SHORT 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

Examination of the transcription factors acting in bone marrow derived macrophages

by **Gergely Nagy**

Molecular Biology MSc

Supervisor: Endre Barta PhD

Doctoral School of Molecular Cell and Immune Biology

University of Debrecen

Head of the **Examination Committee**:	Prof. Gábor Szabó MD, PhD, DSc
Members of the Examination Committee:	Prof. Péter Nagy MD, PhD, DSc
	László Homolya PhD, DSc

The Examination takes place at the discussion room of Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen
on June 21st 2016, at 11 AM.

Head of the **Defense Committee**:	Prof. Gábor Szabó MD, PhD, DSc
Reviewers:	Prof. Imre Boros PhD, DSc
	Zsófia Nemoda PhD
Members of the Defense Committee:	Prof. Péter Nagy MD, PhD, DSc
	László Homolya PhD, DSc

The PhD Defense takes place at the Lecture Hall of Building A, Department of Internal Medicine, Faculty of Medicine, University of Debrecen
on June 21st 2016, at 1 PM.

1. Introduction

1.1. Transcriptional regulation

Transcriptional initiation, elongation and termination are main steps of nascent RNA synthesis. As once RNA polymerase has been launched, it runs along the gene, initiation may provide the most extended regulatory possibilities. In eukaryotes, 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). To launch Pol II, not only the presence of a large number of regulatory proteins, but also their enzymatic activity is needed, e.g. the multiple serine phosphorylation on the carboxy-terminal domain of RNA polymerase II subunit B1 (RPB1).

Histone proteins may carry several posttranslational modifications, among which acetylation and methylation are the most studied ones. There are numerous enzymes creating and eliminating modifications, which eventually build up the epigenetic patterns of the different cell types thus determining the accessibility of DNA. TFs bind specific DNA elements in the promoter and in farther regions called enhancers (or silencers). The operation of these regulatory elements can be explained by the looping of the DNA chain during which the binding sites of the promoter and enhancer(s) are getting close to each other in the 3-dimension chromatin structure.

Co-regulators, which interact with TFs but unable to bind DNA directly, are responsible for the modifications and the release of nucleosomes and ultimately for the gene expression. Co-activators such as P300/CBP proteins may possess intrinsic histone acetyltransferase activity, while co-repressors typically have histone deacetylase function. The acetylation of histones H3 and H4 at multiple lysine residues is generally specific for active enhancers. Histone methyltransferases and demethylases are the other better-known groups of histone modifying co-regulators.

1.2. Promoter and enhancer sequences guiding transcription factors

TFs that transduce signals into the nucleus, inside are directed also by specific DNA patterns (TFBSs). Promoters and enhancers possess distinct motifs. TATA-box was the first regulatory element that has been described in the eukaryotic promoter. It is bound by TATA-binding protein (TBP), which is part of the TFIID complex also composed of TATA associated factor (TAF1-13) proteins. TFIID is bound directly to the TFIIA trimer and the TFIIB protein, which latter can occupy B recognition elements (BREs). Of the "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.

Before the next-generation sequencing (NGS) era, it has been described that most of the promoters are TATA-less, half of them contain Initiator, and half have downstream promoter element or BRE, covering together close to 100% of transcription start sites (TSSs). Nowadays, it seems that these elements are much less abundant. NGS 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 width called transcription start regions (TSRs). TSSs adding up to 22% tend to carry the classical elements, while TSRs have rather CpG island motifs.

The first identified CpG-binding TF was specificity protein 1 (SP1), which is the founding member of SP1-like family. This family is closely related to Krüppel-like factors (KLF1-17), which have the same type of DNA-binding domain (DBD), so are likewise able to bind GC-box. 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. By this and other NGS methods, further motifs were determined: an ETS binding site (EBS), the CCAAT-box, CRE, TRE and the NRF1, GFY, GFX, MYC and YY1 motifs.

The first “erythroblast transformation-specific” (ETS) protein has been identified as a fusion oncogene of the E26 retrovirus leading to leukemia in chicken. There are 27/28 members of the superfamily in mouse and human, respectively, which are separated to 12 phylogenetic groups. In promoters, rather the ccGGAAgt sequence shows enrichment, which is generally bound by the ubiquitous GABP dimer, ETS2, ETV6 and the TCF family proteins, and can be bound cell type specifically by the members of the PEA3 and ELF families. There are two more specific motifs: SPDEF binds mainly the caGGATga sequence, while SPI family tends to rather bind the gaGGAAgt sequence.

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). Shortly later, AP-1 proteins were identified as JUN and its interacting partners, FOS and FOS-related antigens. By now, probably all basic leucine zipper (bZIP) proteins are known, thus it became clear that FOS and JUN proteins form distinct families and form heterodimers with ATF/CREB, BATF and MAF family proteins. Dimers including ATF/CREB are considered as CREB proteins as these usually bind cAMP responsive element (CRE). AP-1 and CREB proteins share in a palindrome sequence built up from two TGA half sites with one and two spacers, respectively.

Nuclear transcription factor Y (NFY) is a heterotrimeric complex that binds CCAAT-box, and although this element is more abundant than the TATA-box, NFY has only a synergistic “enhancer” role. Nuclear respiratory factor 1 (NRF1) is a unique TF forming homodimers and binding a palindrome sequence in the promoter of a large amount of genes. As its element is bound only in the methylated state of the middle cytosines, it was not easy to identify general factor X (GFX) as ZBTB33. For the binding of GFY motif there are two candidates, ZFP143 and THAP11.

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 and to the members of PIC. These ligands 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 the AGGTCA sequence. The C-terminal ligand-binding domain (LBD) provides ligand specificity and selectivity, and also dimerization and other interaction surfaces. In the case of orphan receptors, this is responsible for the constitutive or posttranslational modification dependent regulation.

NRs can be discriminated into four classes. The first cloned NRs, glucocorticoid and estrogen receptor are the founding members of the class of steroid hormone receptors (class I). Class II receptors form heterodimer with retinoid X receptor (RXR) 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, belongs to class III, which collects the dimeric orphan receptors. Monomeric orphan receptors of class IV bind the consensus hexamer. Class II NRs bind a large variety of lipids. Peroxisome proliferator-activated receptor (PPAR) binds several kinds of fatty acids; retinoic acid receptor (RAR) binds all-trans and 9-cis retinoic acid, and thyroid hormone receptor (THR) binds triiodothyronine. Vitamin D receptor (VDR), liver X receptor (LXR), farnesoid X receptor (FXR), pregnane X receptor (PXR) and constitutive androstane receptor (CAR) all respond to steroid derivative ligands.

1.4. Transcription factors in macrophage development

Macrophages are professional phagocytes that are descendant of the myeloid lineage. Although their origin and tissue milieu modulate their function and metabolism, it seems that macrophages always share in the same TFs. During their differentiation, RUNX1 is up regulated by FLI1 (ETS) and together with TAL1 and C/EBP beta, it induces PU.1 (ETS) expression, which, together with C/EBP alpha, ultimately determines the macrophage lineage. There is a further group of TFs essential for macrophage function, the interferon (IFN) regulatory factor (IRF) family, which has 9 members. IFN-stimulated response element (ISRE) is bound by several IRF dimers, but certain IRFs form heterodimers also with PU.1. In mouse bone marrow derived macrophage (BMDM) cells the ETS-IRF composite element (EICE) is bound by the PU.1/IRF8 heterodimer.

1.5. NGS methods in functional genomics

ChIP-seq and RNA-seq are probably the most widely used NGS methods with more or less refined evaluation pipelines; however there are still specific problems that need unique solutions. One of these is the prediction of nucleosome depleted or nucleosome free regions (NFRs). There are only few methods allowing the determination of NFRs from histone modification ChIP-seq data e.g. that of HOMER, which searches for regions with the greatest differential in ChIP-seq signal. There is a special method to map the nascent transcriptome called global run-on sequencing (GRO-seq), for which analysis – beside ours – there are only two published pipelines, a hidden Markov model based transcript prediction approach and the one of HOMER, however these concentrate largely on the elongated transcripts on the dominantly transcribed strand, and thus the resolution of the prediction is poor on the other strand.

2. Aims of the study

Mouse BMDM is a well-characterized cell type thus is a good model to investigate the epigenetic landscapes determined by TFs. In histone modification ChIP-seq data, NFRs are the functionally most important regions, so to determine these, we carried out experiments for H3K4me2, H3K4me3, H3K27ac and H4ac, and developed a novel NFR prediction method. The next goal was to validate our method with motif enrichment analyses and to determine the TFs that are able to bind the elements matching with the found motifs. As the mapped TFBSs might be occupied by several members of different TF groups, 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.

The heterodimerizing partners of RXR activate several pathways, so RXR seems rather a passive, assistant molecule. Upon RXR ligandation of BMDM cells we would have been expected gene activation by LXRs, RARs and PPARs, thus it was a question whether there were distinct pathways specific only for RXR. To answer this question, we carried out ChIP-seq for RXR, PU.1 and P300, and followed the gene expression changes by 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 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. We carried out ChIP-seq also for CTCF and RAD21 proteins, and developed a prediction method to detect the insulator regions that assigned the borders of the regulatory units. 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 MCSF) for 5 days. Cells then were treated for 1 hour with vehicle or 100 nM LG268 (LG100268) ligand, gift from M. Leibowitz (Ligand Pharmaceuticals).

3.2. Chromatin immunoprecipitation coupled with next-generation sequencing

BMDM cells were cross-linked with 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). 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), and were applied for QPCR analysis or library preparation. ChIP-seq libraries were prepared with Ovation Ultralow Library Systems (NuGen) according to the manufacturer's instructions. Libraries were amplified in 16 PCR cycles and then to remove primers were gel-purified with E-Gel systems (Life Technologies). Sequencing was carried out with Illumina HiScanSQ sequencer.

3.3. ChIP-seq analysis

The primary analysis of the ChIP-seq derived raw sequence reads has been carried out using our ChIP-seq analysis command line pipeline. Alignment to the mm9 mouse genome assembly was done by the BWA tool, and BAM files were created by SAMTools. Genome coverage files were generated by HOMER, and used for visualization with IGV2. For the prediction of NFRs from histone modification data, PeakSplitter (EMBL-EBI) was applied to determine nucleosome occupied regions (NORs) by using the genome coverage information. NFRs were defined as the regions between the pairs of NORs. Finally, a summit and an edge based prediction approach were combined to get the fine-positioned NFRs.

ChIP-seq peaks were predicted by MACS2. Two parallels of the control and LG268-treated RXR samples were analyzed by DiffBind v1.0.9: 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 other samples was determined manually. Average read distribution histograms and heat maps centered to the middle of NFRs or peak summits were created by HOMER. Overlaps were defined by BEDTools and visualized by VennMaster-0.37.5. HOMER was used to predict motif enrichments, and fuzznuc (EMBOSS) was used to search for the specific RXR elements.

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

Global run-on assay and library preparation was performed as described earlier (Core et al., Science, 2008; Hah et al., Cell, 2011). 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.

3.5. GRO-seq analysis

The primary analysis of the raw sequencing data has been carried out similarly as detailed for ChIP-seq. The pool of sequence reads obtained from 2x4 macrophage samples was used for transcript prediction and annotation. The strand specific genome coverage files were generated by using HOMER. “Transcibed units” were determined strand-specifically by PeakSplitter (EMBL-EBI). The length of these units was limited in +/-250 bp relative to their summit. Divergent sites were determined based on the pairs of units with divergent direction. BEDTools and other command line programs were used to build up transcripts from those units that were lying closer to each other than 600 bp on the same strand. Read enrichments of H3K4me3 ChIP-seq was used to separate promoter regions from the proximal and intronic enhancers.

Longer transcriptional events were categorized by the following criteria: 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). The remaining H3K4me3 marked intergenic and antisense transcripts were collected in group 2. Putative transcripts 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 and the potential genes overlapping with longer ones in the same direction were identified.

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. The determination of unique read number for each sample was done on those fragments, of which joint length was longer than 0.5 kb. 1/20 RPKM values were used for the expression analysis performed by maSigPro. DiffBind v1.0.9 was applied to determine the significantly changing divergent

sites upon 30-minute LG268 treatment by using the same parameters as for the RXR peaks. 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.

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 molecules were purified with poly-T oligo-attached magnetic beads. 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). Libraries were sequenced with Illumina HiScanSQ sequencer.

3.7. RNA-seq analysis

TopHat and Cufflinks toolkits were used for mapping spliced reads, making transcript assemblies and getting gene expression values in FPKM format.

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. 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 BEDTools. Enrichment analysis of RAD21 on the insulator and the “active/passive” ($GRO^{+/-}$) RXR-bound regions was done similarly as described above.

3.9. Chromosome conformation capture (3C)

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). Tandem primers were designed in the close proximity of the restriction enzyme cutting sites.

3.10. 3C-sequencing

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. 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 of two technical replicates of two biological replicates were sequenced on Illumina MiSeq and HiSeq2000 sequencer.

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. For the frequency analysis of distal interactions, 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.

3.12. Phylogenetic comparison of the 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 MEGA 5.05 software were used. The 56 amino acid long fragment of the basic domains was applied for multiple alignment by using MUSCLE. 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.

4. Results

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

The more than 22,000 enhancer specific NFRs determined by HOMER were much fewer than expected, however, this method worked pretty well for the H3K4me3 derived data. To find as much NFRs as possible, PeakSplitter was used to determine NORs. These NORs enabled the fine-tuning of NFR prediction and to find broader NFRs as well. With our approach, about four times more NFRs could be detected as compared to the results generated by HOMER, and 56,704 regions could be determined both with acetylation and methylation marks, which might be the most active regulatory regions of our macrophages.

The detected continuous NFR width distribution was in agreement with the previously proposed "nuclosome rolling" models. The bigger half of the enhancer specific NFRs showed histone release on 150 +/- 75 bp wide regions, which approximately equals with the length of the DNA double helix, wrapped around one nucleosome core. For H3K4me3, typically broader NFRs were observed compared to those of enhancer marks, which may be due to the PIC assembled on the promoter. The intensity of H3K4me3 highly correlated with the enhancer marks, but the broadest regions showed the highest coverage by trimethylated histones.

4.2. Determining macrophage specific transcription factors from NFR predictions

Top NFRs derived from histone acetylation data were suitable to predict all motifs specific for macrophages: those of PU.1, AP-1, C/EBP, IRF, CREB and RUNX. Interestingly, the top H3K4me2 marked regions gave the same promoter specific motifs as those of the H3K4me3 marked ones: SP1, NFY, ETS, NRF1, GFY, YY1, GFX, CREB and IRF. Ultimately, we could predict almost 30,000 putative enhancers with macrophage specific

elements (NFRs of H3K4me2 and H4ac and/or H3K27ac, without H3K4me3) and more than 11,000 promoters with the well-known TSS/TSR specific motifs.

4.3. Examination of histone patterns near PU.1 binding sites

There was a high correlation between the intensity of histone modifications and PU.1 enrichment. Approximately half of the PU.1 peaks overlapped with NFRs, which surprisingly meant that a significant fraction of peaks overlapped with NORs or potential heterochromatic regions, thus we provided further evidence that PU.1 is able to bind those regions, which are theoretically wrapped into nucleosomes. Beside its own elements, PU.1 frequently bound DNA in the close proximity of AP-1, C/EBP, RUNX and IRF elements. The promoter specific EBS was also likely to be bound directly by PU.1, however with lower frequency as compared to its specific binding sites.

4.4. Combining NFR data to get further putative regulators

The use of a stringent consensus NFR set derived from both histone acetylation data improved the enrichment of the previously detected motifs. Those regions marked also with H3K4me2 but H3K4me3 gave even better motif enrichments, and newer motifs emerged: a HLH motif like hit and the motif specific for the MADS protein family. Based on sequence similarity, the unkown motif might be the M-box specific for the MITF/TFE (MiT) protein group. MiT proteins are members of the bHLH-ZIP family that form dimers only with MiT proteins and are responsible for phagocytic activity. The ancient DBD of MADS family binds the CArG-box. Beside this DBD, MEF2A-D proteins have an additional MEF2 domain, which is also responsible for DNA-binding, dimerization and protein interactions. Interestingly, MEF2 proteins indeed have roles in monocyte/macrophage differentiation.

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, which regulate gene expression through NRs. The best-known NRs in macrophages are the RXRs and their heterodimerizing partners, PPARs, LXRs and RARs. To examine these, we accomplished ChIP-seq for RXR from mouse BMDM cells in the absence or presence of the RXR agonist LG268 and determined the ChIP-seq peaks. The executed statistical approach evidenced a significant increase in DNA occupancy by RXR at 730 regions, while only 83 showed significant decrease. Beside this we got a reasonably sized consensus peak set of 5,206 peaks, which could be used for the following analyses.

PU.1, in average, showed a bit lower coverage in the LG268-treated cells than in control macrophages. At most PU.1 bound sites, a moderated up-regulation could be observed, while quarter of the peaks showed a stronger decrease. At certain regions, ChIP-QPCR experiments were carried out to validate this effect of RXR activation. Upon LG268 treatment, P300 showed a global increase in DNA occupancy. The P300 bound sites in most cases overlapped with the RXR and/or PU.1 peaks. These suggested that P300 functioned as a co-activator highly related to RXR.

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. We predicted all possible divergent transcripts regardless of elongation, and then we used the coverage data of H3K4me3 to separate promoters from enhancers. The strand specific "subpeaks" generated by PeakSplitter gave the possibility to determine all short divergent and elongated transcripts with a very high resolution on both strands.

We predicted 10,586 genes with pausing, H3K4me3 mark and accurate annotation, 4,601 putative genes with H3K4me3 mark and yet unknown TSS, 8,869 other transcripts without H3K4me3 mark and 1504 low expressed genes showing only weak initiation signal. Together we could predict 11,235 known genes and 12,821 yet unknown transcripts including elongated enhancer transcripts and other long ncRNAs. We found 318 up and 423 down regulated genes upon LG268 treatment; however the induced genes changed with a much greater extent.

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

To follow the changes also on the matured mRNA level, we performed RNA-seq upon LG268 treatment for the same time points as in the case of GRO-seq. We also wanted to identify the TFs binding the motifs determined by the NFR prediction, so we collected the expression levels of the different groups of TFs.

Third of the Sp1/Klf genes were expressed in BMDM. Klf6 showed the highest expression, of which protein product had been described as a key factor in pro-inflammatory macrophages. Interestingly, Klf10 showed an immediate induction upon LG268 treatment that was followed by a repression both on the nascent and matured RNA level. KLF10 is specific for the bone marrow derived proangiogenic cells, which induces vascularization.

At the level of nascent RNA, Etv3 was yet comparable with Pu.1, but the mRNA of the latter showed much more stability. 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. Interestingly, there were two induced family members, Fli1 and Ets2, which had been previously connected to angiogenesis in different cell types.

As Thap11 was highly expressed, while Zfp143 showed low expression, it seemed decided what GFY was in our system. Although the promoter specific NFRs gave clear motif enrichment, the mRNA of Zbtb33 (Gfx) showed low level. Nrf1 showed a bit higher, Yy1 a much higher expression than Gfx, which was in agreement with the motif enrichments.

Of the 44 AP-1/CREB-related genes, several showed high expression from almost all family. Of the partners of Jun, Atf3 showed the highest level, while Mafg, Mafk, Batf, Fos and Jdp2 genes showed lower and lower expression. These latter two genes were induced by RXR activation, which meant a quick, transient action of Fos, while Jdp2 needed more time to get its mRNA level elevated. This has a biological sense, as FOS is rather an activator with high affinity to JUN thus can supersede ATF3 and BATF repressors from the TREs, while JDP2 is a repressor of JUN proteins, which can cause then repression on the very same elements. ATF4, of which coding gene showed the highest level of the AP-1-related genes, can dimerize with all JUN partners of the previous model by acting on CREs. This suggests that ATF4 might play a similar role as JUN, in parallel on different regulatory sites. Interestingly, ATF4 activates Vegfa expression and thus also induces vascularization.

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.

Based on our gene expression results, BMDM cells are probably capable to respond to glucocorticoid, thyroid and even estrogen hormones by GR, THRA and ESR1, respectively. BMDM cells 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

The ratio of active and inactive RXR binding sites based on the overlap with divergent sites was 53.4 vs. 46.6. By performing the "differential binding" analysis of divergent sites we found that RXR binding was much more specific for increasing enhancer transcription (673 vs. 45 peaks). Our criteria for the annotation of these sites were the following: the direction of the expressional change of enhancer transcripts with RXR binding had to be identical with the one of a regulated gene having a TSS within 1 Mb. This way, we assigned 387 enhancers and 27 "silencers" to 226 up and 26 down regulated genes, respectively.

We found 25 angiogenesis related genes that were directly induced by LG268, of which 21 had RXR peak(s) nearby. The sequential induction of FOS/ATF4, ETS2, KLF10 and FLI1 might contribute also to the observed transient inductions. The resulted 21 genes included vascular endothelial growth factor A (Vegfa) and angiopoietin-like 4 (Angptl4).

4.9. Putative binding elements of the annotated RXR peaks

The motif enrichments of the annotated RXR peaks gave that 57% of the peaks contained DR1, DR4 or other NR elements. Based on the gene expression 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. 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.

4.10. Examination of functional domains

We carried out ChIP-seq also for the insulator specific 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. Based on our results, 30,290

CTCF and 24,648 RAD21 peaks could be determined. 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. Of the 2786 RXR binding sites expressing divergent transcripts, 223 significantly induced and only 21 repressed could be determined. We determined 10,204 putative functional domains, and their unification resulted almost 700 active topological domains. 80% of the directly regulated genes together with their regulatory site(s) were located on common functional domains, and 33% of these had induced RAD21 coverage on the belonging RXR-bound enhancer(s) and/or CTCF binding site(s).

We carried out NGS coupled chromosome conformation capture (3C-seq) experiments for some selected putative regulatory units. This method gives all interaction (target) sites genome-wide for a chosen region (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. For the Vegfa locus, we predicted an almost 300 kb loop between the promoter proximal upstream and a very distant downstream insulator. 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. 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. 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.

5. Discussion

During processing our NGS data, beside the NFR prediction, we developed a pipeline to process GRO-seq data, as well, which 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 THAP11 in our experimental system. 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.

General TFs of TFII complex and those binding in CpG-rich promoters are essential for the development and maintenance of each cell. Differentiation of the 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, and PU.1 can likewise liberate the linker DNA between the nucleosomes. In macrophages, the main, also dimerizing partner of PU.1 is IRF8, but C/EBPA is similarly determinative as its over-expression could trans-differentiate pre-B cells into macrophages. AP-1 proteins can also join to PU.1 in macrophages, 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, is still present in the matured cells. The further TFs of macrophages are probably more specific for the different functions, the general 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, and found high correlation between the promoter (H3K4me3) and enhancer specific (H3K4me2, H4ac, H3K27ac) modifications. The most frequent NFR length determined for enhancer marks was between 125 and 150 bp, 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. 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 their intensity 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, and by narrowing the regions to the most active enhancers, we got some unexpected motifs, as well. 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. 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. 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. There were several partners of JUN expressed, and ATF4 seemed to occupy the CREs with the very same heterodimerizing partners as of JUN.

Pu.1 showed an extremely high mRNA level, 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. 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. 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 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. 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, and its binding frequency was typically elevated upon treatment with its LG268 agonist. This enrichment seemed to be associated to activator functions, as the co-activator P300 also followed its enrichment. What is more, more than 2/3 of the P300 bound regions were occupied also by RXR, which might indicate a closer, even direct interaction between these proteins. In contrast, PU.1 showed rather lower binding frequency upon RXR activation.

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. Surprisingly, the number of significantly repressed genes (423) exceeded the number of the induced genes (318) 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. 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, Angptl4 is a PPAR target, while Abca1 and Abcg1 are LXR target genes.

Interestingly, several, different types of TFs were also induced by RXR, which are related to angiogenesis. Beside their other functions, KLF10, FLI1, ETS2 and ATF4 have also angiogenic roles. ATF4 did not, but some of its heterodimerizing partners did show induction. 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, 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.

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 Hbegf, the heme oxygenase 1 (Hmox1), the thrombospondin receptor (Cd36) and several other genes playing role in blood vessel formation. 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 bound and probably regulated by RXR. Both the induction of enhancer transcription and the close to 300 kb domain indicated that we found the major regulatory sites of Vegfa in mouse BMDM cells, so this finding was further validated by 3C-seq, which indeed showed interaction between the gene and the diecovered downstream region. 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.

6. Keywords

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

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.

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

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.