs enabled the fine-tuning of NFR prediction and to detect longer NFRs as well, which were difficult to detect with HOMER. We predicted NFRs by applying two, largely independent methods, where the first one was based on the predicted

summits, and the second one used the "inner" edges of the raw NOR predictions. Then finally, the resulted regions were combined (**Figure 6**). To avoid the use of false hits, we extracted the overlapping regions of the two biological replicates that were longer than 100 bp and called them consensus NFRs. Although these regions formed a reasonably large set of potential enhancers (and promoters), for some parts of the analysis, all predicted NFRs were retained (see "Union of NFRs" column in **Table 6**; **Figure 8**).

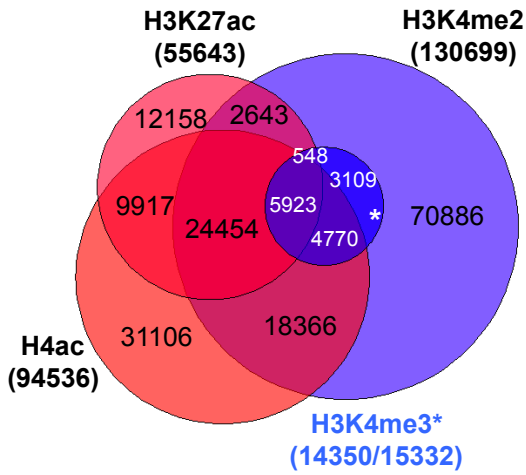


Figure 8. The overlap of the "union" NFR sets determined from H3K27ac, H4ac, H3K4me2 and H3K4me3 ChIP-seq data (982 regions surrounded by H3K4me3 signal could not be determined from the other histone modifications)

From our raw prediction, approximately four times more NFRs could be detected as compared to the results generated by HOMER (**Table 6**). The smaller number of promoter related NFRs was due to the frequently asymmetric nature so the lack of a typical valley-like pattern of the H3K4me3 histone mark. The lack of significant differences between the numbers of the predicted regions of the replicates was an indirect proof indicating the comparability of the data used for the analysis. The most frequent modification was the general enhancer mark H3K4me2 with more than 130,000 predicted regions. 93.6% of the NFRs determined from H3K4me3 ChIP-seq data overlapped with these, representing 11% of the dimethylated regions. On the other hand, more than 40,000 NFRs were marked both by H3K27ac and H4ac, representing 72.4 and 42.6% of the particular NFR sets, respectively. The remaining 27.6% of the NFRs identified only by H3K27 acetylation contained only a few

hundred of NFRs, which could be reproduced in both replicates, suggesting that the majority of these NFRs are likely to be poorly modified. The 56,704 NFRs determined both from acetylation and methylation bordered regions might mark probably the most active enhancer/ promoter regions of macrophages (**Figure 8**).

The width distribution of consensus NFRs revealed a one-nucleosome preference, suggesting mono-nucleosome release at these putative regulatory regions (**Figure 9A**). As the width of the NFRs increased, a consistent decrease could be observed in the number of the NFRs. The continuous NFR width distribution suggests that the complex of the co-binding TFs and co-factors is able to position and "roll" outwards the modified nucleosomes as proposed earlier (Zentner and Scacheri, J Biol Chem., 2012). H3K4me3 showed a different pattern: its preferred NFR length was 200-250 bp wide, and there were just minor protrusions in the NFR number at about 300 and 400 bp widths. This can be due to the PIC assembled on the promoter thus displacing the histone molecules from the DNA.

As of the three enhancer marks H3K4me2 gave the most intense signal, for the following analyses, we used the predicted regions derived from this data. By sorting the NFRs according to their width, it could be seen that both types of histone acetylation followed the pattern of H3K4me2. The bigger half of these sites showed histone release on 150 +/- 75 bp wide regions, which is the approximate length of the DNA double helix wrapped around one nucleosome core (**Figure 9B**) ("nucleosome 1" or "NS1" cluster). The about two-nucleosome wide NFRs (NS2) were also frequent, but the broader regions became more and more rare.

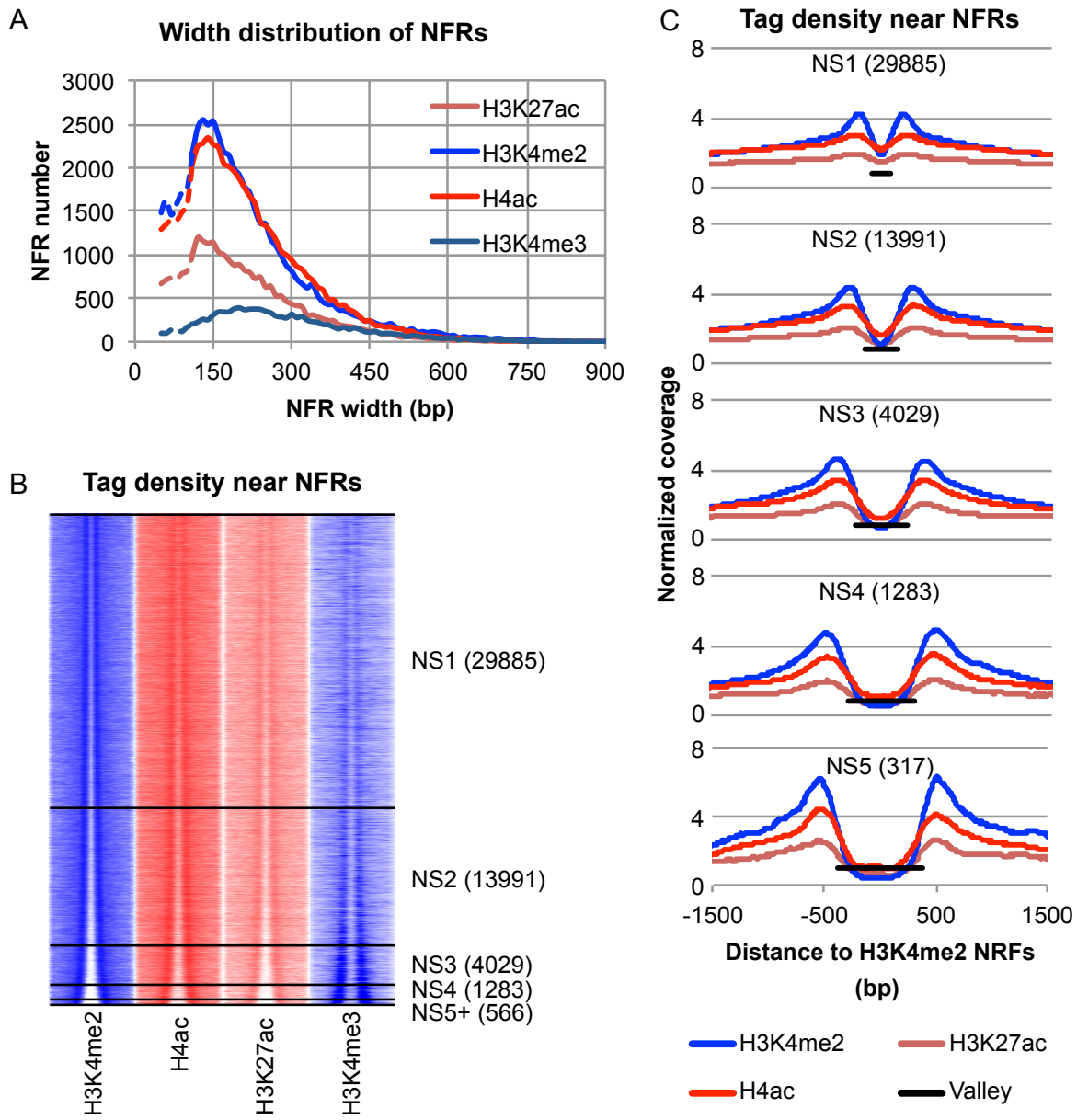


Figure 9. The comparison of the NFRs determined from different histone modifications

A) Width distribution of the "consensus" NFRs (lines are dashed below 100 bp width)

The read distribution heat map (B) and meta profile (C) of the different histone modifications relative to the middle of H3K4me2 derived "consensus" NFRs having at least 100 bp width (NS1-5 refer to the length of the DNA wound on 1 to 5 nucleosome(s) (150*n ± 75 bp), respectively; the number of the regions in the given length range are shown in parentheses)

The intensity of H3K4me3 highly correlated with the enhancer marks but showed a different pattern with a tighter and shallower valley in average. With regard to the width distribution shown in **Figure 9A**, this was mostly due to shifts compared to the middle of the regions used for this analysis. The broadest regions showed the highest coverage by trimethylated histones thus indicating probably the most active promoter regions.

We visualized the data of enhancer marks as read distribution histograms, where the "clusters" were defined also by the (150*n) ± 75 bp NFR widths of the H3K4me2 data (**Figure 9C**). This showed an exact overlap in total between the H4 and H3K27 acetylated histones, but the valleys of H3K4me2 were slightly broader. H4ac represents four lysine modifications, which can partly explain why this protein gives higher IP efficiency (**Table 6**) and therefore a higher "meta-profile" as compared to the other histone acetylation. These analyses reaffirmed that co-activator enzymes on enhancers – similarly, as happens on the promoter regions – are likely to modify the neighboring nucleosomes and reorganize the DNA–histone complexes. The used parameters and thresholds thus seemed applicable, but to further validate our approach, we tried to identify the TFs responsible for these phenomena.

4.2. Determining macrophage specific transcription factors from NFR predictions

According to the ENCODE ChIP-seq analysis guideline, the suggested approach to search for *de novo* motif enrichments is the use of the summit ±50 bp regions of the top 500 peaks as defined by the peak scores (Wang et al., Genome Res., 2012). As we used an indirect binding site detection method and given that HOMER masks out repeat-regions, we used the middle 200 nucleotides of the top 5000 consensus NFRs for each histone modification. The motifs of macrophage determining TFs – PU.1, AP-1 and C/EBP – were enriched from both the H4 and H3K27 acetylation results, while the IRF, CREB/ATF and RUNX motifs were enriched only from the H4ac data (**Figure 10A**).

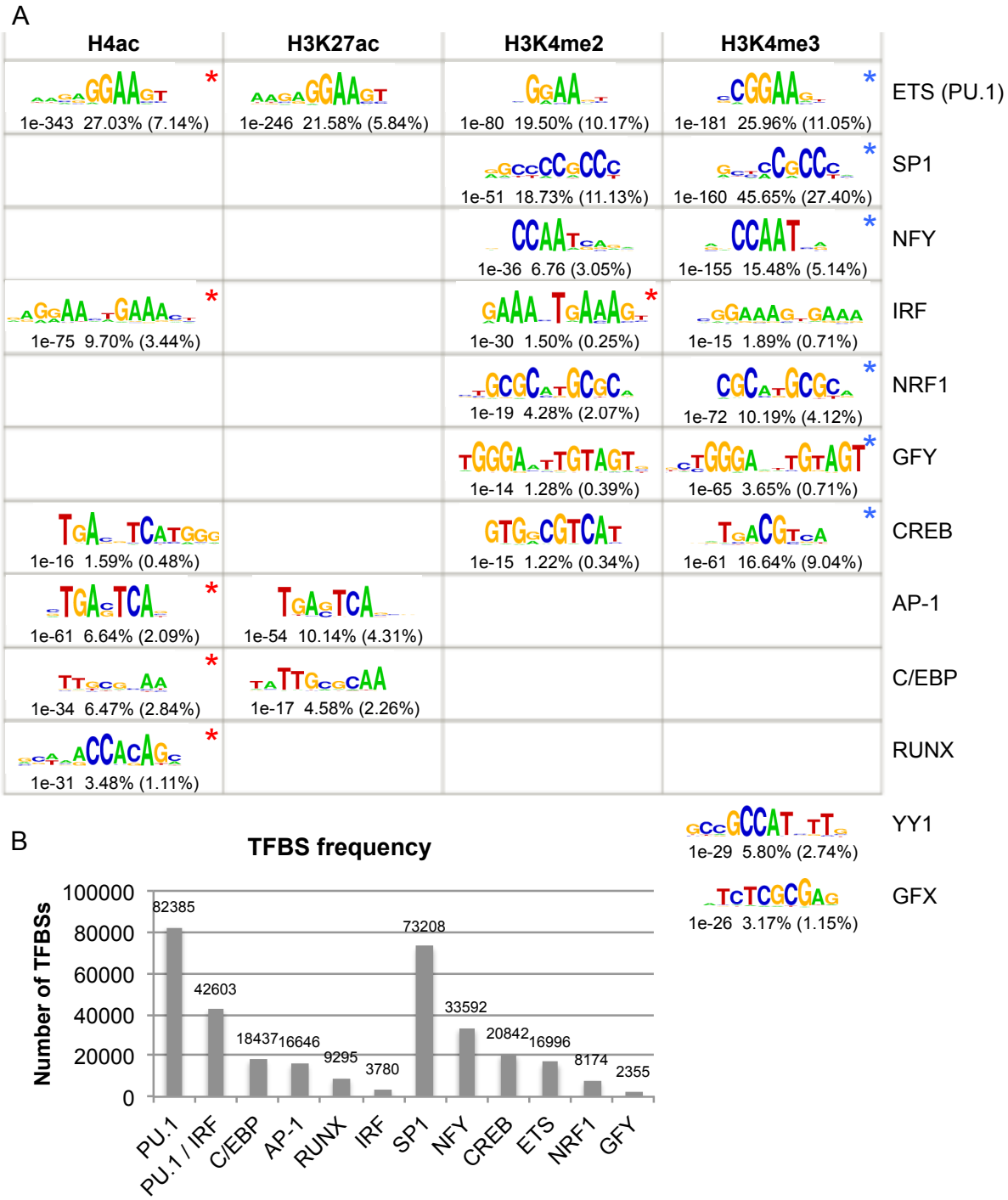


Figure 10. The comparison of the motifs determined from the promoter and enhancer specific NFRs

A) The enrichment of *de novo* motifs determined from the different kinds of NFRs (p-value, target and background enrichments for each motif are shown, respectively)

Motifs were chosen based on their p-value (<1e-10) and fold enrichment (>1.5) compared to the random background. The top motifs specific for enhancer and promoter regions are marked with red and blue asterisks, respectively.

The further part of **Figure 10** *description:*

B) The number of the putative TFBSs mapped onto the united set of the different kinds of NFRs based on the motif matrices marked by asterisks in A)

H3K4me3 marked regions showed the well-known promoter specific enrichments: the GC-box (bound by the SP1 family) and CAAT-box (bound by NFY), as well as the motif of the ETS superfamily, NRF1, GFY, YY1, GFX, CREB/ATF and IRF. Top H3K4me2 NFRs showed the same motifs enriched (ETS, SP1, NFY, IRF, NRF1, GFY and CREB/ATF) with higher p-value (except for ISRE) indicating again that the most intense dimethylation show high correlation with the trimethylation of H3K4. Excepting ISRE, NFRs derived from H3K4me2 data – similarly to the results of H3K4me3 – did not show macrophage specific motif enrichments. The EBS core motif was another exception, which probably was the result of a fusion of the PU-box (gaGGAAgt) and the promoter-specific ETS motif (ccGGAAgt) (**Figure 10A)**.

By using the *de novo* motif matrices, we extracted the putative TFBSs from the predicted NFRs (**Table 7**). The number of these binding sites was comparable with those published in the literature (**Figure 10B**) (e.g. Heinz et al., Mol. Cell, 2010; Ostuni et al., Cell, 2013). The most frequent motifs unsurprisingly belonged to the PU.1 and IRF proteins, while the second most abundant motif (GC-box) was found more than 70,000 times (however it is a less complex element with numerous matches in the random background). The promoter specific CAAT-box, CRE and ETS element showed significant enrichments with 33,592, 20,842 and 16,996 putative binding sites, respectively, however this CRE motif matrix matched frequently also to the background. The number of the predicted C/EBP binding sites was about half of that previously found in C/EBP alpha and beta ChIP-seq (Heinz et al., Mol. Cell, 2010). This was probably due to our motif of which matrix was weaker than the ones

derived from e.g. C/EBP ChIP-seq data. AP-1 and RUNX sites were less abundant, but their number was still significant.

These results demonstrated that based on the genome coverage data of modified histones, our NFR prediction method was suitable to determine the active regulatory regions together with the key TF families acting at these sites. We could predict almost 30,000 putative active enhancers with macrophage specific elements (the NFRs of H3K4me2 and H4ac and/or H3K27ac, without H3K4me3) and more than 11,000 promoters with the well-known TSS/TSR specific motifs (**Table 6**; **Figure 8**).

Motif name	Consensus sequence	Motif length	Score threshold	Number of hits
EBS (ETS)	ARCC**GGAA**GT	4 + 2 x 2	7	16,996
PU-box (PU.1)	AAGA**GGAA**GT	4 + 2 x 2	7	82,385
EICE (PU.1-IRF)	GA**GGAA**CT**GAAA**CT	2 x 4 + 2	8	42,603
ISRE (IRF)	**GAAA**CT**GAAA**GT	2 x 4 + 2	10	3,780
C/EBP	BV**TT**GCGC**AA**	2 x 2 + 4	5.9	18,437
RUNX	GCWA**ACCACA**GC	6	9	9,295
TRE (AP-1)	NKS**TGA**S**TCA**SN	2 x 3 + 1	6.5	16,646
CRE (CREB)	GV**TGA**CG**TCA**	2 x 3 + 2	5.5	20,842
GC-box (SP1)	**GCYCCGCCCH**	2 x 5	6.5	73,208
NFY	YRR**CCAAT**CR	5	7	33,592
NRF1	c**CGCATGCGC**A	2 x 5	8.5	8,174
GFY	TK**CTGGGA**RT**TGTAGT**	2 x 6 + 2	10	2,355

Table 7. The *de novo* motif enrichments of murine macrophages determined from NFRs

The last column shows the number of the motifs mapped on the union of all predicted regions.

4.3. The examination of histone patterns near PU.1 binding sites

After demonstrating that NFRs can assign the macrophage and promoter specific TFs, we tried to validate the predicted binding sites by comparing these results with those found by the analysis of a high-quality PU.1 ChIP-seq data (SRA accession number: SRR042037). This sample was prepared similarly to ours, with the difference that macrophages were processed after 6-8 days of differentiation and at least a few times more cells were used for the IP, which affected positively to the IP efficiency (Heinz et al., Mol. Cell, 2010). Coverage data was processed similarly to the "NOR prediction", but peaks were limited in ±150 nucleotides relative their summits. Putative artifacts defined by the enhancer mark "enrichments" over score 50 were removed from the raw PU.1 peak set. Thus, more than 70,000 PU.1 bound regions could be detected (**Figure 11A**) showing an extremely high IP efficiency (32%). PU.1 motif enrichment also showed an extremely strong p-value based on the 100 bp region around the summit of the top 1000 PU.1 peaks (**Figure 11B**).

Approximately half of the PU.1 peaks overlapped with NFRs determined from the H4ac and/or H3K4me2 modifications (**Figure 11A**), and surprisingly, a significant fraction of PU.1 peaks overlapped with NORs or potential heterochromatic regions (**Figures 11C-D**). On the read distribution histograms of the four PU.1 peak groups there was a high correlation between the intensity of histone modification and PU.1 enrichment (**Figure 11D**). Consensus NFRs marked the highest PU.1 peaks, while the less active regions showed the smaller peaks. These binding sites were either occupied by lower frequency or in a smaller fraction of cells. Direct PU.1 binding to the DNA–nucleosome complex is a known event (Ghisletti et al., Immunity, 2010; Heinz et al., Mol. Cell, 2010), and we provided further evidence that PU.1 is able to bind those regions, which are theoretically wrapped into nucleosomes.

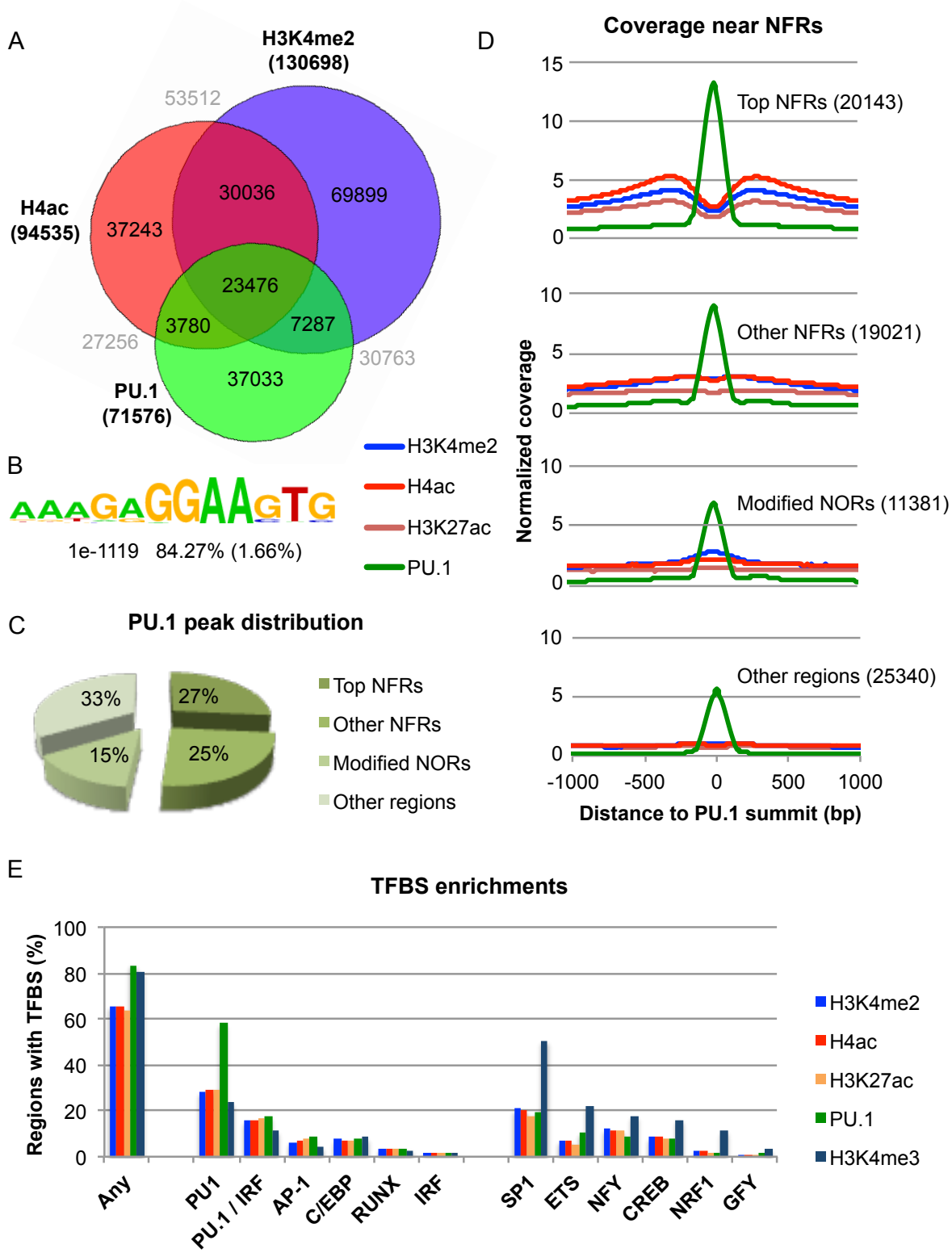


Figure 11. PU.1 binds to compacted DNA

A) The overlap of PU.1 peak summits with the NFRs derived from H4ac and H3K4me2 ChIP-seq data ("union" set)

The further part of **Figure 11** *description:*

B) The PU-box motif enriched under the top PU.1 peaks (p-value, target and background enrichments are shown, respectively)

C) The distribution of PU.1 peaks between different genomic regions: "Top NFR" means any "consensus" NFR of the enhancer specific histone marks; "Other NFR" means any other region of the "union" set of NFRs; "Modified NOR" means any predicted NOR enrichment.

D) The meta profile of PU.1 and the enhancer specific histone modifications relative to the PU.1 peak summits of the same peak groups as shown in C)

E) The ratio of NFRs ("union") and PU.1 peaks having any or the given TFBS (for the total number of motifs mapped see Table 7)

Next, we examined the motif specificities of the regions predicted from the different histone modification data (**Figure 11E**). The determined binding sites covered about 80% of the NFRs of H3K4me3, 83.7% of the PU.1 peaks, as well as about 60% of the regions with enhancer mark. As expected, PU-box was the most enriched within the PU.1 peaks and the least specific within the promoter regions. PU.1-IRF, AP-1 and RUNX elements were about equally specific for the different enhancer regions. C/EBP and IRF elements did not show enhancer specificity: these were present in a comparable amount in promoter regions also. Promoter associated elements showed the expected enrichments on H3K4me3 signed NFRs, while PU.1 peaks showed the least specificity on these, except for the EBS, suggesting that PU.1 is able to bind to the ccGGAAgt sequence, as well.

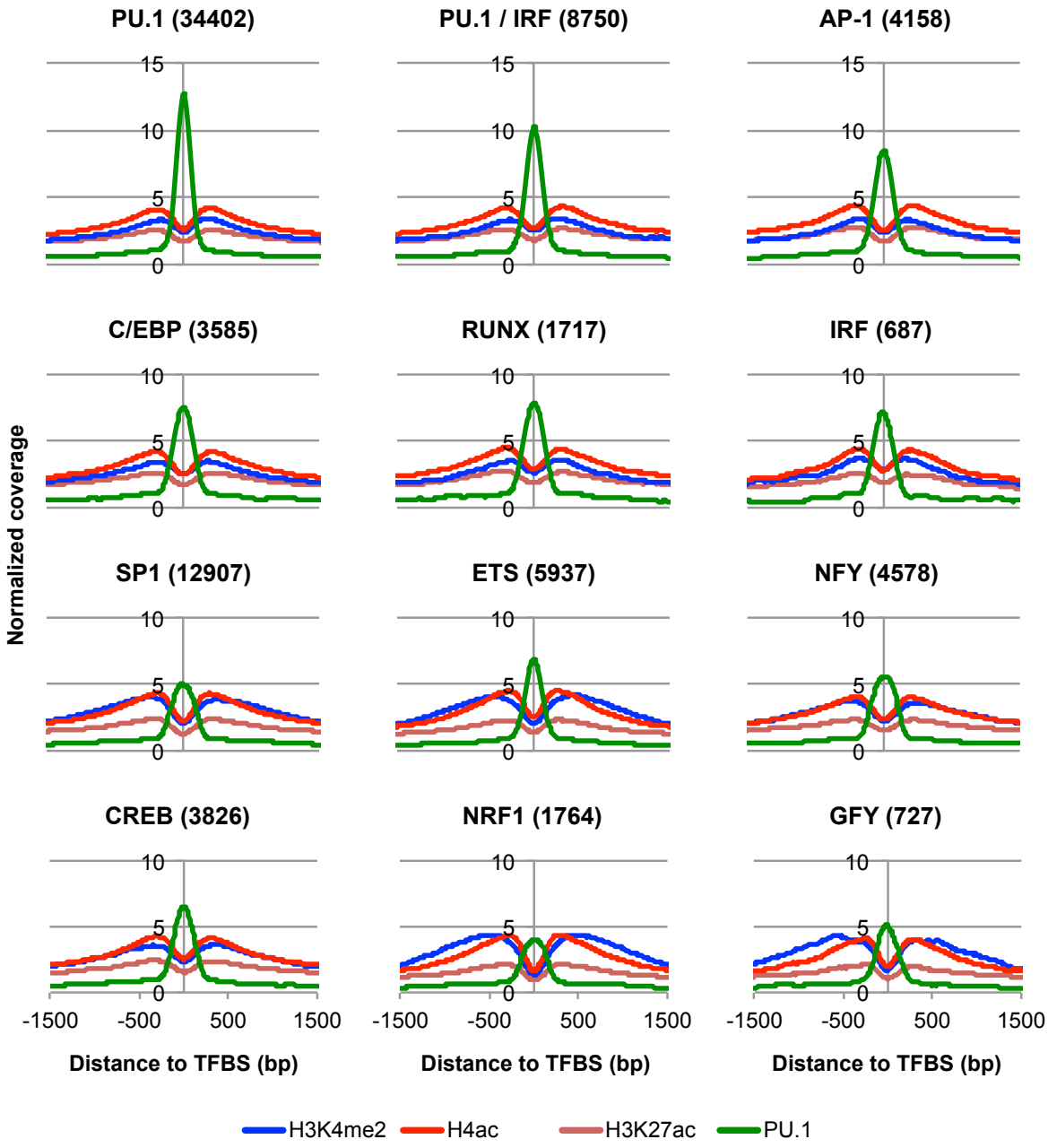


Figure 12. The characterization of the binding and co-binding events of PU.1

The meta profile of PU.1 and the enhancer specific histone modifications relative to the middle of the putative TFBSs located under PU.1 peaks (the motif matrices of **Figure 10A** were used to map the TFBSs; the number of mapped TFBSs is shown in parentheses following the name of the given motif)

To test the binding and co-binding events of PU.1, we investigated its enrichment near the mapped motifs overlapping with PU.1 peaks (**Figure 12**). The shape of the "peaks" drawn by histograms shows the relationship of the co-operating TFs on the DNA: sharper peaks indicate higher PU.1 coverage on the given sites, while broader peaks show larger distance between the co-bound sites. Beside its own (PU.1 and PU.1-IRF) elements, PU.1 frequently bound DNA in the close proximity of AP-1, C/EBP, RUNX and IRF elements. At these sites, H4ac and H3K27ac moved together, and compared to them, H3K4me2 modification was slightly closer to the PU.1 bound sites in average. Near promoter specific elements, we found a similar shape of acetylation as observed at enhancer specific regions; however the read distribution of H3K4me2 was both wider and deeper. Based on this data, EBS was likely to be bound directly by PU.1, however with lower frequency as compared to its specific binding sites. The elements of the CREB and GFY were closer, while those of SP1, NFY and NRF1 were farther from the ETS-boxes. Interestingly, NRF1, GFY, ETS and SP1 elements were best characterized by those histones dimethylated significantly farther outwards from the acetylated ones. These CpG-rich promoter specific elements may be responsible for the broader NFRs (derived also from H3K4me3 data), which indicates that TSR binding proteins indeed supersede several nucleosomes (**Figure 9**).

4.4. Combining NFR data to get further putative regulators

To this point, histone modification results were analyzed in parallel in order to check whether the predicted NFRs were appropriate to identify the enriched motifs and to reveal the potential binding TFs. Then, with the combination of different histone modifications, we wanted to improve the location of the predicted NFRs and thus get the motifs of the less known TFs beside the principal ones (**Figure 13**).

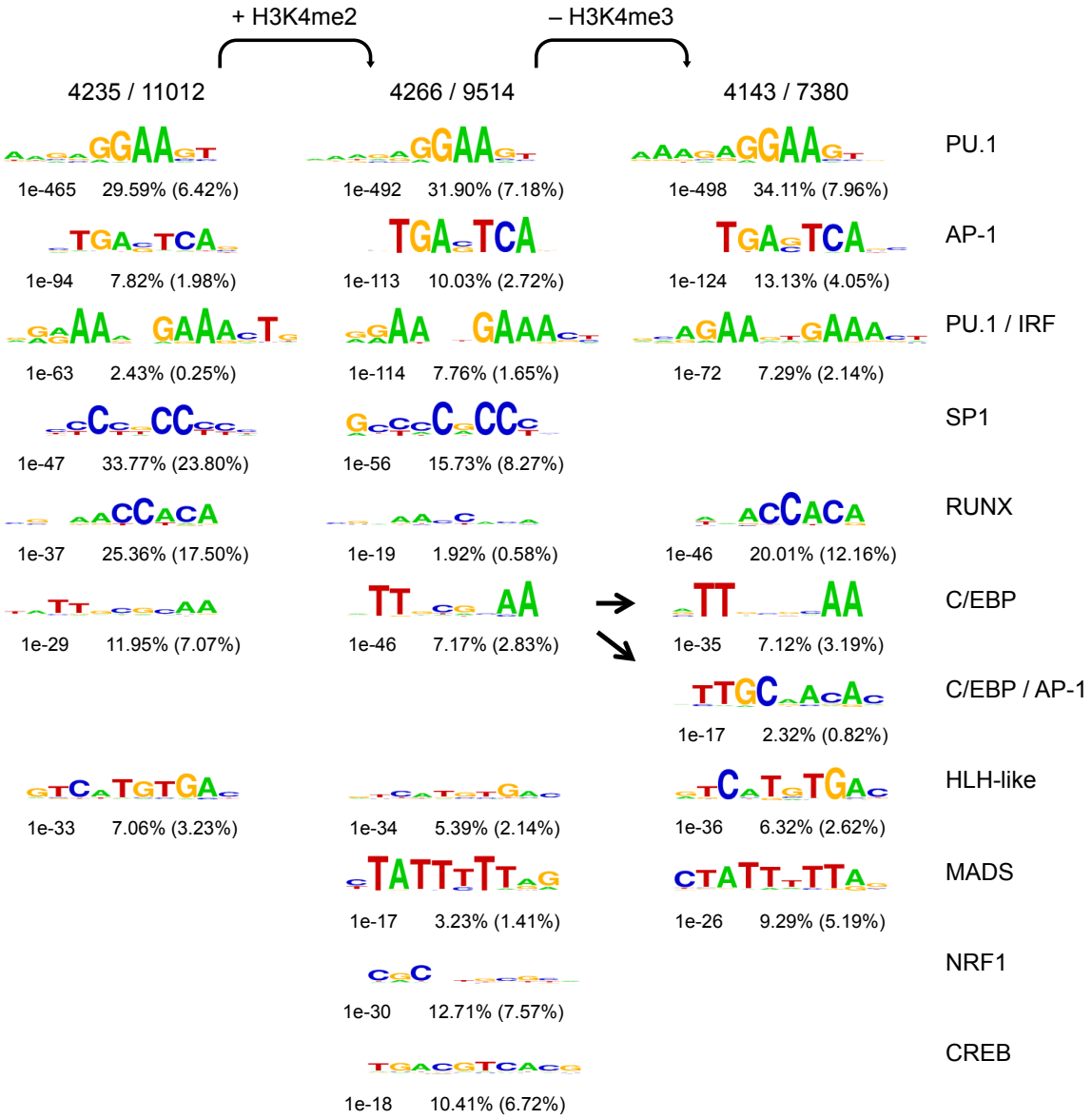


Figure 13. The combination of different histone marks improves the macrophage specific motif enrichments

The longer than 100 bp overlapping regions of the H3K27ac and H4ac derived consensus NFRs (left), the regions (of these) overlapping with any H3K4me2 derived NFRs (middle) and the regions (of the latter set) lacking any H3K4me3 signal (right) were used for the motif enrichment analyses. On the top, the number of the unmasked NFRs (of 5,000 regions) used for the analysis and the size of the whole NFR set are shown. Motifs were chosen based on their p-value (<1e-10). (p-value, target and background enrichments for each motif are shown, respectively)

The use of the even stringent consensus NFR set derived from both histone acetylation data improved the enrichment of the previously detected motifs, gave an SP1 motif, and an additional, HLH motif like enrichment appeared (the left column of **Figure 13**). Examination of those regions marked also with H3K4me2 further improved the motif enrichments, and newer motifs emerged: the promoter specific NRF1 and CREB motifs and those specific for the MADS protein family (the middle column of **Figure 13**). Finally, by removing the H3K4me3 derived NFRs, the promoter specific motifs were efficiently suppressed, and a clear C/EBP motif could be distinguished from the C/EBP-AP-1 motif (the right column of **Figure 13**).

Based on sequence similarity, the enriched "HLH-like" motif tagged by HOMER (CACGTG) belonged to the M-box (TCANNTGA) specific for the microphthalmia transcription factor/transcription factor E (MITF/TFE or MiT) protein group. MiT proteins (together with the MAD/MAX proteins) are members of the bHLH-ZIP family that form dimers only with MiT proteins and are responsible for phagocytic activity. They are key TFs in osteoclast function (Walsh et al., Gene, 2003; Karlström et al., Exp Hematol., 2011), and indeed it has been shown that the lack of their DNA-binding caused osteopetrosis (Meadows et al., J Biol Chem., 2007).

The "maintenance of minichromosome 1 (MCM1, yeast), AGAMOUS (plant), DEFICIENS (plant), serum response factor (SRF)" (MADS) family has an ancient DBD, which binds the CArG-box. Beside this DBD, myocyte enhancer factor 2 (MEF2) A-D proteins have an additional MEF2 domain, which is also responsible for DNA-binding, dimerization and protein interactions. MEF2 proteins indeed have roles in monocyte/ macrophage differentiation. MEF2D together with SP1 and C/EBP activates CD14 expression during monocytic cell differentiation (Park et al., Mol Immunol., 2002), and the MEF2A/D heterodimer activates JUN expression, probably by changing HDACs to P300

acyltransferases, while other genes are repressed by MEF2A/D-HDAC1/7 co-regulator complexes during macrophage differentiation (Aude-Garcia et al., Biochem J., 2010).

4.5. The examination of the RXR cistrome in macrophages

Beside the numerous other tissue specific environmental signals, macrophages are also modulated by lipid molecules such as fat-soluble vitamins and hormones, which regulate gene expression through nuclear receptors (NRs) (Nagy et al., Physiol Rev., 2012). The best-known NRs in macrophages are the RXRs and their heterodimerizing partners, PPARs, LXRs and RARs. We were curious of the cistrome of these NRs, but in the lack of good antibodies or appropriate protocols/buffers for the ChIP-seq, we chose an antibody against RXR, as it theoretically covered most of the PPAR, LXR and RAR TFBSs, and had been working several times in other laboratories (e.g. Nielsen et al., Genes Dev., 2008; Lefterova et al., Genes Dev., 2008; Adhikary et al., PLoS One, 2011).

We accomplished ChIP-seq for RXR from mouse BMDMs in the absence or presence of the RXR agonist LG268 and determined the read enrichments (peaks) as described in **Figure 5**. As based on the coverage files RXR binding showed a global strengthening upon LG268 treatment, the predicted raw peak numbers seemed to be misleading (**Figure 14A and G**). This was why we executed a statistical approach, which evidenced a significant increase in DNA occupancy by RXR at 730 regions, while only 83 showed significant decrease (**Figure 14B**). With this method we got a reasonably sized consensus peak set of 5,206 peaks, which could be used for the following analyses (**Table 8**, **Figure 14C**).

ChIP-seq was also carried out for the DC, macrophage and B-cell specific pioneer and lineage determining factor PU.1 (Guerriero et al., Blood, 2000; Heinz et al., Mol. Cell, 2010; Di Tullio et al., Proc Natl Acad Sci U S A., 2011), the active enhancer mark P300 (Bedford et al., Epigenetics, 2010) and the active promoter specific H3K4me3 (**Table 8**).

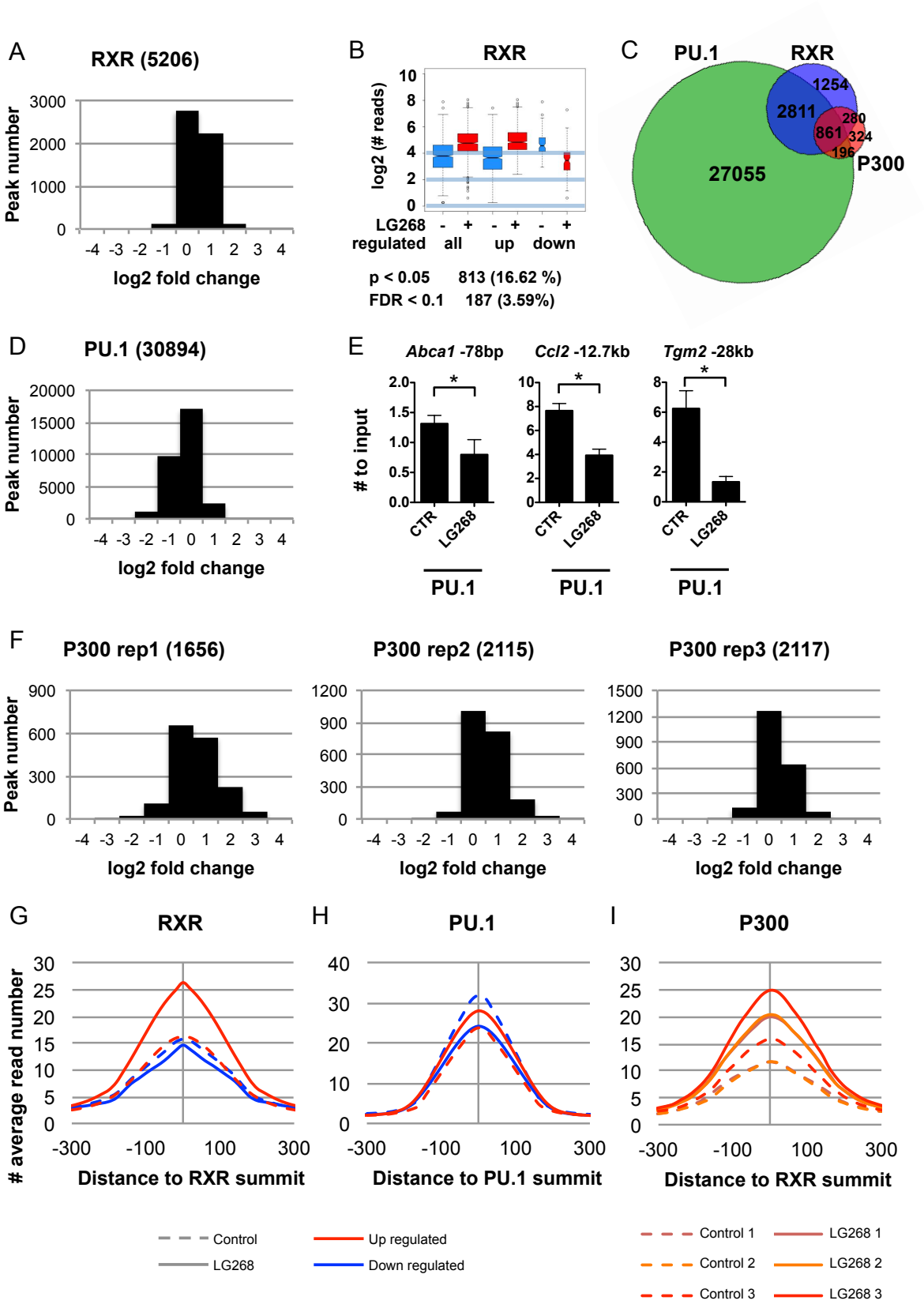


Figure 14. The effect of LG268 treatment on different macrophage cistromes

The description of **Figure 14***:*

The distribution of fold changes upon LG268 treatment in the RXR (A), PU.1 (D) and P300 cistromes (F) (the total number of peaks is shown in parentheses following the name of the given protein; P300 is represented with 3 replicates)

B) Read enrichment of the significantly changing RXR peaks in the presence of LG268 (the number of changing peaks and the statistical stringency applied are indicated below)

C) The overlap of the PU.1, RXR and P300 cistromes

E) The change in PU.1 binding on the indicated individual enhancers using ChIP-QPCR (the mean and SD of three biological replicates are shown; asterisk represents significant difference at $p < 0.05$; n = 3; # symbolizes "normalized")

The meta profile of the increscent and decrescent RXR (G) and PU.1 peaks (H) and three parallels of the increscent P300 peaks (I), which were up regulated in each parallel (# symbolizes "normalized")

Antibody	Peak numbers		
	Control	LG268	Union
PU.1	27,333	23,632	30,923
P300	626	1316	1661
RXR*	5,124	5,185	5,206
H3K4me3	15,788	16,583	17,018
CTCF	25,832	28,697	30,290
RAD21	16,514	20,327	24,548

Table 8. The number of peaks determined for the listed TFs and histone modification

Asterisk indicates that numbers are obtained with DiffBind.

PU.1, by nature, occupied more than 30,000 sites, with a bit lower coverage in the LG268-treated cells than in control macrophages. The overall decrease was confirmed by an occupancy analysis, which resulted the following tendencies: At most PU.1 bound sites, a moderated up-regulation could be observed upon treatment, while quarter of the peaks showed a stronger decrease that shifted the results to down-regulation with respect to the

entire peak set (**Figure 14D and H**). At some regions, ChIP-QPCR experiments were also carried out to validate this effect of RXR activation (**Figure 14E**).

P300 as a co-regulator gave less peaks than the previous TFs probably because of its indirect DNA binding or technical issues. Despite this, upon LG268 treatment, it clearly showed a global increase in DNA occupancy. Although we could not evince statistically significant increments, upon treatment we found 406 peaks of 1661 (24.4%) with higher coverage in all the three replicates produced (**Figure 14F and I**). The P300 bound sites in most cases overlapped with the RXR and/or PU.1 peaks (**Figure 14C**). This and their common enrichment upon ligandation suggested that P300 functioned as a co-activator highly related to RXR in this system.

H3K4me3 "peak" numbers were also included in **Table 8** because this histone modification usually showed peak-like formations in the beginning of the active genes. The predicted "peak" numbers of H3K4me3 were getting higher upon LG268 treatment, which might refer to the direction of gene expressional changes.

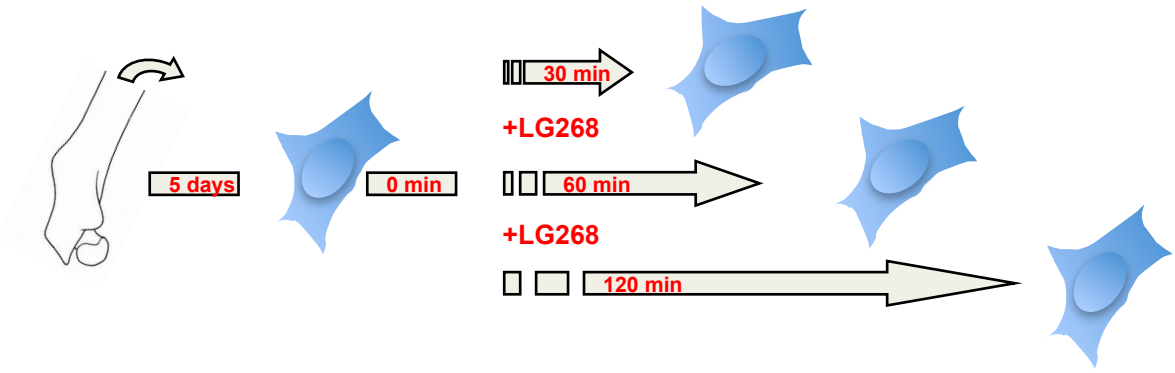


Figure 15. The scheme of macrophage differentiation and time course treatment with the RXR agonist LG268

4.6. The determination of the nascent transcriptome of macrophages

As we wanted to detect the direct effects of RXR activation on gene expression, GRO-seq were carried out in the 0, 30, 60 and 120 minute time points during LG268 treatment (**Figure 15**). This method draws all nascent transcripts genome-wide, including all the expressed mRNA, tRNA, rRNA and other non-coding (nc) RNA molecules, and the expressional changes can be seen nearly immediately after the treatment (Core et al., Science, 2008; Hah et al., Cell, 2011). The resulted coverage landscape is similar to that of a high-quality polimerase ChIP-seq, but this is strand specific, which is very important in the case of simultaneous transcriptional events occurring on both strands. GRO-seq thus gave several novelties about gene expression:

1) Typically, an initiation "peak" in sense direction and at least a short divergent transcript on the other strand could be seen on the promoter(s) of the active genes. 2) Gene body, in most cases, had lower expression compared to the initiation peak; (fold) difference between them is called pausing index. 3) Divergent transcript could be also elongated, and this longer transcript could be even another protein coding gene. 4) Protein coding genes were always transcribed further than their transcription termination sites (TTS), but the relevance or possible function of this 3' overhang is not yet known. 5) Longer non-coding transcripts, by moving to the 3' end, could be characterized by a gradually decreasing expression, which was a similar decrease to those of the 3' overhangs.

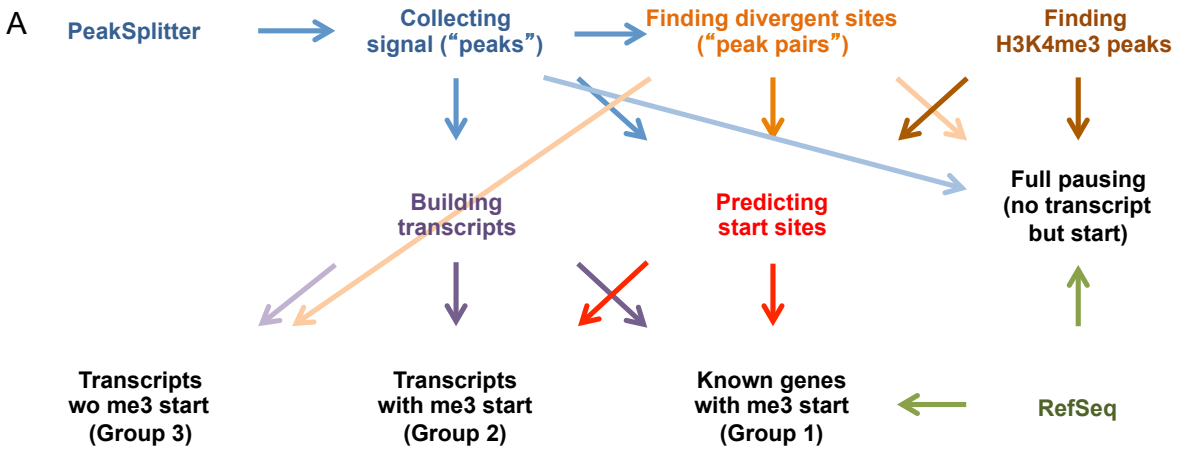
6) We and others (Hah et al., Genome Res., 2013) discovered that divergent transcripts were not only specific for the promoters but also for the active enhancers as they overlapped with the ChIP-seq peaks and histone valleys. This was a striking phenomenon, which raised enhancer transcription to the genome-wide level. GRO-seq made possible to examine this kind of transcripts, but this feature caused also technical problems. Based on the coverage information determined from GRO-seq data, it was often hard to distinguish promoters from

the promoter proximal enhancers. In the case of a flawless genome annotation, this should not be a problem, but there are numerous unknown transcripts that may have several alternative promoters.

This was why we decided to predict all divergent transcripts – including every active regulatory site regardless of elongation –, and then to use the coverage data of H3K4me3 to separate promoters from enhancers. This modification seemed suitable to perform an unbiased promoter mapping, independently of the mouse reference annotation. As both the hidden Markov model based transcript prediction approach (Hah et al., Cell, 2011) and the one of HOMER largely concentrated on elongated transcripts on the dominantly expressed strand, and thus the resolution of the prediction was poor on the other strand, we developed a novel method using PeakSplitter, as well (EMBL-EBI Bertone Group Software). Getting strand specific "subpeaks" gave the possibility to determine short divergent transcripts and elongated ones with a very high resolution on both strands, and pooling all reads from the samples helped to detect the very low expressed transcripts.

Finally, as described in the Materials and Methods section and shown in **Figure 16**, we predicted 10,586 genes with pausing, H3K4me3 mark and accurate annotation (group 1), 4,601 putative genes with H3K4me3 mark and yet unknown TSS (group 2), 8,869 other transcripts without H3K4me3 mark (group 3) and 1504 very low expressed genes showing only weak transcriptional initiation signal ("full pausing", group 4). Both group 2 and group 3 included such transcripts, which showed high overlap with the 3' end of known genes in the sense direction, so these could be annotated as new transcript variants of these genes. Together we could predict 11,235 known genes and 12,821 yet unknown transcripts including elongated enhancer transcripts and other long ncRNAs.

A



B

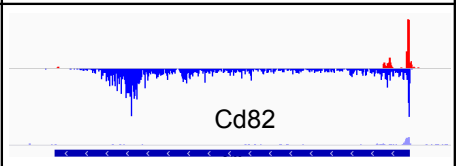
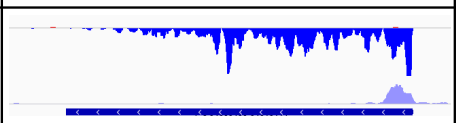
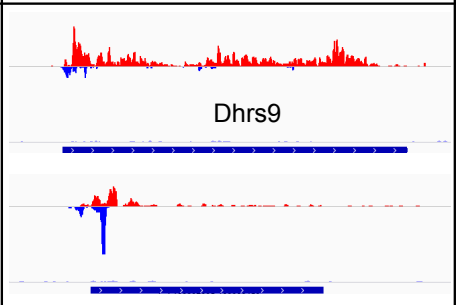
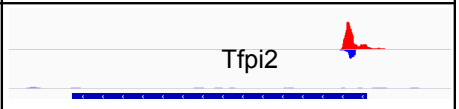
Description	Number	Pausing	H3K4me3	RefSeq	Transcript
Group 1. Known genes with Me3	10586	+	+	+	
Group 2. Unknown transcripts* with pausing	4601	+	+	?	
Group 3. Other transcripts* (>1000 bp)	8869	?	?	?	
Group 4. Full pausing	1504	?	?	+	
All	25560				*coding and non-coding RNAs, e. g. e-transcripts

Figure 16. The prediction and annotation of nascent transcripts using GRO-seq data

A) The algorithm of transcript calling and annotation during the analysis

B) The characteristic features of the predicted transcript categories

On the right, representative transcripts are shown strand specifically by dark blue boxes, and red and blue columns represent the coverage of GRO-seq sequence reads on the positive and negative strand, respectively. Light purple columns represent the H3K4me3 enrichment.

As the polymerase velocity was about 45 b/s (~160 kb/h), and we wanted to determine the directly regulated genes 30 minutes after the treatment, we could use the up to 80 kb long 5' fragments of the transcripts to measure the nascent gene expressional levels. To avoid biases caused by faster transcriptional events, we used the 50 kb fragments of the longer transcripts, and also excluded the 3' overhangs and divergent sites from read counting because these showed different, usually higher coverage than the gene body. To follow gene expressional changes, we calculated an RPKM-like unit from the fragmented transcripts and got 318 up and 423 down regulated genes upon LG268 treatment ($P<0.05$). The average fold difference of the activated genes was 1.81, while the one of the significantly repressed genes was 0.71, which indicated that the induced genes changed much more than the repressed genes.

4.7. The examination of the gene expressional changes of TFs upon RXR activation

The data derived from GRO-seq made possible to determine the immediate expressional changes genome-wide, but in the same time, it gave the unbiased transcriptional frequency of each RNA coding region in macrophages. To follow the changes on the matured mRNA level, we performed RNA-seq upon LG268 treatment for the same time points as for GRO-seq (0, 30, 60 and 120 minutes). These data together allowed observing the consequences of the decay and splicing processes of each RNA product. **Figures 17-20, 22, 24-25 and 29** including nascent RNA levels as heat maps contain some grey boxes also, which indicate the undetermined values. This was usually due to the undetectably low expression level in every time point, but in some cases, it was caused by technical issues if the relatively short genes were totally covered by divergent transcripts. The level of matured mRNA helps to answer, which was the reason.

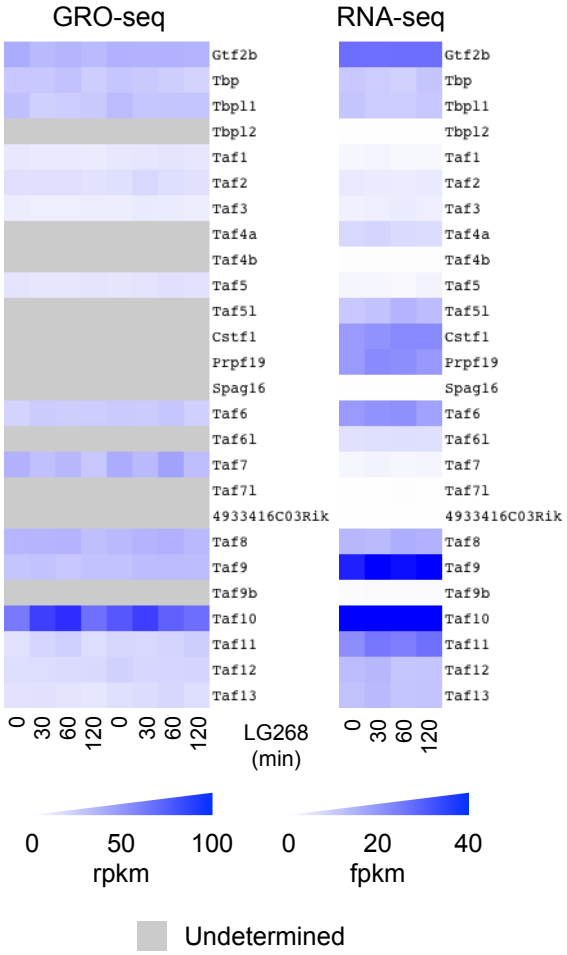


Figure 17. The gene expression changes of TFIIB and D components upon LG268 treatment The level of nascent (GRO-seq) and matured mRNA (RNA-seq) was determined as RPKM and FPKM values, respectively. (Expression values of the two replicates were combined by the RNA-seq analysis pipeline)

Firstly, we compared the expression levels of the classical general TFs (**Figure 17**) and found that their expression was not stoichiometric. This could be due to the different stability of the mRNAs or the different ratio of their protein products in the distinct TFII complexes. Taf10 and Gtf2b showed the highest expression both on the nascent and matured RNA level, while interestingly, Taf1 and Taf7 genes showed the lowest mRNA level. TAF1 and TAF7 proteins are in tight connection (Wang et al., Cell Res., 2014), but our data has shown that their precursors also had similar mRNA levels, however Taf1 had a consistent, while Taf7 had an oscillating expression. This can be explained probably by their size because TAF1 is the largest member of the TFIID complex, and the gene itself is 18.6-times longer than Taf7. Large molecules such as Taf1 derivatives need more energy for the

reproduction, thus probably need to be more stable than the smaller gene products. TAF7 is the repressor of TAF1 as a HAT, which suggests an individual regulation that may be responsible for the detected waves. From TAF1 to 10, the protein size is getting smaller and the RNA level (except for Taf7) was getting higher; this correlation may be indeed related to frugality. Tbpl2, Taf4b, Spag16, Taf7l and 4933416C03Rik, and Taf9b were the non-expressed paralogues of Tbp, Taf4a, Taf5, Taf7 and Taf9 genes, respectively, but other Taf paralogues (Taf5l and Taf6l) were expressed. Their protein products can substitute each other in the complexes, e.g. both TAF6/6L can heterotetramerize with TAF9 (Shao et al., Mol Cell Biol., 2005). The expression of Tbpl1 was unexpected as it has been described to be specific for the sperm (Akhtar and Veenstra, Cell Biosci., 2011).

We also collected the expression levels of the GC-box binding Krüppel-like factors and found that third of these was expressed in BMDM (**Figure 18A**). Klf6 showed the highest expression of them of which protein product had been described as a key factor in pro-inflammatory macrophages (Date et al., J Biol Chem., 2014). Interestingly, Klf10 showed an immediate, significant induction upon LG268 treatment that was followed by a repression both on the nascent and matured RNA level (**Figure 18C-D**). KLF10 has been shown to be specific for the bone marrow derived proangiogenic cells, which induces vascularization (Wara et al., Blood, 2011). Sp1-3 and Klf2, 3 and 13 were also highly expressed and slightly regulated by LG268 (**Figure 18A**). However several activator (Sp1, Klf2-3, 10, 13) and also repressor (Sp2-3) Krüppel-like factors were expressed in the same time; they seemed to have – at least partially – distinct roles in macrophages.

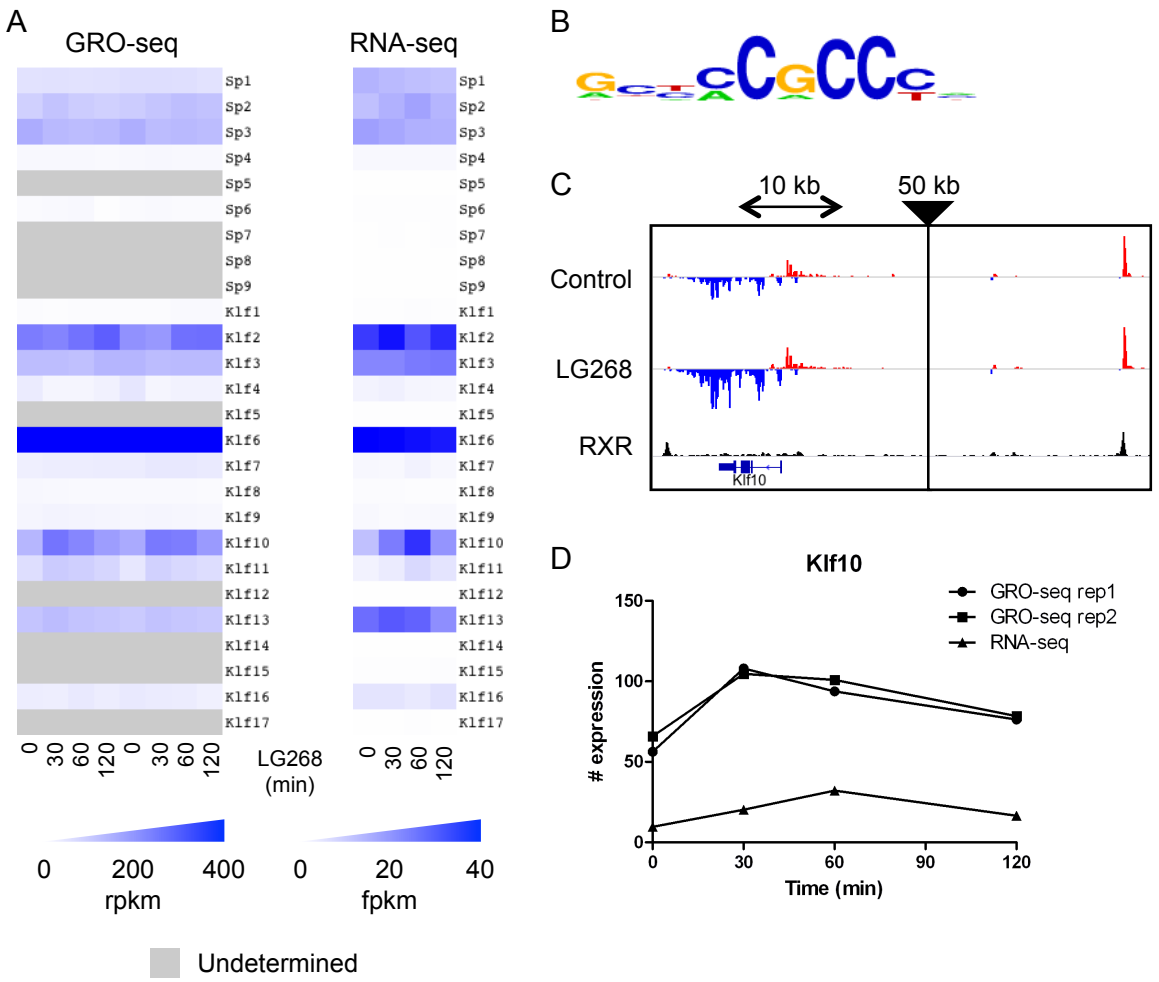


Figure 18. The gene expression changes of Sp1 and Klf families upon LG268 treatment

A) The level of nascent (GRO-seq) and matured mRNA (RNA-seq) was determined as RPKM and FPKM values, respectively. (Expression values of the two replicates were combined by the RNA-seq analysis pipeline)

B) The GC-box motif enriched under the H3K4me3 derived NFRs (**Figure 10A**)

C) The detected induction of nascent RNA transcription on the Klf10 locus (the strand-specific coverage of GRO-seq data is represented by red and blue columns on the positive and negative strands, respectively; the coverage of RXR ChIP-seq data is represented by black columns)

D) The expression profile of Klf10 highlighted from A)

By comparing the expression of the Ets superfamily, we got that Pu.1 (Sfpi1) showed the highest level (**Figure 19A**). This was not surprising, but that was striking that there was even more Pu.1 mRNA than the sum of all other Ets mRNAs together. At the level of nascent RNA, Etv3 was yet comparable with Pu.1, but the mRNA of the latter showed much more stability. Nevertheless, ETV3 of which coding gene showed induction upon LG268 treatment, is known as an antiproliferative TF during terminal macrophage differentiation by repressing e.g. the Myc gene (Klappacher et al., Cell, 2002). The rather promoter specific Etv6, Elf1, 2 and 4, Elk3 and Fli1 were also expressed on high levels, but according to the mRNA levels, it was not surprising that the high amount of PU.1 could supersede their protein products even from the promoters as it has been shown on **Figure 12**. In the light of this data, it is understandable how PU.1 is able to bind to tens of thousands of binding sites, even to the exposed ones of the wrapped EBSs (**Figure 11**). FLI1 of which RNA showed a slight induction upon treatment, is also needed for macrophage differentiation (Suzuki et al., Immunology, 2013), so its presence – similarly to the basic promoter-binding ETS proteins – was also not totally unexpected. Fli1 is a long gene so its transcription and the RNA maturation take more than an hour. This was why we could not see the direct regulation at the mRNA level earlier than 2 hours of LG268 treatment. Ets2 and Etv5 showed a more significant mRNA accumulation, however the one of Ets2 could not be explained by the nascent RNA production. Interestingly, there were two induced family members, Fli1 and Ets2 (**Figure 19C-D**), which had been connected to angiogenesis in different cell types (Morita et al., Proc Natl Acad Sci U S A., 2015; Craig et al., Arterioscler Thromb Vasc Biol., 2015; Wallace et al., PLoS One, 2013). This may suggest that RXR activation enhances the angiogenic potential of macrophages through ETS proteins, which have higher affinity to the promoter specific EBSs than PU.1.

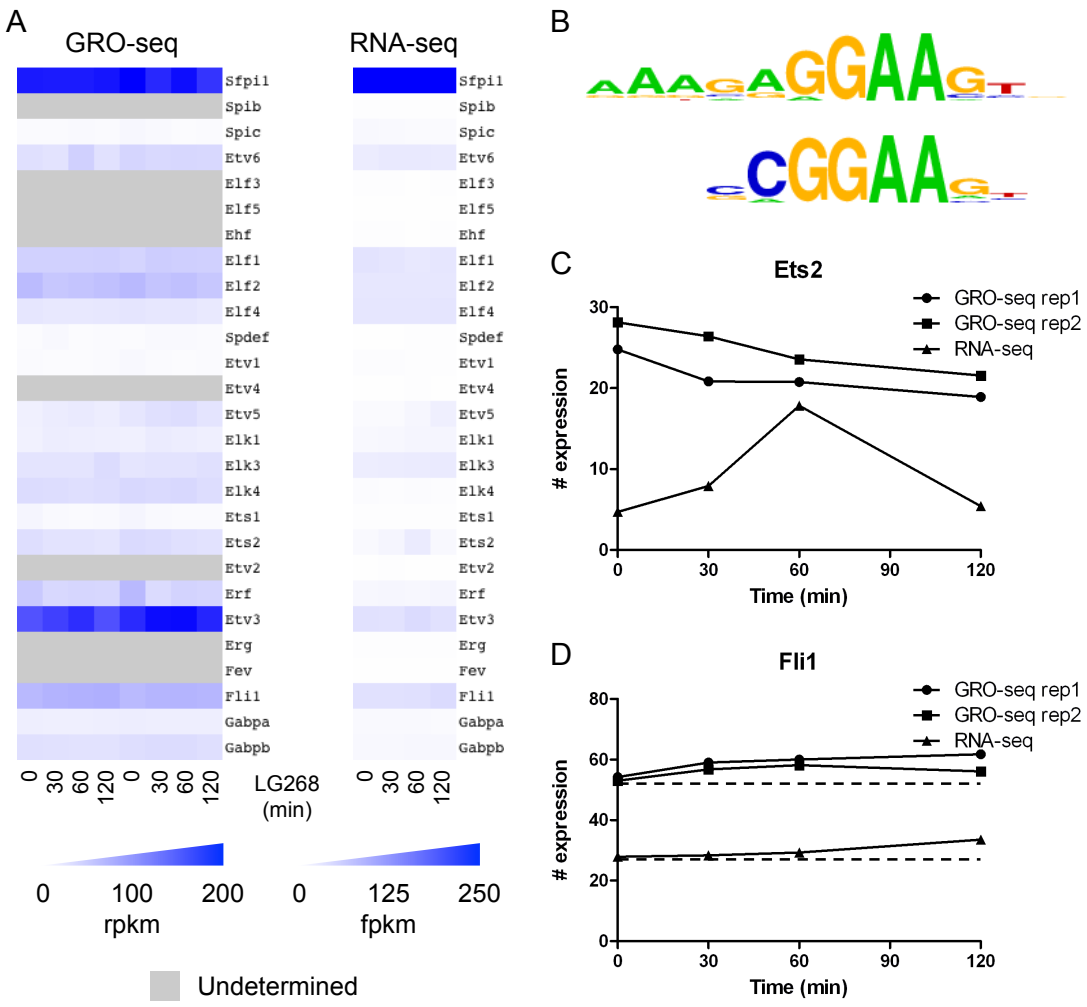


Figure 19. The gene expression changes of the Ets family upon LG268 treatment

A) The level of nascent (GRO-seq) and matured mRNA (RNA-seq) was determined as RPKM and FPKM values, respectively. (Expression values of the two replicates were combined by the RNA-seq analysis pipeline)

B) The PU-box and ETS-box motifs enriched under the H4ac and H3K4me3 derived NFRs, respectively (**Figure 10A**)

The expression profile of Ets2 (C) and Fli1 (D) highlighted from A)

Although we could not find E-box motif enriched, we compared the gene expression of Max, Mlx and their heterodimerizing partners (**Figure 20A**). MAX as the central member of this protein group, together with MXD4 showed the highest mRNA level. MXD4 seemed to be the most dominant partner of both MAX and MLX, but MLXIP and MXD1 might also

form dimers with MLX, and the further partners of MAX were probably the MNT, MXD1 and MXI1 proteins. The nascent RNA level of Myc showed induction upon LG268 treatment, but the effect seemed transient at the level of mRNA, which could be explained by the change of RNA half-life. Ultimately, MAD repressors might be able to countervail the effects of MYC, which gave a possible reason why we did not see E-box motif enrichment – as we examined the active, not the repressive sites.

Then we went further with the promoter specific TFs (**Figure 20A**). The resupply of Nfyb and c was similarly much less than the one of Nfya, but the shortest Nfyb showed the lowest expression at the level of mRNA, which is hard to interpret. The enigmatic motif of "GFY" (**Figure 20B**) has been shown to be bound by two distinct proteins, STAF (ZFP143) (Schaub et al., EMBO J., 1997; Hong et al., Genomics Inform., 2012) and Ronin (THAP11) (Dejosez et al., Cell, 2008). As Thap11 was highly expressed, while Zfp143 showed low expression, it was probably decided what GFY was in our system, however the further THAP family members (THAP3-4 and 7) might also bind this motif (**Figure 20B**). The mRNA of Zbtb33 (Gfx) showed very low expression, but the promoter specific NFRs gave the clear enrichment of its motif. The appearance of GFX motif might be accidental or imply that Gfx had roles thus had long half-life in BMDMs. Nrf1 showed a bit higher, Yy1 a much higher expression than Gfx, which was in agreement with the motif enrichments (**Figure 20B**).

There are 16 AP-1/CREB-related families with 44 members of the bZIP superfamily that bind rather TRE or CRE, respectively (**Tables 3-4**). As we could not find a unified comparison of these proteins, a phylogenetic analysis was applied on their most conserved, 56 amino acid long basic (DNA-binding and dimerization) domains (**Figure 21**). Thus, the known families could be discriminated and their relation with the recently identified members became clearer. Of the 44 genes, several showed high expression from almost all family

(except for the Atf1 and Atf2 families) (**Figure 22**), so it was not easy to identify the interacting members.

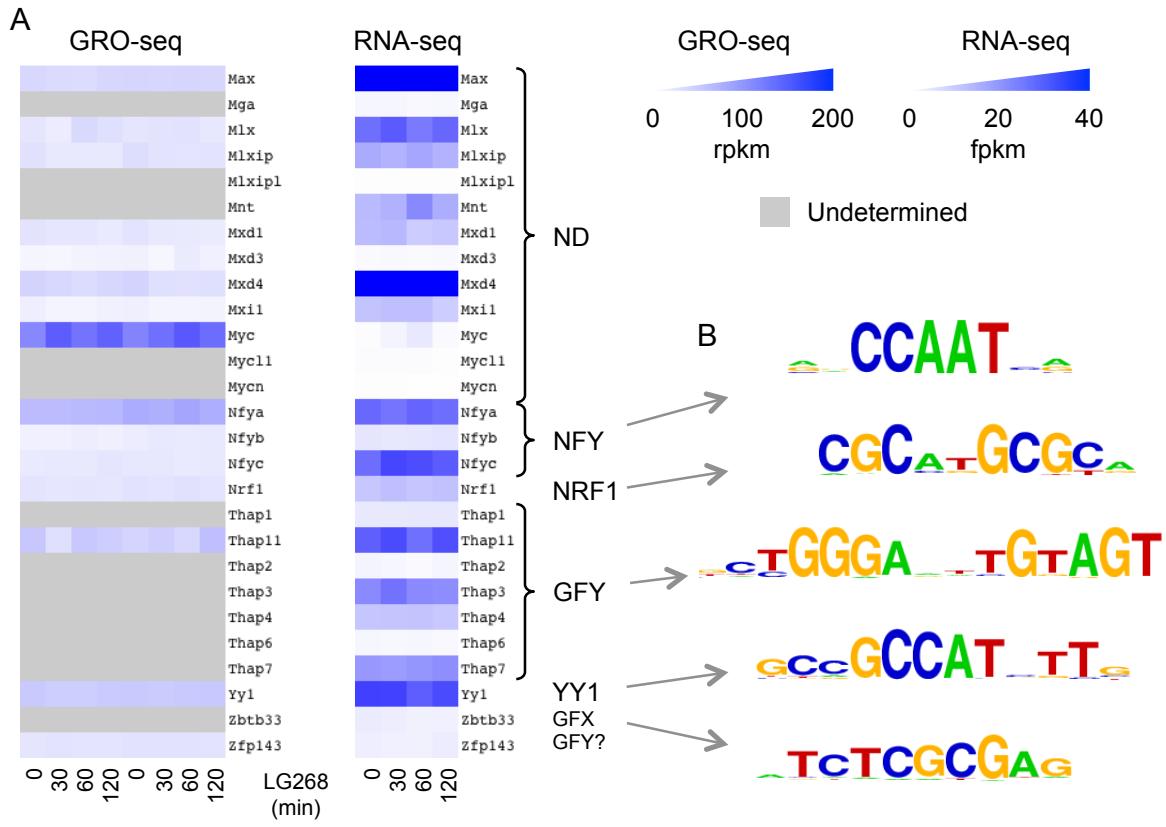


Figure 20. The gene expression changes of promoter binding TF families upon LG268 treatment

A) The level of nascent (GRO-seq) and matured mRNA (RNA-seq) was determined as RPKM and FPKM values, respectively. (Expression values of the two replicates were combined by the RNA-seq analysis pipeline)

B) The promoter specific motifs enriched under the H3K4me3 derived NFRs (**Figure 10A**)

Jun genes of which products are the main components of the AP-1/CREB dimers, were all expressed, however Junb and Jund had really high mRNA levels. Of their partners (Newman and Keating, Science, 2003), Atf3 showed the highest expression level, while Mafg, Mafk, Batf genes, Fos and Jdp2 showed lower and lower levels (**Figure 22**). These latter two genes were induced by RXR activation, which meant a quick, transient action of

Fos, while Jdp2 as a longer gene with 40 kb length needed more time to get its mRNA level elevated (**Figure 23A-B**). This has a biological sense, as FOS is rather an activator with high affinity to JUN proteins thus can supersede ATF3 and BATF repressors from the TREs (**Figure 23C**), while JDP2 is a repressor of JUN proteins, which can cause then repression on the very same elements (**Figure 23D**).

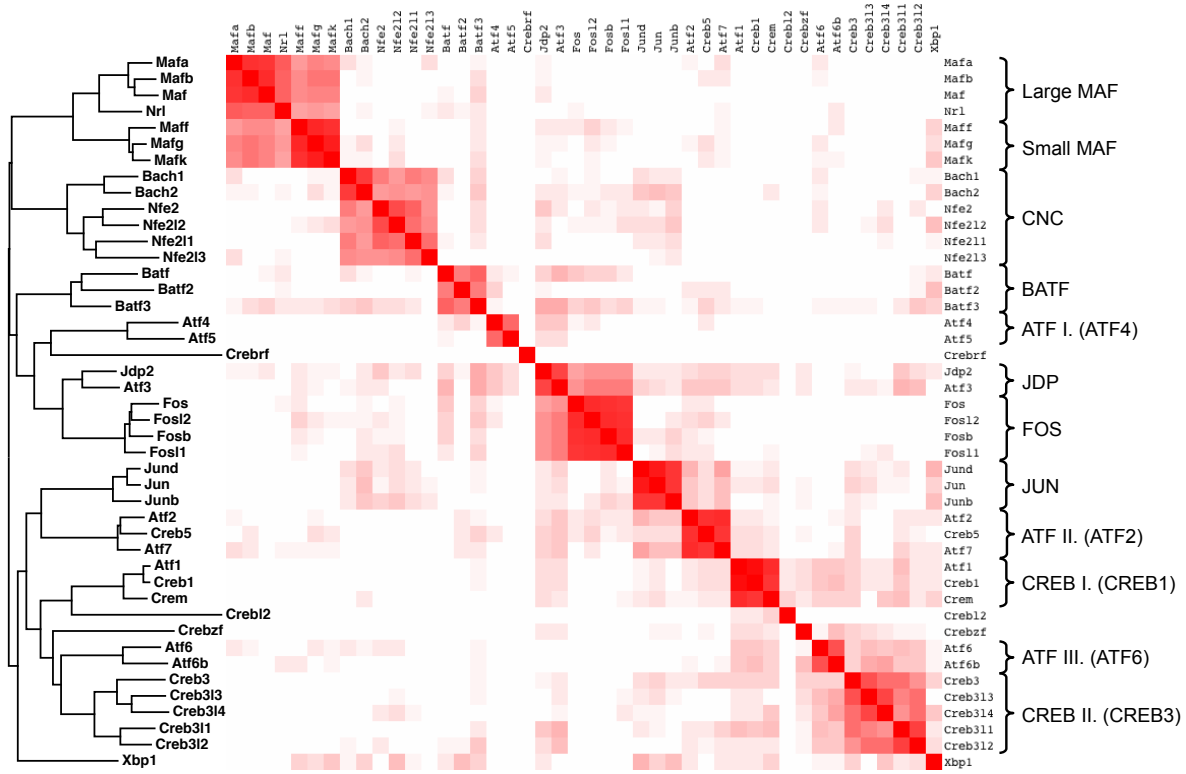


Figure 21. The phylogenetic comparison of the AP-1/CREB related bZIP proteins

The phylogenetic tree (left) and similarity matrix (right) of the basic domain of bZIP proteins (bZIP family names are included; red scale indicates the degree of similarity)

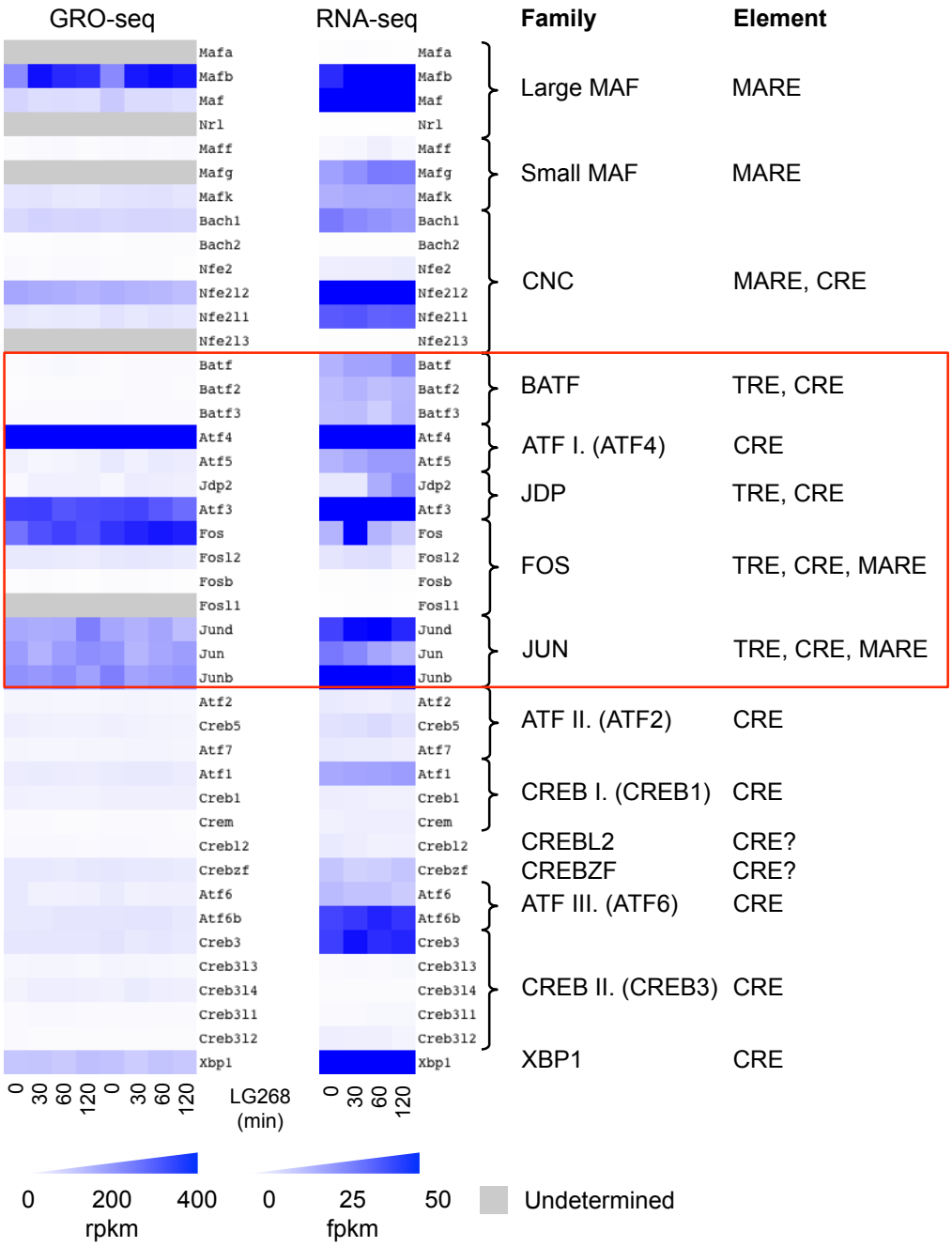


Figure 22. The gene expression changes of the AP-1/CREB related families upon LG268 treatment

The level of nascent (GRO-seq) and matured mRNA (RNA-seq) was determined as RPKM and FPKM values, respectively (left). (Expression values of the two replicates were combined by the RNA-seq analysis pipeline)

The name of the protein families and their response elements (right)

The TRE/CRE binding families are highlighted by red square.

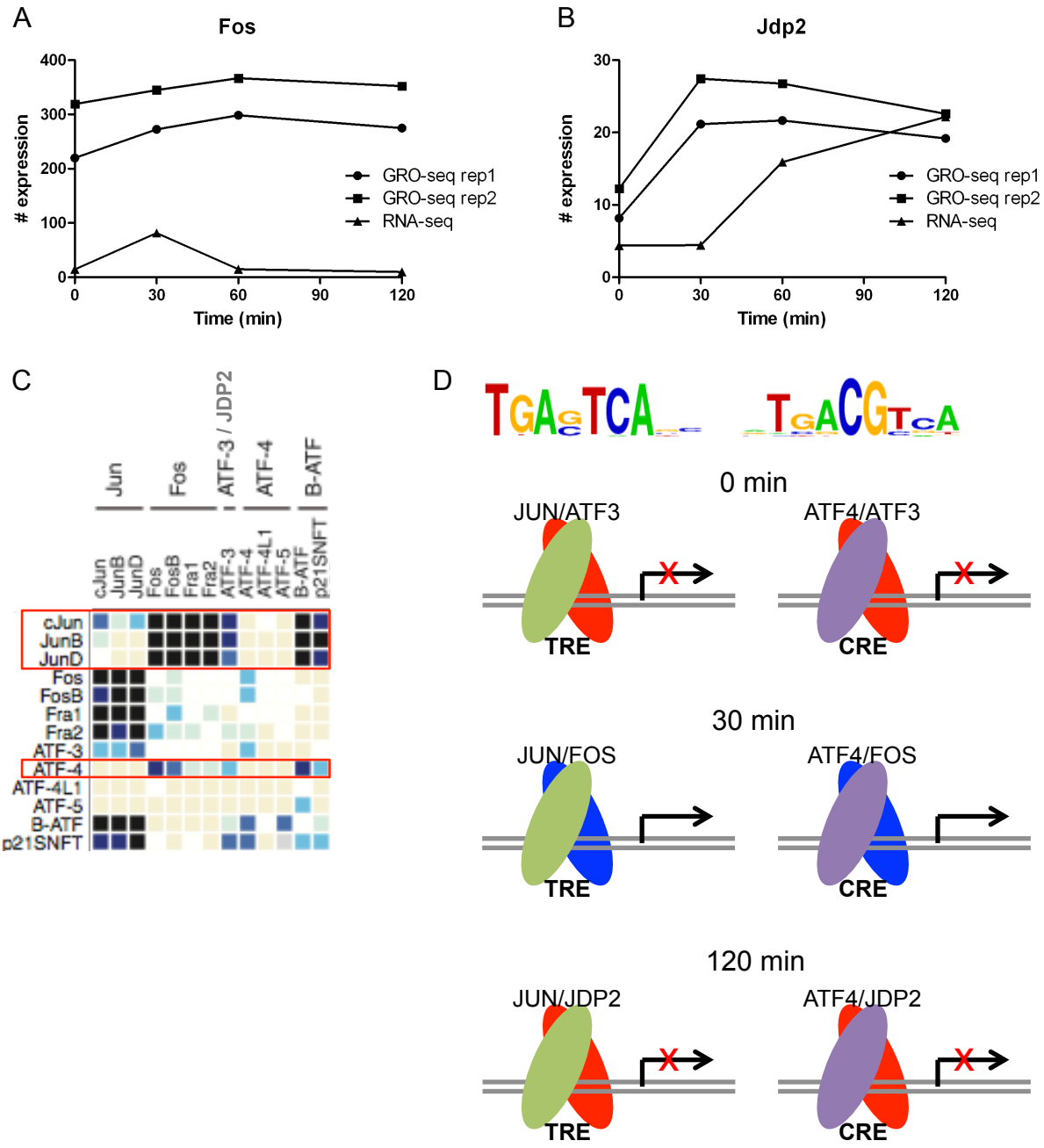


Figure 23. Fos and Jdp2 show induction upon LG268 treatment

The expression profile of Fos (A) and Jdp2 (B) highlighted from **Figure 22**

C) The dimerization affinity of the basic domain of TRE/CRE binding proteins (**Figure 3**, Newman and Keating, Science, 2003)

D) The model of the FOS/JDP2 switch on TRE and CRE

The TRE and CRE motifs enriched under the H4ac and H3K4me3 derived NFRs, respectively (**Figure 10A**)

ATF4 of which coding gene showed the highest transcriptional activity of the AP-1-related genes, can dimerize with all JUN partners of the previous model (but JUN!), the ATF3 and BATF family members, as well as FOS (**Figure 23C**), however these heterodimers act on CREs (**Figure 23D**). This suggested that ATF4 might play a similar role as JUN, in parallel on different regulatory sites. Interestingly, it has been described that ATF4 activated "vascular endothelial growth factor A" (VEGFA) expression and thus induced vascularization (Roybal et al., J Biol Chem., 2004). Beside the induction of Klf10, Fli1 and Ets2, this is a further line of evidence that RXR activation promotes angiogenesis, via FOS/ATF4 activation, as well.

MAFB forms homodimers and also heterodimers with FOS, MAF and BACH (here rather BACH1) proteins (**Figure 3**). Its nascent transcript showed the highest induction (**Figure 22**), but no MARE could be detected in the motif enrichment analyses (**Figure 10A**). Of the CNC family, beside the Bach1 gene, Nfe2l1 and Nfe2l2 were expressed, with increasing mRNA levels in this order. The partners of their protein products are typically small MAFs, which overlap with the preference of JUN proteins. The expression of CREB3 and ATF6B could not affect the transcriptional regulation, as these are membrane bound proteins in the absence of ER stress (Audas et al., Mol Cell Biol., 2008; Yoshida, Cell, 2001). XBP1, which is also able to form heterodimers with ATF6(B), thus probably formed homodimers as its mRNA level was very high.

Of the rather macrophage specific Cebp (bZIP) family, Cebpa and Cebpb showed the highest mRNA level. Cebpd, Cebpg and "DNA-damage-inducible transcript 3" (Ddit3) or C/EBP-homologous protein (Chop) genes were also expressed on higher levels compared to Cebpz and Cebpe (**Figure 24A**).

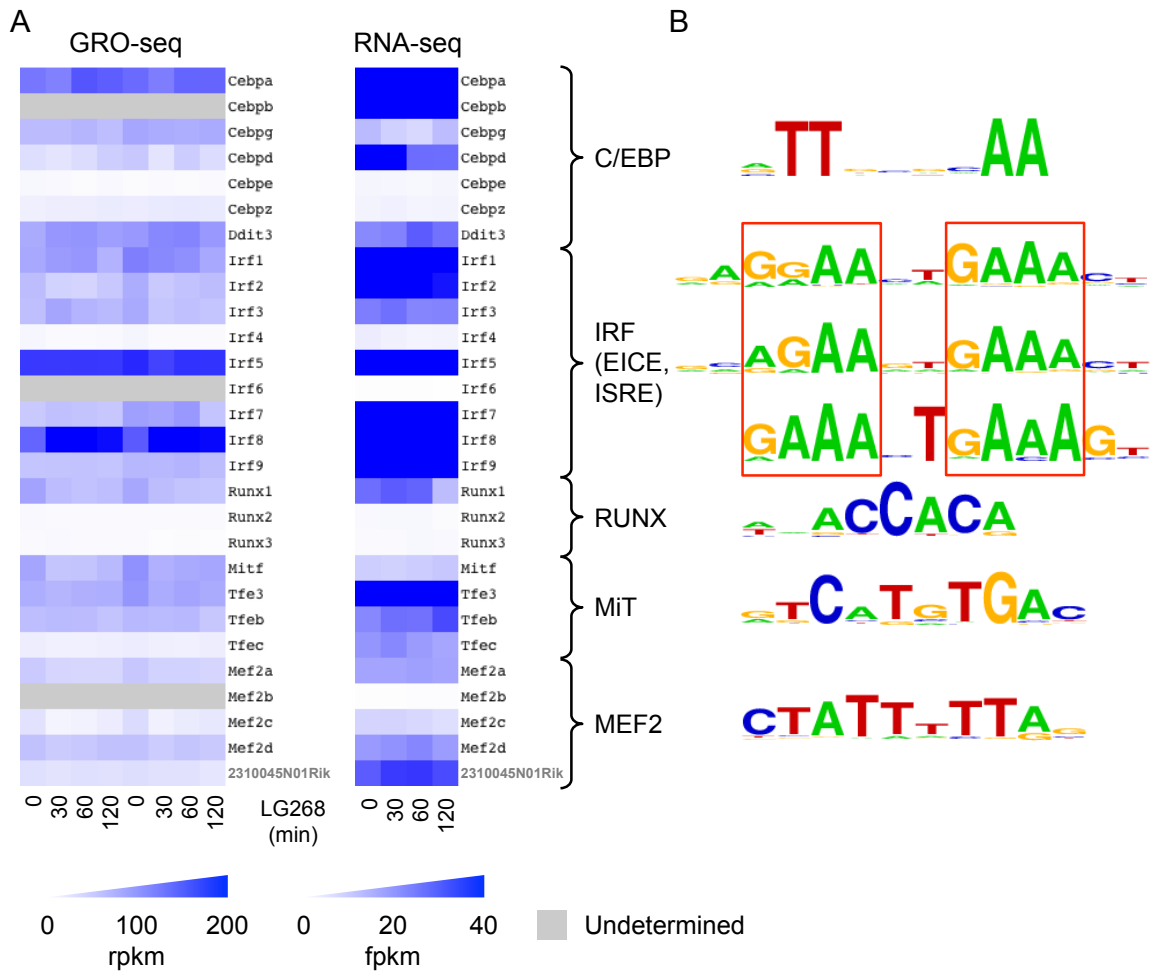


Figure 24. The gene expression changes of macrophage specific TF families upon LG268 treatment

A) The level of nascent (GRO-seq) and matured mRNA (RNA-seq) was determined as RPKM and FPKM values, respectively. (Expression values of the two replicates were combined by the RNA-seq analysis pipeline)

B) The motifs specific for the different TF families enriched under the H4ac and H3K4me2 derived NFRs (**Figures 10A and 13**)

Except for Irf6, which is specific for keratinocytes (Ingraham et al., Nat. Genet., 2006; Richardson et al., Nat Genet., 2006), all Irf genes were found expressed in BMDMs (**Figure 24A**). Among these, the rather lymphoid specific Irf4 showed very low expression level, while Irf8 and Irf5 showed the highest transcriptional activities. Irf1, 2, 7 and 9, which can be considered as ubiquitously expressed genes had similarly high mRNA levels as Irf5 and Irf8. IRF5 is the main transducer of TLR signaling (Takaoka et al., Nature, 2005); and IRF8 of which nascent transcript showed induction upon LG268 treatment, is the main dimerizing partner of PU.1 in macrophages (e.g. Mancino et al., Genes Dev., 2015). IRF1 had been also shown as partner of IRF8, but beside the heterodimer of IRF3 (showing lower mRNA expression) and IRF7 – key factor in macrophage differentiation (Lu and Pitha, J Biol Chem., 2001) –, several other IRF homo- and heterodimers could act in BMDMs. However, in the lack of phosphorylation caused by PRR activation, dimerization and DNA-binding might not be too frequent as the detected ISREs were nor frequent (**Figures 10A and 13**). Although BATF and JUN proteins were expressed, AICEs could not be detected in the motif enrichment analyses, probably because this AP-1/IRF interaction is less frequent.

By using gene expressional data, we identified, which proteins could be responsible for the M-box (HLH-like, bHLH-ZIP) and MEF2 (MADS) motif enrichments (**Figures 13 and 24**). Among the MiT genes, the Tfe3 mRNA was accumulated on the highest level, which, if there is no preferred homo- or heterodimers, suggests that TFE3 homodimers dominate on the M-boxes. Of the Mef2 genes, a less known one, the 2310045N01Rik gene showed high expression at the mRNA level, while Mef2a, d and c were lower expressed. Runx1 was the only expressed member of its family, which was consistent with the literature (e.g. Lichtinger et al., EMBO J., 2012).

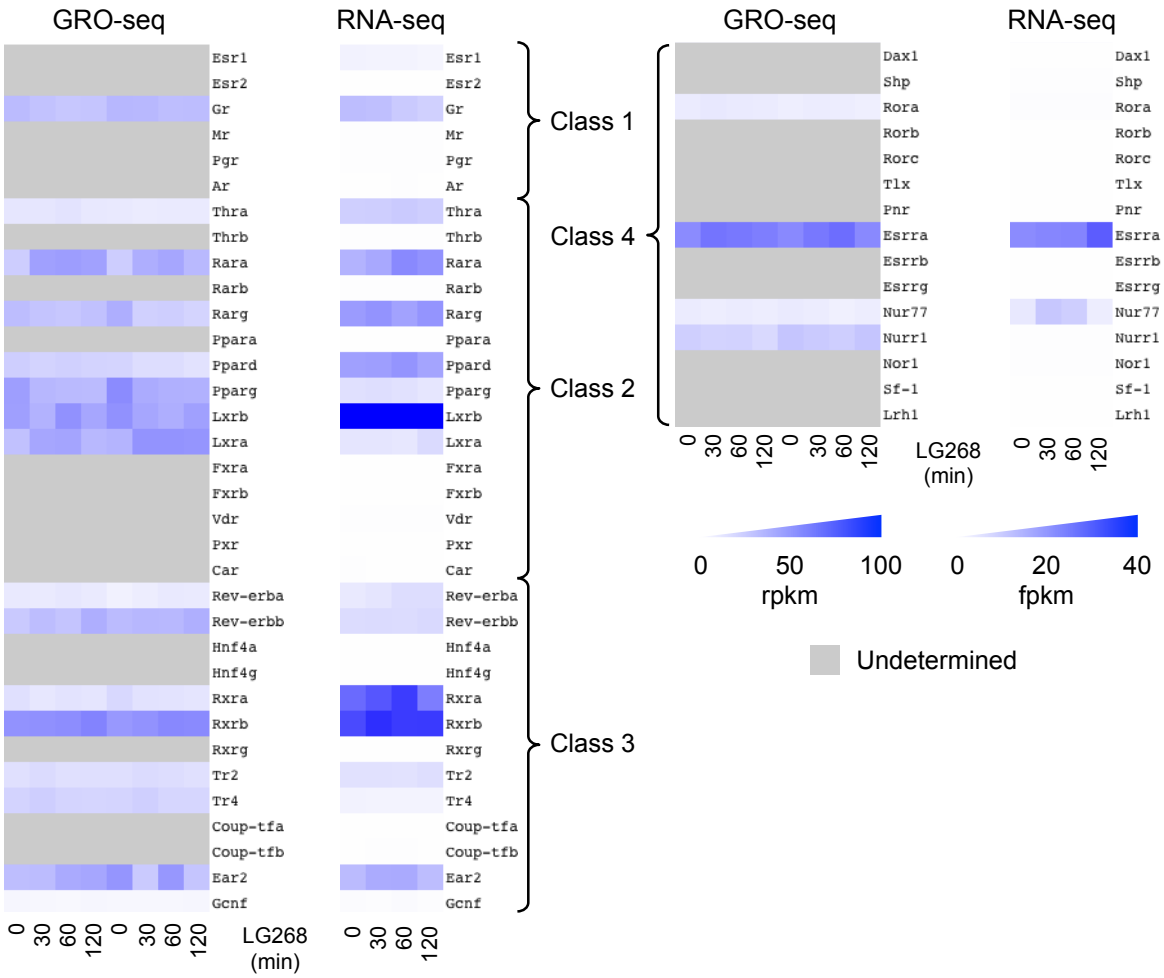


Figure 25. The gene expression changes of the NR superfamily upon LG268 treatment

The level of nascent (GRO-seq) and matured mRNA (RNA-seq) was determined as RPKM and FPKM values, respectively. (Expression values of the two replicates were combined by the RNA-seq analysis pipeline; NR classes are shown)

Based on our gene expression results, BMDM cells are probably capable to respond to glucocorticoid, thyroid and with a lesser extent to estrogen hormones by GR, THRA and ESR1, respectively (**Figure 25**). Class II NRs could heterodimerize with both RXR alpha and beta of which mRNAs got stabilized upon LG268 treatment. Rara was highly induced by LG268 treatment, while Rarg seemed repressed on the nascent RNA level; however in the latter case, this could not be seen on the matured mRNA level. Similarly, both Lxr genes were regulated by LG268: Lxra was induced and Lxrb was repressed upon RXR activation, but this

did not affect their mRNA level: Lxrb was the highest expressed NR, while Lxra showed low expression. Surprisingly, there was a direct Pparg repression upon LG268 treatment, and Ppard dominated over Pparg on the mRNA level. These results suggested that BMDMs crucially might have LXRB/RXR and RAR/RXR heterodimers, but the PPARD/RXR, THRA/RXR and GR/GR dimers might be also specific for these cells. Of the orphan receptors, both Rev-erb and both Tr genes were expressed; and Ear2, Esrra and Nur77 mRNAs showed an induction upon RXR activation.

4.8. The annotation of the putative regulatory regions

Our system could be moved by RXR activation both at the level of DNA-binding and transcription. Previously, we showed that half of the PU.1 bound sites were inactive (**Figure 11**), and the bigger portion of RXR peaks overlapped with PU.1 peaks (**Figure 14**); thus it was a question whether this phenomenon was true also for RXR or most RXR bound sites were active. The presence of divergent transcripts means that polymerase and its loading factors (PIC) are also present together with the additional regulatory proteins. This was not striking on promoters, but it occured on enhancers as well, thus GRO-seq might give more direct evidence for enhancer activity than histone modifications. Thus to respond which portion of the RXR binding events really regulate genes, we used the higher resolution GRO-seq data.

We predicted 51,657 divergent sites of which bigger half was in the close proximity of putative TSSs, most of the remaining regions overlapped with or were close to the 5' end of the predicted transcripts, and only 8% was intergenic, located farther than 10 kb from the beginning of any transcripts (the top left of **Figure 26**). More precisely, the identified 11,235 genes had 17,254 promoter-specific divergent sites, while the 12,821 unknown transcripts had 7,192 such ones, which together meant 21,849 divergent sites – less than the sum of the

previous numbers (because of the 2,597 "common" or nearby promoters); and 268 sites showed just the transcriptional initiation of known genes. Thus together 22,055 divergent sites could be directly connected to the initiation sites of elongated transcripts. The further 5,934 promoter proximal divergent sites could be both unknown alternative TSSs and proximal enhancers. The remaining 23,668 divergent sites were specific for the distal enhancers without elongation. The significant overlap with transcribed regions was due to the large number of detected transcripts, which covered 20.55% of the mouse genome. Strand-specifically this meant that about 11% of the genome, as 1.6% was transcribed on both strands in parallel.

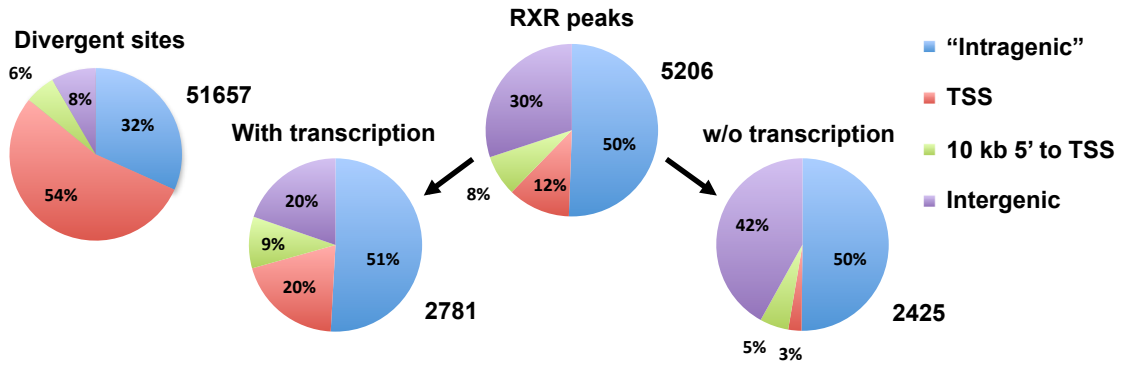


Figure 26. The classification of RXR binding sites based on (enhancer) transcription

The genomic distribution of divergent sites (top left) and all (top right), the transcribed (bottom left) and silent RXR peaks (bottom right) relative to the closest expressed transcripts (over 3 kb length and including 3' overhangs) determined by GRO-seq

In the case of RXR bound sites, 50% of the peaks fell on elongated transcripts, 30% was intergenic and the remaining part coincided with TSSs or element(s) of the 10 kb region upstream to the transcripts (the top right of **Figure 26**). The separation of RXR peaks to possible active regulatory sites and inactive (or poised) regions based on the overlap with divergent sites (dominantly enhancer transcripts) gave a 53.4 vs. 46.6 ratio: 2,781 RXR peaks

fell approximately between divergent transcripts and 2,425 peaks fell onto other location (the bottom of **Figure 26**). The former peak set showed 20% overlap with the TSS proximal regions, a further 9% overlapped with the 10 kb regions upstream to the TSSs, and just 20% were even farther. "Inactive" RXR bound sites showed much less overlap with the active promoters and 42% of them were intergenic.

As we and others (Hah et al., Genome Res., 2013) observed that the expression level of enhancer transcripts showed correlation with the level of the belonging gene, we wanted to annotate the regulatory binding events of RXR with the help of divergent sites. Firstly, we were curious whether these followed the global up regulation tendency upon LG268 treatment that was previously observed in RXR binding. For this, we performed the differential binding analysis of all predicted divergent sites, where 2,190 regions showed significant up regulation, while 1142 regions showed significant decrease upon ligandation (**Figure 27A**). By comparing these two sets based on their overlap with RXR binding, the difference was large but not surprising: RXR binding was much more specific for increasing enhancer transcription (673 vs. 45 peaks), and there was also a difference in PU.1 binding in the favor of induction (902 vs. 310 peaks).

Our criteria for the annotation were the following: the direction of the significant expressional change of enhancer transcripts with RXR binding in between had to be the same as the one of a regulated gene, which TSS was closer than 1 Mb (**Figure 27B**). Thus we assigned 387 enhancers and 27 "silencers" to 226 up and 26 down regulated genes, respectively (**Figure 27C**). Most genes had one or few RXR-bound active enhancers, while Abca1 had 9 such regions. The annotated enhancers distributed roughly symmetrically compared to the TSSs showing a high TSS preference (**Figure 27D**). 226 genes of the 318 up regulated ones meant a high ratio (71%) compared to the ratio of the down regulated genes – 26 vs. 423 (6.1%). These numbers suggested that RXR ligandation caused mainly direct gene

activation, but there were numerous indirect but quick down regulation events, as well. Inhibition could be caused by the transcriptional activation of short TF(s) with repressor function, which could be activated quickly upon LG268 treatment, or e.g. the redistribution and abstraction of TFs and their co-activators from the eventually repressed genes.

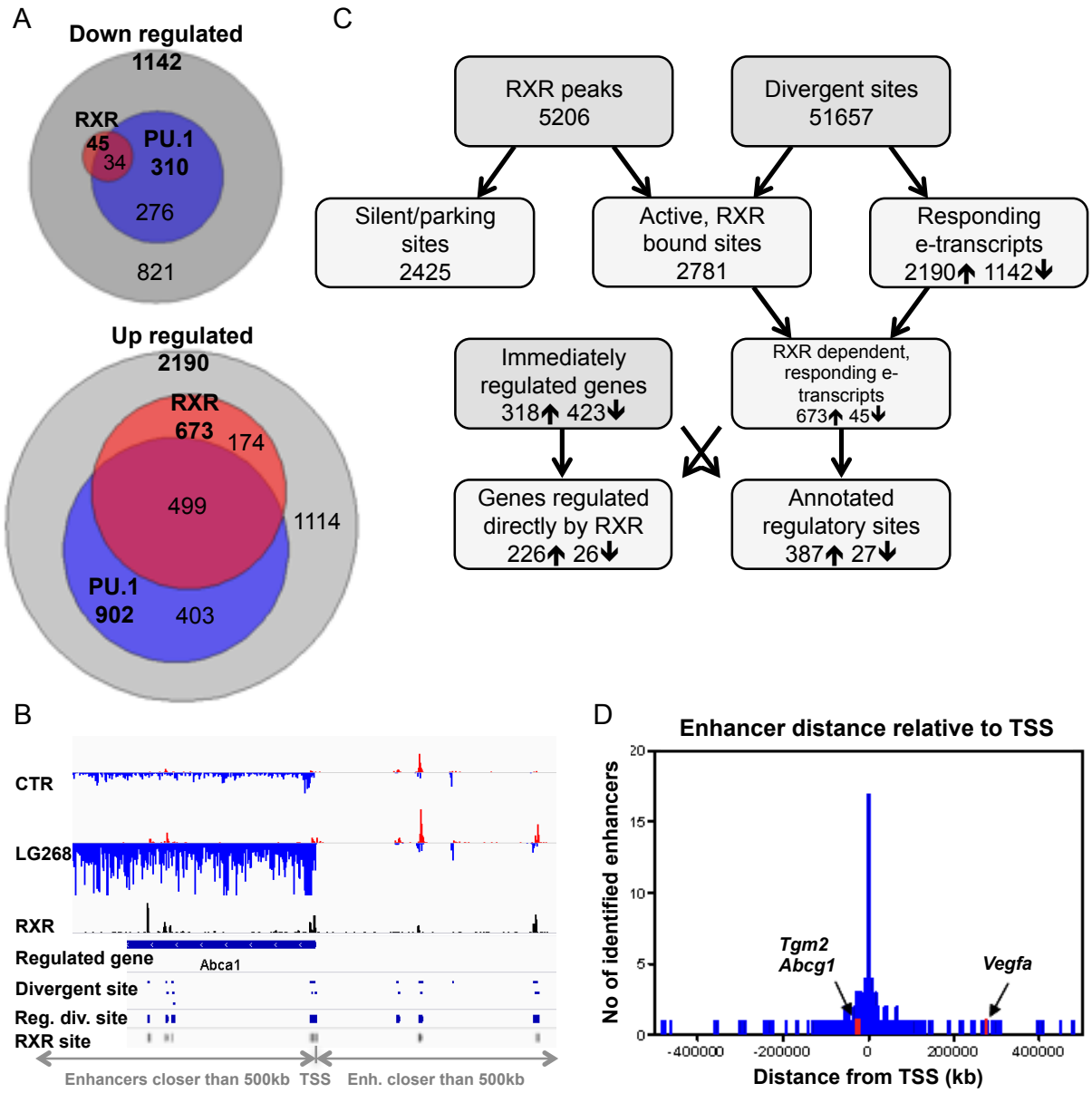


Figure 27. The identification of RXR regulated enhancers and linking those to regulated genes

A) The overlap of the cistrome of liganded RXR and PU.1 and the divergent sites with significantly decreasing (top) and increasing expression (bottom) upon LG268 treatment

The further part of **Figure 27** *description:*

C) The flowchart of pairing active RXR bound enhancers to regulated genes

D) The distribution of RXR driven enhancers relative to the TSS of the identified directly regulated genes

As there were more up regulated TFs affecting angiogenesis, we were curious whether these really had an effect on the genes responsible for this process (**Figure 28A**). We found that 25 of the genes that were directly induced upon RXR ligandation were related to angiogenesis (GO:0001525), morphogenesis (GO:0048514), the development (GO:0001568) or remodeling of blood vessels (GO:0001974) or endothelial cell proliferation (GO:0001935). Gene lists belonging to these GO categories were downloaded from the Mouse Genome Informatics (MGI) database. Most genes (21 of 25) had RXR peak(s) nearby (examples are shown in **Figure 28B**), so we concluded that RXR may also be an angiogenic TF. Its effect lasted for at least an hour on the nascent RNA level (**Figure 28A**), which might be due to the transient induction of its putative co-operating partners, FOS/ATF4, ETS2, KLF10 and FLI1, respectively (**Figure 28C**). The sequential induction of these TFs might explain the observed expression patterns; and the delayed slight induction of FLI1 could have also role in the longer-term regulation of some of the listed or other angiogenic genes. Remarkably, the resulted 21 genes included Vegfa and angiopoietin-like 4 (Angptl4) (**Figure 28B**), which are well known angiogenic genes showing induction also on the mRNA level.

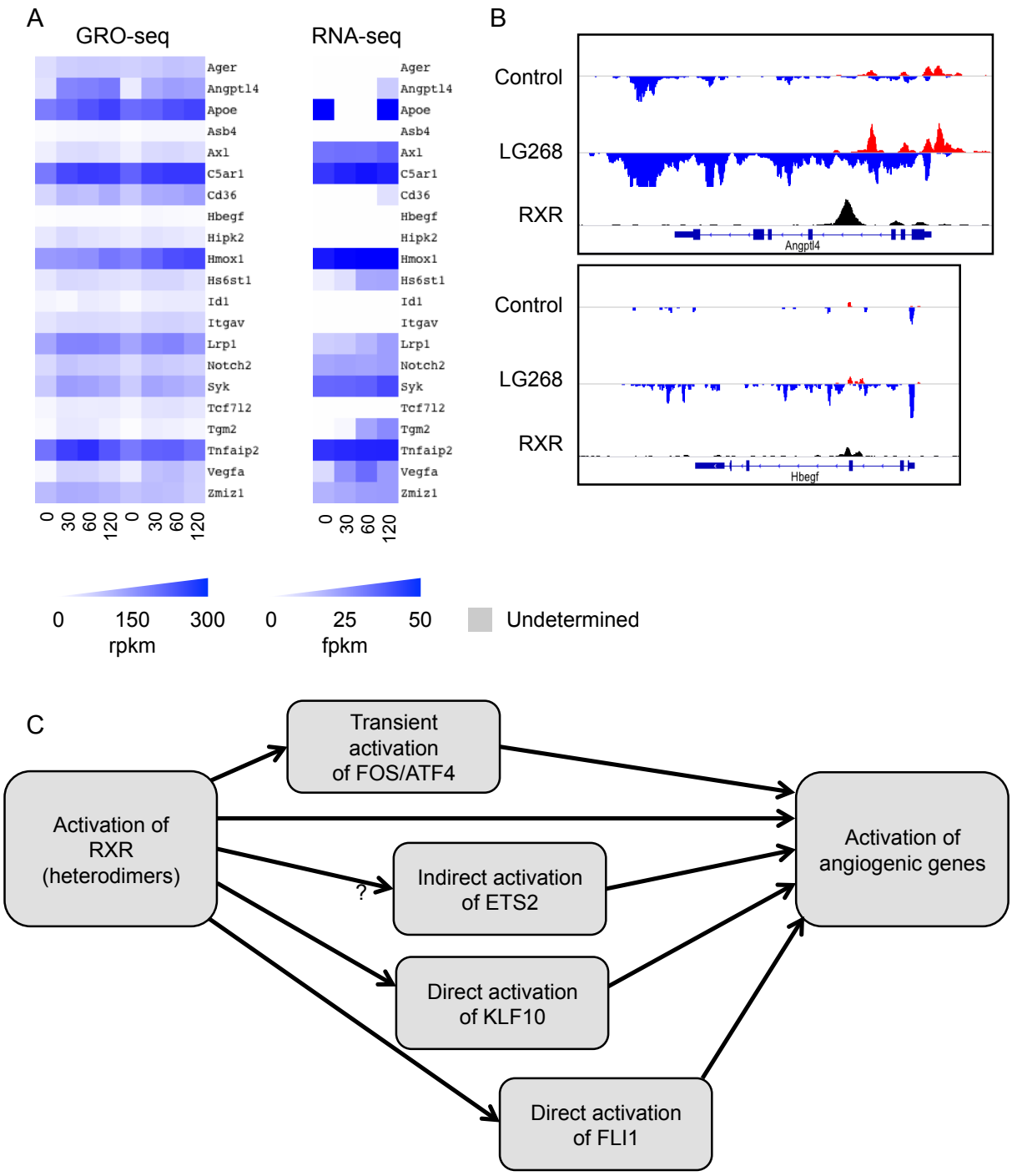


Figure 28. RXR activation induces angiogenic genes

A) The level of nascent (GRO-seq) and matured mRNA (RNA-seq) was determined as RPKM and FPKM values, respectively. (Expression values of the two replicates were combined by the RNA-seq analysis pipeline)

The further part of **Figure 28** *description:*

B) The detected induction of nascent RNA transcription on the Angptl4 and Hbegf locus (the strand-specific coverage of GRO-seq data is represented by red and blue columns on the positive and negative strands, respectively; the coverage of RXR ChIP-seq data is represented by black columns)

C) The model of sequential TF induction affecting angiogenesis upon RXR activation

4.9. Putative binding elements of the annotated RXR peaks

We carried out a motif enrichment analysis for the RXR-bound regulatory sites to get the possible heterodimerizing partners responsible for the induction. The motif distribution of the annotated RXR peaks gave that 57% of the peaks contained NR motifs (**Figure 28**): DR1 and DR4 were the most abundant, 17% contained other repeats and 7% carried weaker elements seemed half sites. DR1 is typically bound by PPAR/RXR heterodimers, but beside the DR5 and DR2 elements, RARs are also able to bind DR1 as dimers formed with RXR (Durand et al., Cell, 1992) (**Table 5**). Based on the gene expressional data, PPAR delta, RAR alpha and RAR gamma were the best candidates for DR1 binding, and DR4s were probably bound by the LXRB/RXR or THRA/RXR heterodimers (**Figure 25**). The DR1 binding of testicular orphan receptor homodimers was imaginable but not probable because of the lower expression of Tr2 and 4 genes and the lack of known protein-protein interactions with RXR. COUP-TF gamma (coded by Ear2) is theoretically also able to bind DR1 and 4 (Kadowaki et al., Biochem Biophys Res Commun., 1992), but we have even less knowledge about this protein and coding gene.

The smaller half of the 387 enhancers did not contain any NR binding sites: 20% carried only PU-box, and the remaining part lacked these elements too (**Figure 28**). This result showed that RXR could be detected not only on its binding sites but also in connection with PU.1 or other TFs (such as AP-1 or C/EBP) like a co-regulator. This would not be unprecedented in the NR superfamily because ER and GR are known to function as co-

regulators, e.g. in cooperation with the FOX pioneer factors even in the lack of their binding sites (Cirillo et al., EMBO J., 1998; Gao et al., Mol Endocrinol., 2003; Carroll et al., Cell, 2005); and most REV-ERB alpha bound sites are unaffected by the DBD KO (Zhang et al., Science, 2015). Another explanation is the phenomenon of tethering, during which RXR – detected without its binding site at one genomic region – binds directly to another region. The complex containing RXR and PU.1 may be the basis of the formation of functional loops between the promoter and enhancer(s).

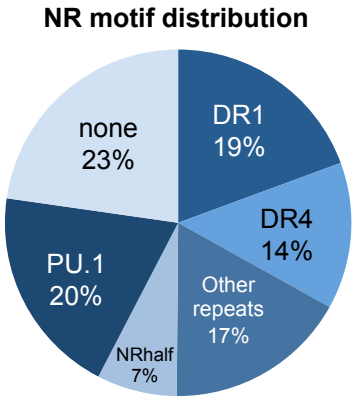


Figure 29. Motif distribution of the annotated 387 RXR-bound enhancers (Additional PU.1 sites can be found in conjunction with other NR motifs)

4.10. The examination of functional domains

To highlight looping in our macrophage system, we carried out ChIP-seq also for the insulator specific CCCTC-binding factor (CTCF) and RAD21, which latter is part of the cohesin complex. Recently, these proteins have been functionally connected, as they co-occupy the topological domain borders (Merkenschlager and Odom, Cell, 2013; Sofueva et al., EMBO J., 2013). Based on our results, 30,290 CTCF and 24,648 RAD21 consensus peaks could be determined and both proteins seemed to bind more regions upon RXR activation (**Table 8**).

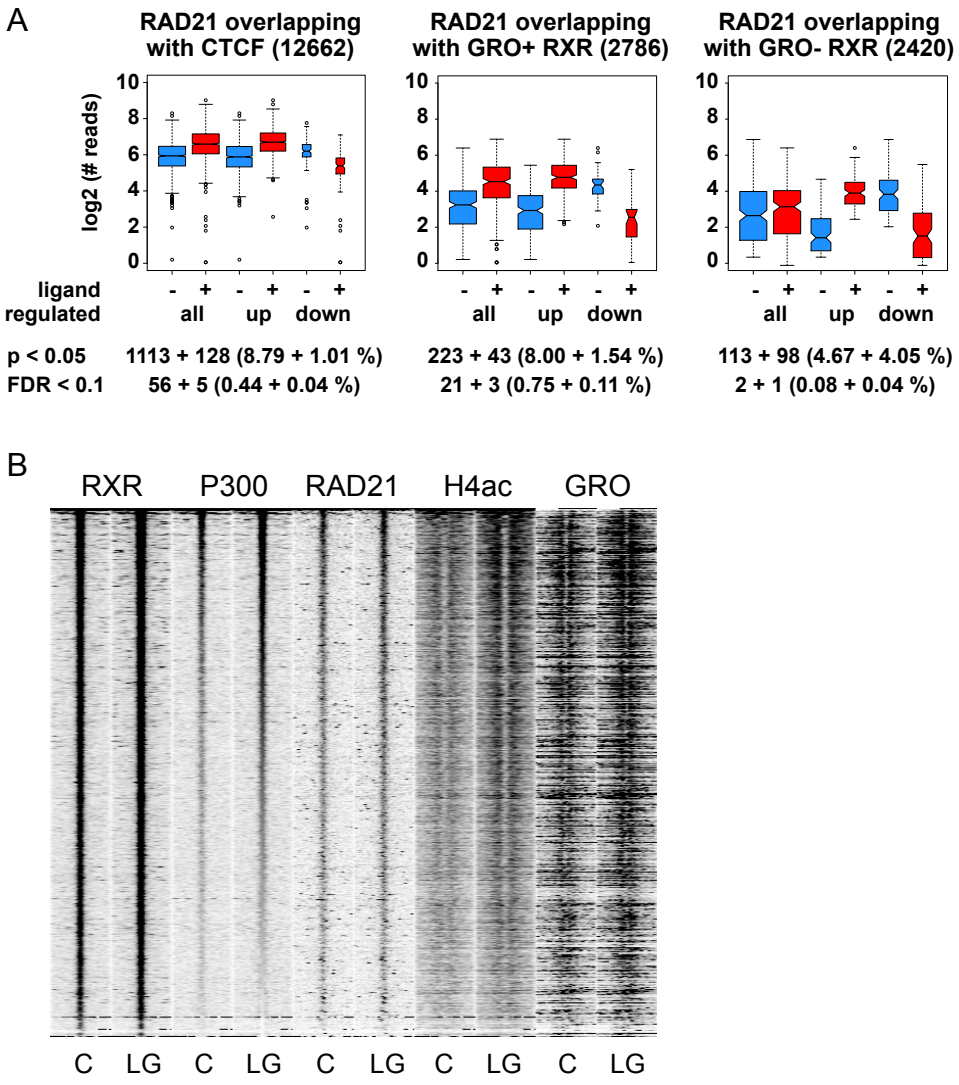


Figure 30. RAD21 is affected by RXR activation

A) The significantly changing RAD21 occupancy upon LG268 treatment on CTCF binding sites (left) and RXR binding sites with (middle) or without (enhancer) transcription (right) (The number of changing peaks and the statistical stringency applied are indicated below each plot)

B) The read distribution heat map of RXR, P300, RAD21 and H4ac ChIP-seq data and GRO-seq data relative to the 5206 RXR peak summits in 3 kb ranges sorted by the fold change of P300 occupancy upon LG268 treatment

To validate the changes, firstly, we selected the insulator related RAD21 peaks and compared their binding in the two conditions. Of the 12,662 CTCF/RAD21 co-peaks, 1113 showed significant induction in response to LG268 treatment, while 128 showed decrease according to the p-values (the left of **Figure 30A**). As we observed RAD21 binding at the most active regulatory regions including those occupied by RXR (**Figure 30B**), we were curious whether we could measure the induction at these sites. Out of the 2786 RXR binding sites expressing divergent transcripts, 223 significantly induced and only 21 repressed ones could be determined, while the 2420 "inactive" RXR peaks gave a more balanced ratio with 113 and 98 regions, respectively (the middle and right of **Figure 30A**, respectively). Putting these together, the results suggest that the induction of cohesin binding both at insulator and active enhancer regions followed the changes seemed in RXR binding.

We determined all neighboring co-peak pairs having similar coverage within 1 Mb range to include even large domains (parameters are detailed in the Methods section). By keeping only the closest pair for each co-peak, we determined 10,204 putative functional domains with the median length of 81.15 kb (some examples are shown in **Figure 31**). The unification of these regions gave similar number and size to those of topological domains found previously by Hi-C (Dixon et al., Nature, 2012; Sofueva et al., EMBO J., 2013): almost 700 active domains, having a median length of 1.1 Mb. The domains identified this way covered close to half of the mouse genome. 80% (203/252) of the directly regulated genes by LG268 together with their respective regulatory site(s) were located on their common functional domains. Furthermore, 33% (84/252) of the regulated genes had induced RAD21 coverage on the belonging RXR-bound enhancer(s) and/or CTCF binding site(s). These data suggest that most RXR driven enhancers act inside functional domains/loops, and that the activation of the receptor contributes to loop stabilization thus inducing the interaction frequency of enhancers with promoters.

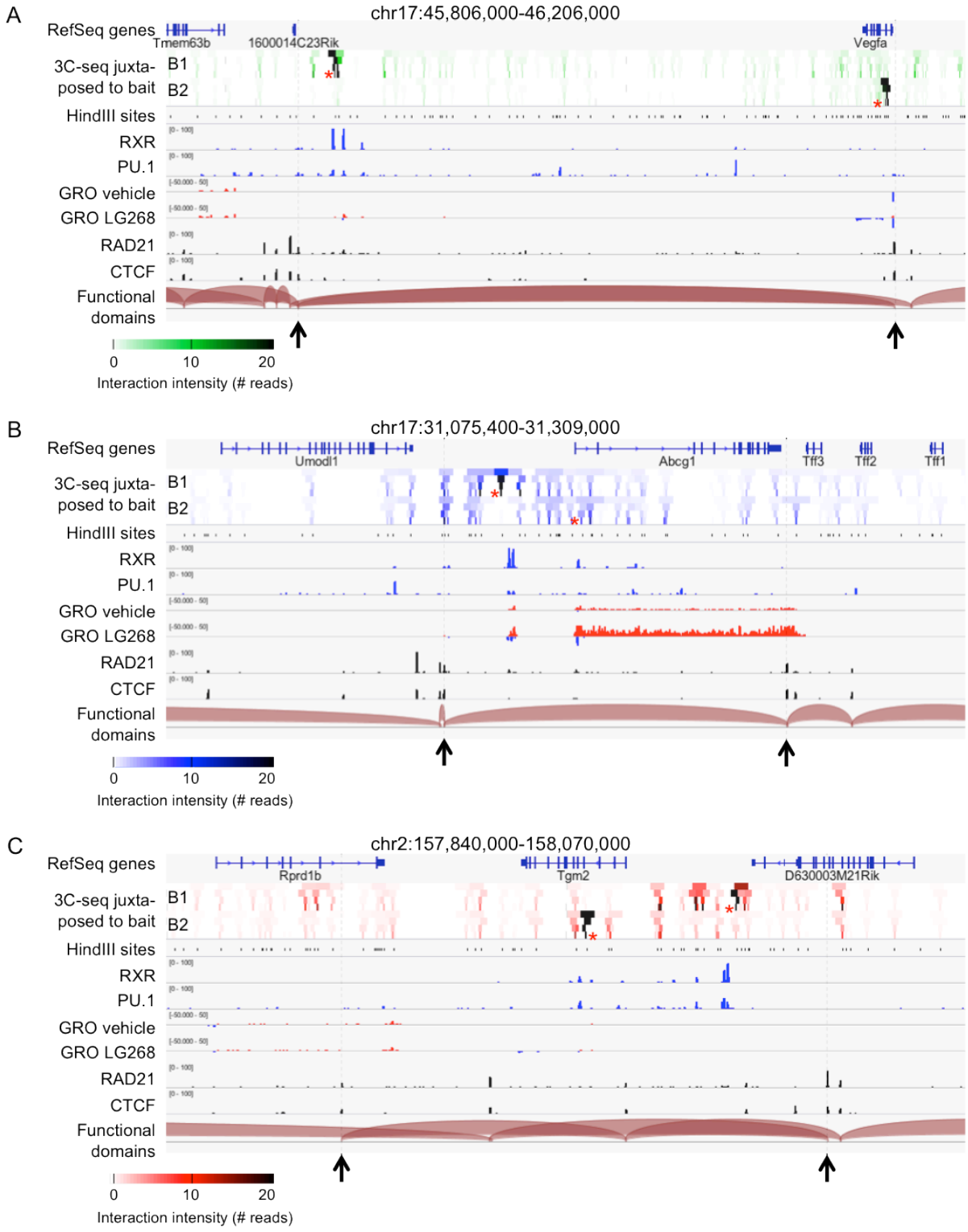


Figure 31. Functional domains under the control of RXR

The description of **Figure 31***:*

The genome browser view of the Vegfa (A), Abcg1 (B) and Tgm2 locus (C) containing the proximal interacting regions of the intergenic (B1) and intronic (B2) baits and the loop predictions generated based on CTCF/RAD21 co-bound regions (asterisks show the sites of the specific baits; GRO-seq and ChIP-seq for the indicated factors are shown; black arrowheads and gray dashed lines indicate the predicted domain borders)

As loop prediction was based on only two-dimensional ChIP-seq information, for the purpose of further confirmation, we applied a three-dimensional method, called chromosome conformation capture (3C) (Miele et al., Curr Protoc Mol Biol., 2006; Hagege et al., Nat Protoc., 2007). To validate the predicted functional domains consistent with the RXR ChIP-seq and expressional results, we carried out 3C-seq (Stadhouders et al., Nat Protoc., 2013) for some selected putative regulatory units. This method gives all interaction (target) sites genome-wide for a chosen region (called bait), so two relatively distant baits for a domain provide a good control to one another. We used bait pairs located in the first intron of Vegfa and Abcg1, and the third intron of Tgm2 gene (B1), and a respective distal enhancer (B2) within the given functional domain (**Figure 31**). We chose Vegfa for several reasons: i) it was induced upon RXR activation; ii) its known upstream regulators were also induced in this system; iii) based on our results, it seemed that Vegfa had an unusually distant downstream regulatory region; and iv) angiogenesis is one of the most studied topic related to e.g. development and cancer formation. ATP-binding cassette subfamily G member 1 (Abcg1) – together with Abca1 (Repa et al., Science, 2000) – is a well-known LXR target in macrophages (Venkateswaran et al., J Biol Chem., 2000); and now we had the tools to reveal its controlling sites. The regulation of Tgm2 is dependent both on RAR and LXR in macrophages, and it is also related to angiogenesis (**Figure 28**); this is why we chose this as our third model gene (Rébé et al., Circ Res., 2009).

For the Vegfa locus, we predicted an almost 300 kb loop between the promoter proximal upstream and a very distant downstream insulator (**Figure 31A**). Based on the target sequences, bait 2 really showed interactions in the proximity of the gene, and beside the several common targets, bait 1 could also find the neighboring restriction site of bait 2 (**Figure 31B**). At the Abcg1 locus, we saw a more "classical" case, a "gene loop", in which the regulatory elements located upstream or intronic compared to the gene (**Figure 31C**). In this ~100 kb domain, most of the interacting regions of the two baits were common and covered the regulatory regions. The Tgm2 locus had both kinds of the previous loops: a gene loop and its "reflection" similar to the one of Vegfa. Based on the insulator binding, the regulatory regions and the 3C-seq enrichments, it seems that the Tgm2 gene loop had a dominant regulatory function over the other one, as the downstream signals were much smaller. As a result, we could show that chromatin interactions were typically inside the predicted loops enriched near the regulatory regions.

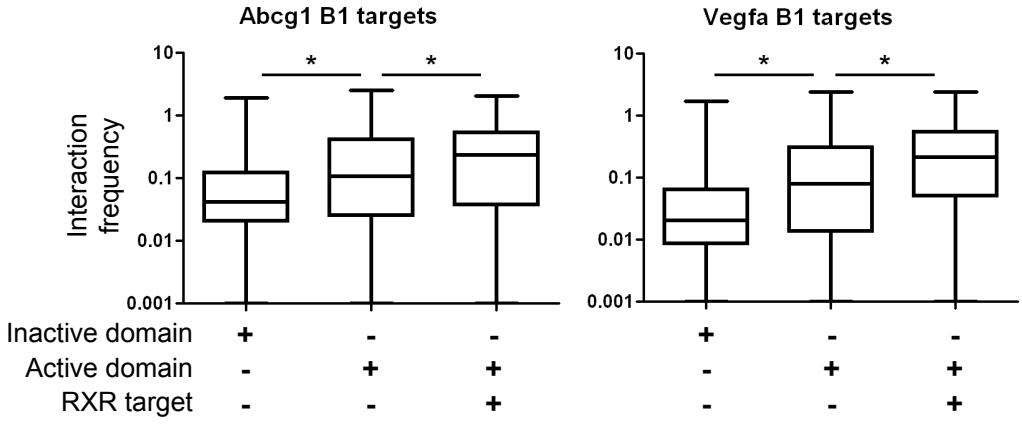


Figure 32. RXR bound active domains show higher interdomain interaction frequency with each other than with other regions

The description of **Figure 32***:*

The distribution of interaction frequency of the Abcg1 B1 (chr17: 31,172,998-31,173,656) and Vegfa B1 enhancer (chr17: 45,890,060-45,890,829) determined by 3C-seq (Unpaired two-tailed t-test analysis was used to determine significant differences; asterisk represents significant difference at $p < 0.0001$)

Finally, we examined the farther interaction points whether there were enrichments suggesting the presence of "transcription factories". Theoretically, nuclear protein foci can be formed where numerous transcriptional complexes act together on numerous regulatory sites throughout the genome. But our bait pairs never found the same distant targets, and the long-range and interchromosomal interactions showed much less frequency compared to the proximal ones. In order to provide statistical context to these findings, we compared the interaction frequency of our baits with inactive topological domains, active domains devoid of RXR regulated regions, and active domains with RXR regulated regions (excluding the loops of the given genes). As shown in **Figure 32**, in case of the intronic baits of Abcg1 and Vegfa, there was significant difference between the frequencies of these kinds of interactions, and similar results were obtained with the enhancer specific baits and those of Tgm2. This suggested that the RXR-bound active enhancers interacted with other active genomic regions and with a higher likelihood with other RXR regulated ones; however this still was a rare event.

5. Discussion

During processing our NGS data, beside the NFR prediction (**Figure 6**), we developed a pipeline to process GRO-seq data, as well (**Figure 16A**). This included the determination and annotation of the different kinds of nascent RNA transcripts. As there was an overlap between the short divergent and elongated transcripts, we distinguished these types of polymerase activities. Thus we could detect more than 50,000 active regulatory sites, 11,235 known genes and 12,821 yet unknown transcripts, which means that fifth of the mouse genome showed transcription on at least one strand. Together with RNA-seq data, upon RXR activation, we could observe the dynamics of the mRNA maturation of different genes, and e.g. identify GFY as Ronin (THAP11) in our experimental system (**Figure 20**). We developed a further pipeline to determine chromatin domains bordered by insulators, and finally one for the 3C-seq analysis to validate some domains. These methods helped us to map the nascent and matured transcriptome and the regulatory network of macrophages also with regard to the RXR specific functions.

The general TFs of TFII complex and those binding in CpG-rich promoters are essential for the development and maintenance of each cell. The differentiation of distinct cell types is driven by another special group of TFs termed as master regulators. Some of these, the so-called pioneer factors are able to loosen compacted chromatin. Forkhead (FOX) proteins by superseding linker histones can open up the DNA (Cirillo et al., EMBO J., 1998), and PU.1 can likewise liberate the linker DNA between the nucleosomes (Ghisletti et al., Immunity, 2010; Heinz et al., Mol. Cell, 2010). In macrophages, the main, also dimerizing partner of PU.1 is IRF8 (Eklund et al., J Biol Chem., 1998), but C/EBP alpha is similarly determinative as its over-expression could trans-differentiate pre-B cells into macrophages (Di Tullio et al., Proc Natl Acad Sci U S A., 2011). AP-1 proteins can also join to PU.1 in macrophages (Mancino et al., Genes Dev., 2015), but as there are several members of this

protein group, it is hard to tell what the partner of e.g. JUN is in these complexes. RUNX1, which plays role during macrophage differentiation (Lichtinger et al., EMBO J., 2012), is still present in the matured cells (Heinz et al., Mol. Cell, 2010). The further TFs of macrophages are probably more specific for the different functions, the general, e.g. metabolic pathways or those sensing the environmental signals.

TFs by acting together with co-regulators arrange the required chromatin environment, loop the DNA and recruit PIC on the promoters. This chromatin environment means different epigenetic modifications on promoters and enhancers. To test this, we developed a method to determine all regulatory sites surrounded by active histone marks (**Figure 6**), and found high correlation between the promoter (H3K4me3) and enhancer specific (H3K4me2, H4ac, H3K27ac) modifications (**Figure 9B**). The most frequent NFR length determined for enhancer marks was between 125 and 150 bp (**Figure 9A**), which approximately equals with the length of the DNA wound on one nucleosome core. This means that it managed to set feasible parameters for the prediction. The most active NFRs looked promoters, as these were typically broad and marked with all modifications examined, including H3K4me3. The co-existence of H3K4me2 and me3 at the same location was not surprising as these represent the same protein with different methylation rate. The presence of enhancer marks on promoters was in turn more interesting, as the intensity of their signal was similarly high as those detected with much lower H3K4me3 coverage. Thus "active enhancer" marks were specific for all active regulatory regions and H3K4me3 became the best histone mark enabling the separation of promoters from enhancers.

The motif enrichment analyses also confirmed the previous findings (**Figure 10A**), and by narrowing the regions to the most active enhancers, we got some unexpected motifs, as well (**Figure 13**). Beside the motifs of the known lineage determining factors, we got those of MiT and MEF2 protein families, which indeed had been related to macrophages (Walsh et

al., Gene, 2003; Karlström et al., Exp Hematol., 2011; Park et al., Mol Immunol., 2002; Aude-Garcia et al., Biochem J., 2010). Based on the gene expressional data, probably TFE3 dominated on the M-boxes, and beside the MEF2D, an uncharacterized MEF2 protein bound the MADS-boxes (**Figure 24**). In the case of TRE/CRE binding proteins, it was not easy to determine the different heterodimers as all the 16 AP-1/CREB families were represented with at least one member (**Figure 22**). There were several partners of JUN expressed, and ATF4 seemed to occupy the CREs with the very same heterodimerizing partners as of JUN. The members of the lineage determining Cebp family were all expressed, but of course Cebpa and Cebpb showed the highest mRNA level (**Figure 24**). From the Irf family, most genes showed high expression with the exception of the lymphocyte specific Irf4 and the keratinocyte specific Irf6 (**Figure 24**). From the Runx family, only Runx1 was detectable.

Pu.1 showed an extremely high mRNA level (**Figure 19A**), which explains the large amount of occupied EBSs including the PU-boxes and the promoter specific ETS elements. Half of the PU.1 bound sites seemed inactive or poised regulatory regions lacking the typical histone patterns (**Figure 11A,C-D**). Interestingly, the other half showed the very same pattern of H4ac, H3K27ac and H3K4me2 in average, which suggested that the presence of any of these histone marks could distinguish the active regulatory sites from inactive ones. These PU.1 peaks were significantly larger than the inactive ones, which indicated a higher binding frequency at active regulatory sites. Co-binding analyses showed that PU.1 co-operated with both the promoter and macrophage specific TFs (**Figure 12**). The smaller PU.1 enrichment on the promoter specific EBSs compared to the enrichment on PU-boxes might be due to the different affinity to these elements or the competition with other ETS proteins for these sites. But the presence of PU.1 was clear, as all the main promoter specific elements showed the proximity of PU.1.

The lack of NR motif enrichment in NFRs was probably a technical issue because the role of NRs is known in macrophages (Nagy et al., Physiol Rev., 2012); nevertheless it indicated that these TFs are not lineage determining but have fine-tuning roles in BMDM cells. Beside Rxra and Rxrb, we found Lxrb highly expressed, and from class II, Ppard, Rara/g and Thra showed higher expression (**Figure 25**). From the other classes, Gr, Esrra and Ear2 showed yet remarkable expression level. This indicates that RXR heterodimers dominate over the other NRs; and indeed, RXR bound to more than 5,000 regions (**Table 8**), and its binding frequency was typically elevated upon treatment with its LG268 agonist (**Figure 14A-B**). This enrichment seemed to be associated to activator functions, as the co-activator P300 also followed its enrichment (**Figures 14F,I and 30B**). What is more, more than 2/3 of the P300 bound regions were occupied also by RXR (**Figure 14C**), which might indicate a closer, even direct interaction between these proteins. In contrast, PU.1 showed rather lower binding frequency upon RXR activation (**Figure 14D-E**).

GRO-seq, by showing the direct regulatory effects both at the level of enhancer and gene transcription, was the most suitable method to follow expressional changes upon LG268 treatment. According to the nascent RNA expression, almost 2-times more divergent sites were significantly induced than repressed, and this ratio was close to 15 for the RXR-bound regions, which meant 673 directly induced regulatory sites (**Figure 27A**). Surprisingly, the number of significantly repressed genes (423) exceeded the number of the induced genes (318) (**Figure 27C**); however the average extent of induction was ~2.8-times greater than the one of the repression. Finally, based on proximity we assigned 387 enhancers and 27 silencers to 226 up and 26 down regulated genes, respectively. Then, by using insulator-specific CTCF/RAD21 co-peaks, we predicted functional domains, which covered 80% of the regulated genes thus further confirmed the applicability of annotation. Interestingly, RAD21 behaved as a co-activator, marking the RXR-bound regions and showing induction upon

LG268 treatment. In summary, we observed that RXR activation directly affected (mostly induced) more than 200 genes, and probably indirectly regulated (mostly repressed) further 500 genes.

More than half of the 387 RXR-bound enhancers possessed NR binding sites: DR1 was the most common, which could be bound by PPARD/RXR and RAR/RXR heterodimers if we excluded the COUP-TF gamma, TR2/4 and RXRA/B dimers, while from the DR4 elements LXRB/RXR probably superseded the THRA/RXR heterodimers (**Figures 29 and 25**). It is hard to tell which NRs could bind the further half and composite elements beside RXR, but it seemed sure that the sites without NR elements were bound indirectly by RXR e.g. through PU.1, C/EBP, AP-1, RUNX1 or other proteins. Nevertheless, the presence of RXR at enhancers with LG268 inducible transcription in the proximity of induced genes indicated that RXR may have functions distinct from those of its partners; however there were overlaps between their target genes, e.g. Tgm2 is an RAR/LXR target (Rébé et al., Circ Res., 2009), Angptl4 is a PPAR target (Yoon et al., Mol Cell Biol., 2000), while Abca1 and Abcg1 are LXR target genes (Repa et al., Science, 2000; Venkateswaran et al., J Biol Chem., 2000).

Interestingly, several, different types of TFs were also induced by RXR, which are related to angiogenesis (**Figure 28C**). Beside their other functions, KLF10 (Wara et al., Blood, 2011), FLI1, ETS2 (Morita et al., Proc Natl Acad Sci U S A., 2015; Craig et al., Arterioscler Thromb Vasc Biol., 2015; Wallace et al., PLoS One, 2013) and ATF4 have also angiogenic roles (Roybal et al., J Biol Chem., 2004). ATF4 did not, but some of its heterodimerizing partners did show induction at the RNA level upon treatment (**Figure 22**). Based on the expressional results, it seemed that ATF4 had a similar central role as of JUN(s) as these proteins does not interact with each other (Newman and Keating, Science, 2003), both showed high expression, and several common partners were also expressed. Fos – of which phosphorylated protein product is rather an activator – and Jdp2 – of which protein

product is rather a repressor – were simultaneously induced by LG268 treatment. As Jdp2 is a much longer gene, its effect is shifted in time, which might cause a transient FOS/JUN and FOS/ATF4 effect (**Figure 23**).

And indeed, numerous angiogenic genes were induced upon RXR activation. Not only the ATF4 target Vegfa, the RAR/LXR target Tgm2 and the PPAR target Angptl4, but also the heparin-binding epidermal growth factor-like growth factor (Hbegf), the heme oxygenase 1 (Hmox1), the thrombospondin receptor (Cd36) and several other genes playing role in blood vessel formation (**Figure 28A**). This seemed a coordinated cooperation driven by RXR and its downstream regulators. The most studied and maybe most important angiogenic gene, Vegfa seemed to have an unusually distant enhancer group ("super-enhancer") bound and probably regulated by RXR (**Figure 31A**). Both the induction of enhancer transcription and the close to 300 kb domain (predicted based on the insulator usage) indicated that we found the major regulatory sites of Vegfa in mouse BMDM cells, so this finding was further validated by 3C-seq, which showed interaction between the gene and the discovered downstream region (**Figure 31A**). Tgm2 did not show such extremity in distances rather a simple gene loop, in which the regulatory sites may stabilize similar size of DNA fingers in the 3-dimension structure (**Figure 31C**).

6. Keywords / kulcsszavak

Macrophage, transcription factors, nucleosome-free region, PU.1, RXR, angiogenesis

Makrofág, Transzkripciós faktorok, nukleoszómamentes régió, PU.1, RXR, érképzés

7. Summary

To examine the transcriptional regulation of BMDM cells, we used several NGS methods including ChIP-seq, GRO-seq, RNA-seq and 3C-seq. For data processing, as there is a large amount of tools but still no widely used standard for a significant part of the analyses, we needed new approaches and the development of pipelines to answer the more or less specific questions. GRO-seq data provided diverse information, which needed a complex system for transcript prediction, annotation and the calculation of expressional levels. In this pipeline, we incorporated the data derived from H3K4me3 ChIP-seq to distinguish the promoters from enhancers thus the gene transcripts from enhancer transcripts, respectively. We used a novel pipeline for NFR and domain prediction from ChIP-seq data and one also for the 3C-seq data analysis.

The major TF families affecting in macrophages were determined based on their motifs enriched from the predicted NFRs, but the identification of the individual TFs needed gene expressional data. We were interested in the direct effects of RXR ligandation, and found that RXR – with or without its heterodimerizing partners – acted rather as an activator. It showed P300 and RAD21 enrichment, which both might prepare the chromatin environment for gene induction. It induced several angiogenic genes (e.g. Vegfa, Angptl4, Tgm2 and Hbegf) and TFs (ATF4, ETS2, KLF10 and FLI1) also involved in angiogenesis. FOS and JDP2, the heterodimerizing partners of JUNs and ATF4 were also regulated, which might cause a transient induction. For the most important angiogenic gene, Vegfa, we found an unusually distant group of enhancers, which was associated to the gene by a 300 kb loop of which ends indeed showed interaction based on the 3C-seq data.

Összefoglalás

A BMDM sejtek transzkripciós szabályozásának vizsgálatára számos NGS módszert használtunk, beleértve a ChIP-seq-et, GRO-seq-et, RNA-seq-et és 3C-seq-et. Mivel nagy mennyiségű eszköz van, de az elemzések jelentős részéhez még mindig nincs széleskörben használt standard, az adatfeldolgozáshoz új megközelítésekre és új stratégiák kidolgozására volt szükség a többé-kevésbé egyedi kérdések megválaszolására. A GRO-seq szerteágazó információt biztosított, amely egy összetett rendszert igényelt a transzkriptumok meghatározására, annotálására és az expressziós szintek kiszámítására. A promóterek enhanszerektől való, tehát a gének enhanszer transzkriptumoktól való elválasztásához beépítettük a műveletek közé a H3K4me3 ChIP-seq-ből származó adatokat is. Új stratégiát alkalmaztunk az NFR- és doménpredikciókra ChIP-seq adatokból, és a 3C-seq elemzéshez is.

A makrofágokban működő főbb TF családokat a prediktált NFR-ek motívum-feldúsulásai alapján határoztuk meg, de az egyes TF-ok azonosítása génexpressziós adatokat igényelt. Érdekeltek az RXR ligandkezelés közvetlen hatásai, és azt találtuk, hogy az RXR – a heterodimerizáló partnereivel vagy azok nélkül – inkább aktivátorként működött. P300 és RAD21 feldúsulást mutatott, melyek előkészíthetik a kromatinkörnyezetet a génindukcióra. Számos angiogenikus gént (pl. Vegfa, Angptl4, Tgm2 és Hbegf) és szintén érképzésben érintett TF-t (ATF4, ETS2, KLF10 and FLI1) indukált. A FOS and JDP2, a JUN-ok és ATF4 heterodimerizáló partnerei szintén szabályozódtak, amely egy átmeneti indukciót okozhatott. A legfontosabb angiogenikus génhez, a Vegfa-hoz egy szokatlanul távoli enhanszercsoportot találtunk, amely egy 300 kb-os hurkon keresztül közelíti meg a gént, mely huroknak a végei valóban kölcsönhatást mutattak a 3C-seq adatok alapján.

8. Abbreviations

(E)P300: E1A protein 300 kDa

3C: Chromosome Conformation Capture

3C-seq: 3C sequencing

Abca1: ATP-binding cassette subfamily A member 1

Abcg1: ATP-binding cassette subfamily G member 1

AICE: AP-1/IRF composite element

Angptl4: Angiopoietin-like 4

AP-1: Activator protein 1

AR: Androgen receptor

ASV 17: Avian sarcoma virus 17

ATAC-seq: Assay for transposase accessible chromatin coupled with NGS

ATF: Activating transcription factor

BACH: BTB and CNC homology (protein)

BAI: binary alignment/map index

BAM: binary alignment/map

BATF: bZIP transcription factor, ATF-like

BED: browser extensible data

bHLH-ZIP: basic helix-loop-helix leucine zipper

BLAST: basic local alignment search tool

BLAT: BLAST-like alignment tool

BMDM: bone marrow derived macrophage

BRD: bromodomain

BRE: B recognition element

BTB: broad complex-tramtrack-bric-a-brac (domain)

BWA: Burrows-Wheeler alignment (tool)

BWT: Burrows-Wheeler transformation

bZIP: basic leucine zipper

C/EBP: CCAAT-box and enhancer-binding protein

CAGE: cap analysis of gene expression

CAK: CDK-activating kinase

CAR: Constitutive androstane receptor

CBF: CCAAT-binding factor

CBP: CREB binding protein

CDK: Cyclin-dependent kinase

cDNA: complementary DNA

C-ERBA: Cellular avian erythroblastosis virus A

ChIP: Chromatin immunoprecipitation

ChIP-chip: ChIP coupled with chip

ChIP-exo: ChIP-seq coupled with exonuclease treatment

ChIP-seq: ChIP coupled with NGS

Chop: C/EBP-homologous protein

CITCO: 6-(4-chlorophenyl)imidazo(2,1-b)(1,3)thiazole-5-carbaldehyde O-(3,4-dichlorobenzyl)oxime

CLD: cytoplasmic localization domain

CLR: C-type lectin receptor

CNC: Cap'n'collar-type (bZIP protein)

CoREST: REST co-repressor

COUP-TF: Chicken ovalbumin upstream promoter-transcription factor

CRE: cAMP-response element

CREB: CRE binding (protein)

CREB3L: CREB3-like

CREBL2: CRE-binding protein-like 2

CRE-BP1: CRE binding protein 1

CREBRF: CREB3 regulatory factor

CREBZF: CREB/ATF bZIP transcription factor

CREM: CRE modulator

CSAR: ChIP-seq analysis in R

CTCF: CCCTC-binding factor

CTD: carboxy-terminal domain

DAD: deacetylase activation domain

DAX1: Dosage-sensitive sex reversal-adrenal hypoplasia congenital critical region on the X chromosome, gene 1

DBD: DNA-binding domain

DC: dendritic cell

DCE: downstream core element

Ddit3: DNA-damage-inducible transcript 3

DiffBind: differential binding analysis of ChIP-seq peak data

DMEM: Dulbecco's modified eagle medium

DNase-seq: the sequencing of fragments generated by DNase

DPE: downstream promoter element

DR: direct repeat (element)

DSG: di(N-succinimidyl) glutarate

E1A: Early-region 1A

EAR2: v-ErbA-related protein 2

EBI: European Bioinformatics Institute

E-box: enhancer box

EBS: ETS binding site

EcR: Ecdysone receptor

EHF: ETS homologous factor

EICE: ETS-IRF composite element

ELF: ETS-like factor

ELG/ELK: ETS-like genes

EMBL: European Molecular Biology Laboratory

EMBOSS: European Molecular Biology Open Software Suite

ENCODE: Encyclopedia of DNA Elements

ER: endoplasmic reticulum

ER: everted repeat (element)

ER/ESR: estrogen receptor

ER71: ETS-related protein 71

ERF: ETS2 repressor factor

ERG: ETS related gene

ERR/ ESRR: Estrogen related receptor

ESE: Epithelium specific ETS

ETS: E26, erythroblast transformation-specific (protein)

ETV: ETS variant

ETV3L: ETS variant 3-like

EWG: Erect wing gene

FAIRE: formaldehyde-assisted isolation of regulatory elements

FAIRE-seq: FAIRE coupled with NGS

FBJ: Finkel-Biskis-Jinkins (virus)

FEV: Fifth Ewing variant

FLI1: Friend leukemia integration 1

FOS: FBJ murine osteosarcoma viral oncogene homolog, p55, phosphoprotein 55 kDa

FOSL: FOS-like

FOX: Forkhead box

FPKM: fragments per kb of exon per million mapped fragments

FRA: FOS-related antigen

FXR: Farnesoid X receptor

GABP: GA repeat binding protein

GCNF: Germ cell nuclear factor

GFF: general feature format

GFX: General factor X

GFY: General factor Y

GR: Glucocorticoid receptor

GRO-seq: global run-on sequencing

GTF: general transfer format

H3K27ac: Histone H3 lisine 27 acetylation

H3K4me2: Histone H3 lisine 4 dimethylation

H3K4me3: Histone H3 lisine 4 trimethylation

H4ac: Histone H4 acetylation

HAT: histone acetyltransferase

Hbegf: Heparin-binding epidermal growth factor-like growth factor

HDAC: histone deacetylase

HDM: histone demethylase

HFM: histone fold motif

HLH: helix-loop-helix

Hmox1: Heme oxygenase 1

HMT: histone methyltransferase

HNF4: Hepatocyte nuclear factor 4

HOMER: Hypergeometric optimization of motif enrichment

HTH: helix-turn-helix

IAD: IRF-associated domain

IFN: Interferon

IgG: Immunoglobulin G

IGV: Integrative Genomics Viewer

Inr: Initiator

IP: immunoprecipitation

IR: inverted repeat (element)

IRF: IFN regulatory factor

ISRE: IFN-stimulated response element

JDP: JUN dimerization protein

JUN: ASV17 protooncogene, 17: ju-nana (Japanese)

KAT3: lysine acetyltransferase 3

KIX: kinase-inducible domain interacting (domain)

KLF: Krüppel-like factor

KNRL: KNIRPS-like

KO: knock-out

LG268: LG100268

LBD: ligand-binding domain

lgMAF: large MAF

LPS: lipopolysaccharide

LRF: Luman recruitment factor

LRH1: Liver receptor homolog-1

LXR: Liver X receptor

MACS: Model-based analysis of ChIP-seq

MAD: MAX dimerization protein

MADS: MCM1, AGAMOUS, DEFICIENS, SRF (family)

MAF: Musculoaponeurotic fibrosarcoma (protein)

MARE: MAF-recognition element

MAST: Motif alignment search tool

MAX: MYC-associated factor X

MCM1: Maintenance of minichromosome 1

MCSF: Macrophage colony-stimulating factor

MED1: Mediator 1

MEF2: Myocyte enhancer factor 2

MEGA: Molecular evolutionary genetics analysis

MEME: Multiple expectation-maximization for motif elicitation

MGA: MAX gene associated (protein)

MGI: Mouse Genome Informatics (database)

MiT: MITF/TFE

MITF: Microphthalmia transcription factor

MLX: MAX-like protein X

MLXIP: MLX interacting protein

MNase: micrococcal nuclease

MNase-seq: MNase sequencing

MNT: MAX network transcriptional repressor

MR: Mineralcorticoid receptor

mRNA: messenger RNA

MTE: motif ten element

MUSCLE: Multiple sequence comparison by log-expectation

MXD: MAX dimerization protein

MXI1: MAX interactor 1

MYC: avian Myelocytomatosis viral oncogene

NCoR: Nuclear receptor co-repressor

ncRNA: non-coding RNA

NFE2: Nuclear factor erythroid 2

NFE2L: NFE2-like

NFR: nucleosome-free region

NFY: Nuclear transcription factor Y

NGFI-B: Nerve-growth-factor-induced gene

NGS: next-generation sequencing

NLR: NOD-like receptor

NLS: nuclear localization signal

NOD: nucleotide-binding oligomerization domain

NOR: nucleosome occupied region

NOR1: Neuron-derived orphan receptor 1

NR: nuclear receptor

NRC: nuclear receptor co-activator interlocking (domain)

NRF: NFE2-related factor

NRF1: Nuclear respiratory factor 1

NRL: Neural retina leucine zipper protein

NS: nucleosome

NURD: Nucleosome remodeling deacetylase

NURR1: Nur-related factor 1

NURR77: Nuclear hormone receptor 77

OBF: Open Bioinformatics Foundation

PAGE: polyacrylamide gel electrophoresis

PAMP: pathogen-associated molecular pattern

PCR: polymerase chain reaction

PDEF: Pointed domain containing ETS transcription factor

PEA3: Polyomavirus enhancer activator 3

PHD: plant homeodomain

PIC: pre-initiation complex

PNK: polynucleotide kinase

PNR: Photoreceptor-cell-specific nuclear receptor

PNT: pointed (domain)

Pol II: RNA polymerase II

PPAR: Peroxisome proliferator-activated receptor

PPARGC1: PPARg coactivator 1

PPRC1: PPARg coactivator 1

PR/PGR: Progesterone receptor

PRR: pattern-recognition receptor

PU.1: Purine-rich nucleic acid binding protein 1

PUFA: polyunsaturated fatty acid

PXR: Pregnane X receptor

QPCR: quantitative PCR

RACE: rapid amplification of 5' cDNA ends

RAID: redundant array of independent disks

RAR: Retinoic acid receptor

RE: response element

REST: RE1-silencing transcription factor

REV-ERB: Reverse-ERB

RIG-I: Retinoic acid inducible gene-I

RLR: RIG-I-like receptor

RNA-seq: RNA sequencing

ROR: RAR-related orphan receptor

RPB1: RNA polymerase II subunit B1

RPKM: reads per kb of a region per million mapped reads

rRNA: ribosomal RNA

RUNX: Runt-related transcription factor

RXR: Retinoid X receptor

SAM: sterile alpha motif (domain)

SAM: sequence alignment/map (file)

SAS: serial attached SCSI (disk)

SCSI: small computer system interface

SD: standard deviation

SF-1: Steroidogenic factor 1

SFFV: Spleen focus forming virus

SHP: Short heterodimeric partner

SICER: Spatial clustering for identification of ChIP-enriched regions

SID: SIN3 interaction domain

SIN3: Switch independent 3

SIN3A: Switch independent 3 analogue A

SIR: Silent information regulator

SIRT1: Sirtuin 1

smMAF: small MAF

SMRT: Silencing mediator for retinoid or thyroid-hormone receptors

SOAP: Short oligonucleotide alignment program

Sono-seq: Sonication of cross-linked chromatin sequencing

SP1: Specificity protein 1

SPDEF: SAM pointed domain containing ETS transcription factor

SPI: SFFV proviral integration (oncogene)

SRA: Sequence Read Archive

SRF: Serum response factor

STAF: Selenocysteine tRNA gene transcription activating factor

SXR: Steroid X receptor

T3: triiodothyronine

TAD: trans-activation domain

TAF: TATA associated factor

TAL1: T-cell acute lymphocytic leukemia protein 1

TAP: tobacco acid pyrophosphatase

TAZ: transcription adaptor putative zinc finger (domain)

TBP: TATA-binding protein

TBPL1-2: TBP-like factors 1-2

TCF: Ternary complex factor

TCPOBOP: 1,4-bis[2-(3,5-dichloropyridyloxy)]benzene

TDF: tiled data file

TEL: Translocation E26 transforming-specific leukaemia (gene)

TF: transcription factor

TFBS: transcription factor binding site

TFE: Transcription factor E

TFIIA-H: Transcription factor II A-H

Tgm2: Transglutaminase 2

THAP: Thanatos (Greek death god) associated protein

TLR: Toll-like receptor

TLX: Tailless homolog orphan receptor

TPA: phorbol 12-O-tetradecanoate 13-acetate

TR/THR: Thyroid hormone receptor

TR2,4: Testicular orphan receptor 2,4

TRE: TPA response element

tRNA: transfer RNA

TSR: transcription start region

TSS: transcription start site

TSV: tab separated value

TTS: transcription termination sites

UCSC: University of California, Santa Cruz

VDR: Vitamin D receptor

VEGFA: Vascular endothelial growth factor A

WIG: wiggle (file format)

XBP1: X-box binding protein 1

xONR1: *Xenopus* Orphan nuclear receptor 1

YY1: Yin-yang 1

ZBTB33: Zinc-finger and BTB domain containing 33 (gene)

ZFP: Zinc-finger protein

ZINBA: Zero inflated negative binomial (peak caller)

ZNF: Zinc-finger factor

Registry number: DEENK/67/2016.PL
Subject: Ph.D. List of Publications

Candidate: Gergely Nagy
Neptun ID: AOKH5U
Doctoral School: Doctoral School of Molecular Cell and Immune Biology

List of publications related to the dissertation

1. Dániel, B., **Nagy, G.**, Hah, N., Horváth, A., Czimmerer, Z., Póliska, 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

2. **Nagy, G.**, Dániel, B., Jonás, D., Nagy, L., Barta, E.: A novel method to predict regulatory regions based on histone mark landscapes in macrophages.
   Immunobiology. *218* (11), 1416-1427, 2013.
   DOI: http://dx.doi.org/10.1016/j.imbio.2013.07.006
   IF:3.18

List of other publications

3. Cuaranta-Monroy, I., Simándi, Z., Kolostyák, Z., Doan-Xuan, Q., Póliska, 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

4. Dániel, B., **Nagy, G.**, Nagy, L.: The intriguing complexities of mammalian gene regulation: How to link enhancers to regulated genes. Are we there yet?
FEBS Lett. 588 (15), 2379-2391, 2014.
DOI: http://dx.doi.org/10.1016/j.febslet.2014.05.041
IF:3.169

5. Széles, L., Póliska, S., **Nagy, G.**, Szatmári, I., Szántó, A., Pap, A., Lindstedt, M., Santegoets, S.J.A.M., Rühl, R., Dezső, B., Nagy, L.: Research resource: Transcriptome profiling of genes regulated by RXR and its permissive and nonpermissive partners in differentiating monocyte-derived dendritic cells.
Mol. Endocrinol. 24 (11), 2218-2231, 2010.
DOI: http://dx.doi.org/10.1210/me.2010-0215
IF:4.889

Total IF of journals (all publications): 25,729
Total IF of journals (publications related to the dissertation): 13,978

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of Web of Science, Scopus and Journal Citation Report (Impact Factor) databases.

18 March, 2016

DEBRECENI EGYETEM KENÉZY ORVOSTUDOMÁNYI EGYETEMI ÉS NEMZETI KÖNYVTÁRA

9. Acknowledgements

I would like to thank my supervisor Dr. Endre Barta for introducing me into the world of bioinformatics.

This thesis would not have been possible without my former supervisor and present chief Prof. László Nagy.

I would like to thank Dr. Bence Dániel for providing me a huge amount to good quality NGS data to work with.

I am greatful to Prof. László Fésüs and Prof. József Tőzsér, the former and recent heads of the Department of Biochemistry and Molecular Biology for the opportunity to work in a well-equipped environment.

I am thankful to all the past and present members of the Nuclear Receptor Research Group for their help and scientific discussions.

Special thanks to Dr. Szilárd Póliska, Attila Horváth and Dávid Jónás.

I also wish to express my gratitude to Dóri and my family for their support and patience.

10. References

1. Achatz et al., Mol Cell Biol., 1997 Sep;17(9):4914-32.
2. Adhikary et al., PLoS One, 2011 Jan 19;6(1):e16344.
3. Ai et al., J Biol Chem., 2004 Mar 5;279(10):8684-93.
4. Akhtar and Veenstra, Cell Biosci., 2011 Jun 27;1(1):23.
5. Albert et al., Nature, 2007 Mar 29;446(7135):572-6.
6. Allfrey et al., Proc Natl Acad Sci U S A., 1964 May;51:786-94.
7. Allis et al., Cell, 2007 Nov 16;131(4):633-6.
8. Altschul et al., J Mol Biol., 1990 Oct 5;215(3):403-10.
9. Andersson and Scarpulla, Mol Cell Biol., 2001 Jun;21(11):3738-49.
10. Aravantinou-Fatorou et al., Stem Cell Reports, 2015 Sep 8;5(3):405-18.
11. Aronheim et al., Mol Cell Biol., 1997 Jun;17(6):3094-102.
12. Audas et al., Mol Cell Biol., 2008 Jun;28(12):3952-66.
13. Aude-Garcia et al., Biochem J., 2010 Sep 1;430(2):237-44.
14. Auerbach et al., Proc. Natl. Acad. Sci. U S A., 2009 Sep 1;106(35):14926-31.
15. Ayer et al., Cell, 1993 Jan 29;72(2):211-22.
16. Baar et al., FASEB J., 2003 Sep;17(12):1666-73.
17. Baes et al., Mol Cell Biol., 1994 Mar;14(3):1544-52.
18. Bailey et al., J. Steroid Biochem. Mol. Biol., 1997 May;62(1):29-44.
19. Bailey et al., PLoS Comput Biol., 2013;9(11):e1003326.
20. Balwierz et al., Genome Biology, 2009;10(7):R79.
21. Banerji et al., Cell, 1981 Dec;27(2 Pt 1):299-308.
22. Baranello et al., Transcription, 2013 Sep-Dec;4(5):232-7.
23. Barish et al., Genes Dev., 2010 Dec 15;24(24):2760-5.
24. Barish et al., Mol Endocrinol., 2005 Oct;19(10):2466-77.
25. Barta, EMBnet.Journal, 2011
26. Bedford et al., Epigenetics, 2010 Jan 1;5(1):9-15.
27. Billin et al., J Biol Chem., 1999 Dec 17;274(51):36344-50.
28. Billin et al., Mol Cell Biol., 2000 Dec;20(23):8845-54.
29. Blackwood and Eisenman, Science, 1991 Mar 8;251(4998):1211-7.
30. Blumberg et al., Genes Dev., 1998 Oct 15;12(20):3195-205.
31. Bohmann et al., Science, 1987 Dec 4;238(4832):1386-92.
32. Boyle et al., Cell, 2008 Jan 25;132(2):311-22.
33. Brass et al., Genes Dev., 1996 Sep 15;10(18):2335-47.
34. Bucher, J. Mol. Biol., 1990 Apr 20;212(4):563-78.
35. Buenrostro et al., Nat Methods, 2013 Dec;10(12):1213-8.
36. Burrows and Wheeler, Digital Equipment Corporation, 1994
37. Carroll et al., Cell, 2005 Jul 15;122(1):33-43.
38. Chan et al., Proc. Natl. Acad. Sci. U S A., 1993 Dec 1;90(23):11366-70.
39. Chang et al., J Biol Chem., 2004 Sep 1;2(3):341-350.
40. Chen et al., Mol Cell Proteomics, 2007 May;6(5):812-9.
41. Chen et al., Mol Endocrinol., 1994 Oct;8(10):1434-44.
42. Chen et al., Nat Struct Mol Biol., 2008 Nov;15(11):1213-20.
43. Choi et al., J Biol Chem., 1997 Sep 19;272(38):23565-71.

44. Chrivia et al., Nature, 1993 Oct 28;365(6449):855-9.

45. Cirillo et al., EMBO J., 1998 Jan 2;17(1):244-54.

46. Clarke et al., Am J Clin Nutr., 1999 Oct;70(4):566-71.

47. Cock et al., Nucleic Acids Research, 2010 Apr;38(6):1767-71.

48. Comb et al., Nature, 1986 Sep 25-Oct 1;323(6086):353-6.

49. Conesa et al., Bioinformatics, 2006 May 1;22(9):1096-102.

50. Cooper et al., Neuromuscul Disord., 2007 Apr;17(4):276-84.

51. Core et al., Science, 2008 Dec 19;322(5909):1845-8.

52. Cowdry, J Exp Med., 1925 Aug 31;42(3):323-33.

53. Craig et al., Arterioscler Thromb Vasc Biol., 2015 Apr;35(4):865-76.

54. Croce and Nowell, Blood, 1985 Jan;65(1):1-7.

55. Curran and Franza, Cell, 1988 Nov 4;55(3):395-7.

56. Curran and Teich, J Virol., 1982 Apr;42(1):114-22.

57. Daniel and Reynolds, Mol Cell Biol., 1999 May;19(5):3614-23.

58. Date et al., J Biol Chem., 2014 Apr 11;289(15):10318-29.

59. Dawson and Xia, Biochim Biophys Acta., 2012 Jan;1821(1):21-56.

60. Dejosez et al., Cell, 2008 Jun 27;133(7):1162-74.

61. Dejosez et al., Genes Dev., 2010 Jul 15;24(14):1479-84.

62. Deng and Roberts, Chromosoma, 2007 Oct;116(5):417-29.

63. Dixon et al., Nature, 2012 Apr 11;485(7398):376-80.

64. Di Tullio et al., Proc Natl Acad Sci U S A., 2011 Oct 11;108(41):17016-21.

65. Dorn et al., Proc. Natl. Acad. Sci. U S A., 1987 Sep;84(17):6249-53.

66. Dorsey et al., Oncogene, 1995 Dec 7;11(11):2255-65.

67. Duesberg et al., Proc Natl Acad Sci U S A., 1977 Oct;74(10):4320-4.

68. Durand et al., Cell, 1992 Oct 2;71(1):73-85.

69. Dvir, Biochim Biophys Acta., 2002 Sep 13;1577(2):208-223.

70. Dynan and Tjian, Cell, 1983a Mar;32(3):669-80.

71. Dynan and Tjian, Cell, 1983b Nov;35(1):79-87.

72. Echlin et al., Oncogene, 2000 Mar 30;19(14):1752-63.

73. Eckner et al., Genes Dev., 1994 Apr 15;8(8):869-84.

74. Edgar, BMC Bioinformatics, 2004 Aug 19;5:113.

75. Edwards et al., J Lipid Res., 2002 Jan;43(1):2-12.

76. Efstratiadis et al., Cell, 1980 Oct;21(3):653-68.

77. Eklund et al., J Biol Chem., 1998 May 29;273(22):13957-65.

78. Evans and Mangelsdorf, Cell, 2014 Mar 27;157(1):255-66.

79. Evans and Scarpulla, J Biol Chem., 1989 Aug 25;264(24):14361-8.

80. Fazio et al., J Biol Chem., 2001 Jun 1;276(22):18710-6.

81. Feng et al., Proc Natl Acad Sci U S A., 2008 Apr 22;105(16):6057-62.

82. Finnin et al., Nat Struct Biol., 2001 Jul;8(7):621-5.

83. Frank et al., J Mol Biol., 2005 Feb 18;346(2):505-19.

84. Fujisawa et al., J. Biochem, 2000 Mar;127(3):373-82.

85. Gaertner and Zeitlinger, Development, 2014 Mar;141(6):1179-83.

86. Gao et al., Mol Endocrinol., 2003 Aug;17(8):1484-507.

87. Gardiner-Garden and Frommer, J Mol. Biol., 1987 Jul 20;196(2):261-82.

88. Gaulton et al., Nat. Genet., 2010 Mar;42(3):255-9.

89. Geissmann et al., Science, 2010 Feb 5;327(5966):656-61.

90. Gershenzon and Ioshikhes, Bioinformatics, 2005 Apr 15;21(8):1295-300.

91. Ghisletti et al., Immunity, 2010 Mar 26;32(3):317-28.

92. Giglioni, Biochem Biophys Res Commun., 1989 Jul 14;162(1):326-33.

93. Glasmacher et al., Science, 2012 Nov 16;338(6109):975-80.

94. Gottlieb, Can Med Assoc J., 1934 Mar;30(3):256-8.

95. Green et al., Nature, 1986 Mar 13-19;320(6058):134-9.

96. Green, Trends Biochem Sci., 2000 Feb;25(2):59-63.

97. Greschik et al., Mol Cell Biol., 1999 Jan;19(1):690-703.

98. Grosdidier et al., Mol. Endocrinol., 2012 Jul;26(7):1078-90.

99. Gruber et al., Cell, 2003 Mar 21;112(6):765-77.

100. Grünberg and Hahn, Trends Biochem Sci., 2013 Dec;38(12):603-11.

101. Guerriero et al., Blood, 2000 Feb 1;95(3):879-85.

102. Guillemette et al., PLoS Genet., 2011 Mar;7(3):e1001354.

103. Gutierrez-Hartmann et al., Trends Endocrinol Metab., 2007 May-Jun;18(4):150-8.

104. Hadjur et al., Nature, 2009 Jul 16;460(7253):410-3.

105. Hagege et al., Nat Protoc., 2007;2(7):1722-33.

106. Hagiwara et al., Biochem Biophys Res Commun., 2002 Jan 25;290(3):979-83.

107. Hah et al., Cell, 2011 May 13;145(4):622-34.

108. Hah et al., Genome Res., 2013 Aug;23(8):1210-23.

109. Hai and Curran, Proc. Natl. Acad. Sci. U S A., 1991 May 1;88(9):3720-4.

110. Hai et al., Genes Dev., 1988 Oct;2(10):1216-26.

111. Harper et al., Proc Natl Acad Sci U S A., 1996 Aug 6;93(16):8536-40.

112. Hassig et al., Cell, 1997 May 2;89(3):341-7.

113. Hay et al., J Am Chem Soc., 2014 Jul 2;136(26):9308-19.

114. Haynes et al., Nucleic Acids Res., 1992 May 25;20(10):2603.

115. Heidemann et al., Biochim Biophys Acta., 2013 Jan;1829(1):55-62.

116. Heinz et al., Mol. Cell, 2010 May 28;38(4):576-89.

117. Hildmann et al., Appl Microbiol Biotechnol., 2007 Jun;75(3):487-97.

118. Hogan et al., FEBS Letters, 2006 Jan 9;580(1):58-62.

119. Hollenberg et al., Nature, 1985 Dec 19-1986 Jan 1;318(6047):635-41.

120. Hollenhorst et al., Annu. Rev. Biochem., 2011;80:437-71.

121. Holtschke et al., Cell, 1996 Oct 18;87(2):307-17.

122. Honda and Taniguchi, Nat Rev Immunol., 2006 Sep;6(9):644-58.

123. Hong et al., Genomics Inform., 2012 Sep;10(3):145-52.

124. Hortega, 1920

125. Hossain et al., J Biol Chem., 2009 Mar 27;284(13):8621-32.

126. Hu et al., J Biol Chem., 2000 May 19;275(20):15254-64.

127. Hubbard et al., Nucleic Acids Res., 2002 Jan 1;30(1):38-41.

128. Hudson et al., J Biol Chem., 2015 Jul 17;290(29):18237-44.

129. Hurlin et al., EMBO J., 1999 Dec 15;18(24):7019-28.

130. Hurlin et al., Genes Dev., 1997 Jan 1;11(1):44-58.

131. Hurst and Jones, Genes Dev., 1987 Dec;1(10):1132-46.
132. Imataka et al., EMBO J., 1992 Oct;11(10):3663-71.
133. Ingraham et al., Nat. Genet., 2006 Nov;38(11):1335-40.
134. Inohara et al., Annu. Rev. Biochem., 2005;74:355-83.
135. Itoh et al., Nucleic Acids Res., 2015 Feb 27;43(4):2033-44.
136. Ji et al., Nat Biotechnol., 2008 Nov;26(11):1293-300.
137. Kadonaga et al., Cell 1987 Dec 24;51(6):1079-90.
138. Kadowaki et al., Biochem Biophys Res Commun., 1992 Mar 16;183(2):492-8.
139. Kahle et al., Mol Cell Biol., 2005 Jul;25(13):5339-54.
140. Karlström et al., Exp Hematol., 2011 Mar;39(3):339-350.e3.
141. Kataoka et al., Mol Cell Biol., 1994 Jan;14(1):700-12.
142. Kato et al., Genes Dev., 1992 Jan;6(1):81-92.
143. Kawai et al., Virology, 1992 Jun;188(2):778-84.
144. Kent et al., Genome Res., 2002 Jun;12(6):996-1006.
145. Kent, Genome Res., 2002 Apr;12(4):656-64.
146. Kestler et al., Bioinformatics, 2005 Apr 15;21(8):1592-5.
147. Khan et al., Proteomics, 2006 Jan;6(1):123-30.
148. Klappacher et al., Cell, 2002 Apr 19;109(2):169-80.
149. Kliewer et al., Cell, 1998 Jan 9;92(1):73-82.
150. Knuppel et al., J. Comput. Biol., 1994 Fall;1(3):191-8.
151. Koelle et al., Cell, 1991 Oct 4;67(1):59-77.
152. Kohl et al., Cell, 1983 Dec;35(2 Pt 1):359-67.
153. Kolliker, 1873
154. Kouzarides, Cell, 2007 Feb 23;128(4):693-705.
155. Kuhn and Xu, Prog Mol Biol Transl Sci., 2009;87:299-342.
156. Kupffer, 1876
157. Lagrange et al., Genes Dev., 1998 Jan 1;12(1):34-44.
158. Landry et al., Blood, 2005 Oct 15;106(8):2680-7.
159. Landschulz et al., Science, 1988 Jun 24;240(4860):1759-64.
160. Langerhans, 1868
161. Langmead et al., Genome biology, 2009;10(3):R25.
162. Lazar et al., Mol Cell Biol., 1989 Mar;9(3):1128-36.
163. Lee et al., Cell, 1987 Jun 19;49(6):741-52.
164. Lee et al., Nat. Genet., 2007 Oct;39(10):1235-44.
165. Lee et al., Nucleic Acids Res., 2010 Oct;38(18):6045-53.
166. Lee, J Cell Sci., 1992 Sep;103 (Pt 1):9-14.
167. Lefterova et al., Genes Dev., 2008 Nov 1;22(21):2941-52.
168. Leprince et al., Nature, 1983 Nov 24-30;306(5941):395-7.
169. Levine and Tjian, Nature, 2003 Jul 10;424(6945):147-51.
170. Levine et al., Cell, 2014 Mar 27;157(1):13-25.
171. Levy et al., Genes Dev., 1988 Apr;2(4):383-93.
172. Li and Durbin, Bioinformatics, 2009 Jul 15;25(14):1754-60.
173. Li et al., Bioinformatics, 2008 Mar 1;24(5):713-4.
174. Li et al., Bioinformatics, 2009 Aug 15;25(16):2078-9.
175. Li et al., Genome Res., 2008 Nov;18(11):1851-8.

176. Li et al., Nature, 2012 Oct 25;490(7421):543-6.
177. Lichtinger et al., EMBO J., 2012 Nov 14;31(22):4318-33.
178. Lifton et al., Cold Spring Harb Symp Quant Biol., 1978;42 Pt 2:1047-51.
179. Liu et al., EMBO Rep., 2015 May;16(5):654-69.
180. Loudig et al., Biochem J., 2005 Nov 15;392(Pt 1):241-8.
181. Lu and Pitha, J Biol Chem., 2001 Nov 30;276(48):45491-6.
182. Luger et al., Nature, 1997 Sep 18;389(6648):251-60.
183. Ma et al., Biochem. J., 2011 Oct 1;439(1):27-38.
184. Maekawa et al., EMBO J., 1989 Jul;8(7):2023-8.
185. Maglich et al., J Biol Chem., 2003 May 9;278(19):17277-83.
186. Maile et al., Science, 2004 May 14;304(5673):1010-4.
187. Maki et al., Proc Natl Acad Sci U S A., 1987 May;84(9):2848-52.
188. Mancino et al., Genes Dev., 2015 Feb 15;29(4):394-408.
189. Mandal et al., PNAS, 2004 May 18;101(20):7572-7.
190. Mangelsdorf et al., Cell, 1995 Dec 15;83(6):835-9.
191. Mangelsdorf et al., Nature, 1990 May 17;345(6272):224-9.
192. Meadows et al., J Biol Chem., 2007 Jan 19;282(3):1891-904.
193. Medzhitov and Janeway, Science, 2002 Apr 12;296(5566):298-300.
194. Meraro et al., J Immunol., 1999 Dec 15;163(12):6468-78.
195. Merkenschlager and Odom, Cell, 2013 Mar 14;152(6):1285-97.
196. Metchnikoff, 1882
197. Michaud et al., Genome Res., 2013 Jun;23(6):907-16.
198. Miele et al., Curr Protoc Mol Biol., 2006 May;Chapter 21:Unit 21.11.
199. Mignotte et al., Nucleic Acids Res., 1989 Jan 11;17(1):37-54.
200. Misra et al., J. Biol. Chem., 2005 Apr 15;280(15):15257-66.
201. Mitchell et al., PLoS One, 2012;7(11):e49274.
202. Mittrücker et al., Science, 1997 Jan 24;275(5299):540-3.
203. Miyamoto et al., Cell, 1988 Sep 9;54(6):903-13.
204. Morita et al., Proc Natl Acad Sci U S A., 2015 Jan 6;112(1):160-5.
205. Muiño et al., Plant Methods, 2011 May 9;7:11.
206. Müller and Herrmann, Nature, 1997 Oct 23;389(6653):884-8.
207. Murray, Biochemistry, 1964 Jan;3:10-5.
208. Myers and Kornberg, Annu Rev Biochem., 2000;69:729-49.
209. Nagalakshmi et al., Science, 2008 Jun 6;320(5881):1344-9.
210. Nagy et al., Physiol Rev., 2012 Apr;92(2):739-89.
211. Natoli et al., Curr Opin Drug Discov Devel., 2009 Sep;12(5):607-15.
212. Nau et al., Nature, 1985 Nov 7-13;318(6041):69-73.
213. Nelson et al., Cell, 2006 Sep 8;126(5):905-16.
214. Newman and Keating, Science, 2003 Jun 27;300(5628):2097-101.
215. Nguyen and Zhang, Genes Dev., 2011 Jul 1;25(13):1345-58.
216. Nielsen et al., Genes Dev., 2008 Nov 1;22(21):2953-67.
217. Nikolov et al., Proc Natl Acad Sci U S A., 1996 May 14;93(10):4862-7.
218. Nishizuka et al., J Biol Chem., 1968 Jul 10;243(13):3765-7.
219. Northrup and Zhao, Immunity, 2011 Jun 24;34(6):830-42.

220. Nowak and Corces, Trends Genet., 2004 Apr;20(4):214-20.

221. Nowling et al., Mol Immunol., 2008 Jan;45(1):1-12.

222. Nuclear Receptor Nomenclature Committee, Cell, 1999 Apr 16;97(2):161-3.

223. Oh et al., Biochim Biophys Acta., 2012 Aug;1826(1):1-12.

224. Oikawa and Yamada, Gene, 2003 Jan 16;303:11-34.

225. Okada et al., PLoS One, 2011;6(9):e24837.

226. Ostuni et al., Cell, 2013 Jan 17;152(1-2):157-71.

227. Oyake et al., Mol Cell Biol., 1996 Nov;16(11):6083-95.

228. Panne et al., Cell, 2007 Jun 15;129(6):1111-23.

229. Park et al., Mol Immunol., 2002 Sep;39(1-2):25-30.

230. Pearson and Lipman, Proc. Natl. Acad. Sci. U S A., 1988 Apr;85(8):2444-8.

231. Piper et al., Nucleic Acids Res., 2013 Nov;41(21):e201.

232. Prokhortchouk et al., Genes Dev., 2001 Jul 1;15(13):1613-8.

233. Quinlan and Hall, Bioinformatics, 2010 Mar 15;26(6):841-2.

234. Raghav et al., Mol. Cell, 2012 May 11;46(3):335-50.

235. Raghuram et al., Nat Struct Mol Biol., 2007 Dec;14(12):1207-13.

236. Rashid et al., Genome Biol., 2011 Jul 25;12(7):R67.

237. Rauscher et al., Science, 1988 May 20;240(4855):1010-6.

238. Razin et al., FEBS Lett., 2013 Jun 27;587(13):1840-7.

239. Rébé et al., Circ Res., 2009 Aug 14;105(4):393-401.

240. Repa et al., Science, 2000 Sep 1;289(5484):1524-9.

241. Rhead et al., Nucleic Acids Res., 2010 Jan;38(Database issue):D613-9.

242. Rhee and Pugh, Cell, 2011 Dec 9;147(6):1408-19.

243. Richardson et al., Nat Genet., 2006 Nov;38(11):1329-34.

244. Robertson et al., Nat Methods, 2007 Aug;4(8):651-7.

245. Rock et al., Proc. Natl. Acad. Sci. U S A., 1998 Jan 20;95(2):588-93.

246. Ross-Innes et al., Nature, 2012 Jan 4;481(7381):389-93.

247. Roussigne et al., Oncogene, 2003 Apr 24;22(16):2432-42.

248. Roussigne et al., Trends Biochem Sci., 2003 Feb;28(2):66-9.

249. Roybal et al., J Biol Chem., 2004 Apr 9;279(15):14844-52.

250. Rühl et al., PLoS Genet., 2015 Jun 1;11(6):e1005213.

251. Ruprich-Robert and Thuriaux, Nucleic Acids Res., 2010 Aug;38(14):4559-69.

252. Ryden and Beemon, Mol Cell Biol., 1989 Mar;9(3):1155-64.

253. Sandelin et al., Nucleic Acids Res., 2004 Jan 1;32(Database issue):D91-4.

254. Schaub et al., EMBO J., 1997 Jan 2;16(1):173-81.

255. Schuster et al., EMBO J., 1995 Aug 1;14(15):3777-87.

256. Seila et al., Cell Cycle, 2009 Aug 15;8(16):2557-64.

257. Shao et al., Mol Cell Biol., 2005 Jan;25(1):206-19.

258. Shi et al., Cell, 1991 Oct 18;67(2):377-88.

259. Shiio and Eisenman, Proc Natl Acad Sci U S A., 2003 Nov 11;100(23):13225-30.

260. Shiraki et al., Proc Natl Acad Sci U S A., 2003 Dec 23;100(26):15776-81.

261. Siersbaek et al., Cell Rep., 2014 Jun 12;7(5):1443-55.

262. Sieweke and Allen, Science, 2013 Nov 22;342(6161):1242974.

263. Smith et al., Nucleic Acids Res., 1994 Jan 11;22(1):66-71.

264. Sofueva et al., EMBO J., 2013 Dec 11;32(24):3119-29.

265. Song et al., PLoS One, 2012;7(11):e49026.

266. Spyrou et al., BMC Bioinformatics, 2009 Sep 21;10:299.

267. Stadhouders et al., Nat Protoc., 2013 Mar;8(3):509-24.

268. Stanley et al., Eur J Biochem., 2001 Oct;268(20):5424-9.

269. Sun et al., PLoS One, 2009;4(3):e4721.

270. Suzuki et al., Immunology, 2013 Jul;139(3):318-27.

271. Szántó et al., Mol Cell Biol., 2004 Sep;24(18):8154-66.

272. Takahashi and Yamanaka, Cell, 2006 Aug 25;126(4):663-76.

273. Takaoka et al., Nature, 2005 Mar 10;434(7030):243-9.

274. Takeuchi and Akira, Cell, 2010 Mar 19;140(6):805-20.

275. Tamura et al., Mol Biol Evol., 2011 Oct;28(10):2731-9.

276. Tan et al., Cell, 2011 Sep 16;146(6):1016-28.

277. Tanaka et al., Cell Struct Funct., 2013 Jul 6;38(2):145-54.

278. Taniguchi et al., Annu Rev Immunol., 2001;19:623-55.

279. Thomas and Chiang, Crit Rev Biochem Mol Biol., 2006 May-Jun;41(3):105-78.

280. Thompson et al., Science, 1991 Aug 16;253(5021):762-8.

281. Thorvaldsdottir et al., Brief Bioinform., 2012 Mar;14(2):178-92.

282. Trapnell et al., Bioinformatics, 2009 May 1;25(9):1105-11.

283. Trapnell et al., Nat Biotechnol., 2010 May;28(5):511-5.

284. Trapnell et al., Nat Protoc., 2012 Mar 1;7(3):562-78.

285. Trievel et al., Cell, 2002 Oct 4;111(1):91-103.

286. Tsai and Nussinov, Biochem J., 2011 Oct 1;439(1):15-25.

287. Tsai et al., Cell, 2014 Jun 5;157(6):1430-44.

288. Tsuchihara et al., Nucleic Acids Res., 2009 Apr;37(7):2249-63.

289. Tzameli et al., Mol Cell Biol., 2000 May;20(9):2951-8.

290. Umesono et al., Cell, 1991 Jun 28;65(7):1255-66.

291. Valen et al., Genome Res., 2009 Feb;19(2):255-65.

292. Vannini et al., Proc Natl Acad Sci U S A., 2004 Oct 19;101(42):15064-9.

293. Venkateswaran et al., J Biol Chem., 2000 May 12;275(19):14700-7.

294. Vennström and Bishop, Cell, 1982 Jan;28(1):135-43.

295. Vennström et al., J Virol., 1982 Jun;42(3):773-9.

296. Virbasius et al., Genes Dev., 1993 Dec;7(12A):2431-45.

297. Vliet et al., Genomics, 2006 Apr;87(4):474-82.

298. Wallace et al., PLoS One, 2013 Aug 16;8(8):e71533.

299. Walsh et al., Gene, 2003 Mar 27;307:111-23.

300. Wang et al., Cell Res., 2014 Dec;24(12):1433-44.

301. Wang et al., Genome Res., 2012 Sep;22(9):1798-812.

302. Wang et al., J Biol Chem., 1987 Nov 25;262(33):16080-6.

303. Wang et al., Science, 2004 Oct 8;306(5694):279-83.

304. Wara et al., Blood, 2011 Dec 8;118(24):6450-60.

305. Watson et al., Nature, 2012 Jan 9;481(7381):335-40.

306. Wei et al., EMBO J., 2010 Jul 7;29(13):2147-60.

307. Whyte et al., Cell, 1989 Jan 13;56(1):67-75.

308. Whyte et al., Cell, 2013 Apr 11;153(2):307-19.

309. Wilhelm et al., Nature, 2008 Jun 26;453(7199):1239-43.

310. Willment and Brown, Trends Microbiol., 2008 Jan;16(1):27-32.

311. Wisely et al., Structure, 2002 Sep;10(9):1225-34.

312. Wondrak et al., Biochem J., 2000 Nov 1;351 Pt 3:769-77.

313. Woo et al., J. Mol. Biol., 2007 Oct 26;373(3):735-44.

314. Wu et al., Drug Discov Today, 2013 Jun;18(11-12):574-81.

315. Xie et al., Nature, 2005 Mar 17;434(7031):338-45.

316. Xu et al., Curr Opin Genet Dev., 1999 Apr;9(2):140-7.

317. Xue et al., Mol Cell, 1998 Dec;2(6):851-61.

318. Yamamoto et al., Nature, 1988 Aug 11;334(6182):494-8.

319. Yant et al., Nucleic Acids Res., 1995 Nov 11;23(21):4353-62.

320. Yoneyama and Fujita, Immunol Rev., 2009 Jan;227(1):54-65.

321. Yoon et al., Mol Cell Biol., 2000 Jul;20(14):5343-9.

322. Yoshida, Cell, 2001 Dec 28;107(7):881-91.

323. You et al., Proc Natl Acad Sci U S A., 2001 Feb 13;98(4):1454-8.

324. Yuan et al., Science, 2005 Jul 22;309(5734):626-30.

325. Zang et al., Bioinformatics, 2009 Aug 1;25(15):1952-8.

326. Zentner and Scacheri, J Biol Chem., 2012 Sep 7;287(37):30888-96.

327. Zervos et al., Cell, 1993 Jan 29;72(2):223-32.

328. Zhang et al., Genome Biol., 2008;9(9):R137.

329. Zhang et al., Mol Cell, 2002 Mar;9(3):611-23.

330. Zhang et al., Science, 2015 Jun 26;348(6242):1488-92.

331. Zhang, Genes Dev., 2003 Nov 15;17(22):2733-40.

332. Zhao and Meng, Develop. Growth Differ., 2005 May;47(4):201-11.

333. Zhu et al., Cell Res., 2006 Sep;16(9):780-96.

334. Zou et al., J Biol Chem., 2011 Aug 12;286(32):28019-25.