SHORT THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

Genome-wide study of DNA binding sites of interferon regulatory factors in dendritic cells

by Mária Nagyné Csumita

Supervisor: Lajos István Széles, PhD

UNIVERSITY OF DEBRECEN
DOCTORAL SCHOOL OF MOLECULAR CELL AND IMMUNE BIOLOGY

DEBRECEN, 2022

Genome-wide study of DNA binding sites of interferon regulatory factors in dendritic cells

By Mária Nagyné Csumita, MSc

Supervisor: Dr. Lajos István Széles, PhD

Doctoral School of Molecular Cell and Immune Biology, University of Debrecen

Head of the **Examination Committee**:	Prof. Dr. Gábor Szabó, PhD, DSc
Members of the Examination Committee:	Prof. Dr. Imre Miklós Boros, PhD, DSc
	Dr. Árpád Lányi, PhD

The Examination took place at the Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen; at 12 p.m.; 1st of July, 2019

Head of the **Defense Committee**:	Prof. Dr. László Fésüs, PhD, DSc, MHAS
Reviewers:	Prof. Dr. Tamás Emri, PhD, DSc
	Dr. Tibor Pankotai, PhD
Members of the Defense Committee:	Dr. Ida Gálné Miklós, PhD
	Prof. Dr. Imre Miklós Boros, PhD, DSc

The PhD Defense takes place at the Department of Emergency Care and Oxyology, lecture room, Faculty of Medicine, University of Debrecen; at 11 a.m.; 3rd of February, 2023

Introduction

1.1. Dendritic cells

In 1973 Zanvil Cohn and Ralph Steinman published their discovery of previously unknown cell type with microscope among mouse spleen cells, which was named dendritic cell (DC) from the Greek word dendron (tree) after its tree-shaped. Since then, several subtypes have been described and DCs have been shown to play a key role in triggering and regulating multiple immune responses. Based on their origin, we distinguish myeloid and lymphoid DCs. Myeloid DCs are the most effective professional antigen-presenting cells of the immune system, participating in formation of immunogenic and tolerogenic immune responses, while the role of lymphoid plasmacytoid DCs is primarily the formation of the antiviral immune response by the production of type I interferon (I-IFN). DCs are activated in the presence of pathogens, as they express pattern recognition receptors (PRR) that detect conserved structures on pathogens (pathogen-associated molecular pattern(s), PAMP). PRRs are primarily transmembrane protein, but cytoplasmatic PRRs are also known.

1.2. Subtypes of dendritic cells in mice

Several subpopulations of DCs have been identified in humans and mice. In our work, we used mice DCs. Differentiation of different DC populations is regulated by cytokines and transcription factors (TF). The differentiation of precursor DCs from myeloid and lymphoid plasmacytoid DC (pDC) begins in the bone marrow.

The role of pDCs is primarily the formation of antiviral immune response and the production of large amounts of I-IFN as a result of viral infection, but they also have the ability to present antigens. Langerhans cells are located in the epidermis of skin. Similar to the migratory DCs after the activation and antigen uptake they migrate to lymph nodes where the present antigen and regulate expression stimulus-dependently.

Monocyte-derived DCs differentiate from circulating monocytes in an inflammatory state. They share several properties of DCs. Monocytes express macrophage colony-stimulating factor receptor (MCSFR), which is essential for their differentiation.

Migratory DCs are present in different tissues of the body. They are specialized for antigen processing and presentation, acting as antigen recognizing sentinels. Different subpopulations are distinguished based on the expression of cell surface molecules. Migratory CD11+ and CD103+ DCs undergo maturation after pathogen uptake and PAPM recognition, which initiates they migration to the nearby lymph nodes. During maturation, their morphology changes, their ability to process antigens and their cell-activating capacity increases. Meanwhile, pathogen antigens are

being processed and proteins are broken down into peptides. The fundamental difference between migratory CD11+ DCs and migratory CD103+ DCs is that migratory CD11+ CDs present antigen on MHC-II (direct presentation), while migratory CD 103+ DCs primarily via MHC-I (cross-presentation).

Lymphatic tissue-resident DCs are found in different lymphatic tissues, such as the spleen, thymus and lymph nodes. They may be classified based on the expression of CD4 and CD8α cell surface markers. CD8+ cell surface molecule is a co-receptor helping the interaction between MHC-I and T cell receptor. Lymphatic tissue-resident DCs are activated by PAMPs found in blood, and PAMPs transported by migratory DCs into lymphatic tissues. This is also how they take up the antigens that are to be presented. CD8+ cells are similar to CD103+ migratory DCs in their ability to cross-present antigens via MHC-I membrane proteins. CD8+ DCs differentiate from precursor DCs found in the lymphatic tissues. Flt3L is a key cytokine in their development. Several transcription factors are essential for their development, such as PU.1, IRF8 and Batf3. CD8+ DCs help in maintaining tolerance to the body's own tissues. Immature cells mainly express a large number of receptors promoting the recognition and uptake of the pathogen. Its main PRRs are Toll-like receptors (TLR) (TLR3 and TLR9) and C-type lectin receptors (CLR). After stimulation and maturation, the expression of MHC and co-stimulating molecules increases, and large quantities of cytokines are produced, such as IL-12 and other inflammatory cytokines. Meanwhile the antigens are presented via MHC-I, therefore they have an important role in activating cytotoxic CD8+ T-cells, and, in inducing antiviral immune response.

1.3. MutuDC CD8$^+$ cell line

DCs occur in relatively small amounts in mice and humans, and are sensitive after isolation. Precursors can be isolated from mice bone marrow (or human blood), but '*in vitro*' differentiation requires several days and the use of cytokine, thus it is time-consuming and expensive. To overcome these difficulties, Hans Acha-Orbea and his collegues (Unil, Lausanne, Switzerland) created an immortalized cell line (MutuDC1940) from mice spleen-derived CD8$^+$ DCs, which is a homogenous cell population that can be extracted in large numbers. In a direct comparison they demonstrated that the MutuDC cell line retained the phenotypic and functional characteristics of primary CD8+ DCs. This cell line was used for our experiments.

1.4. Stimulus-regulated transcription factors

DCs are able to sense and integrate signals from their environment and induce immune responses based on them. The induction of immune response is accompanied by changes in the expression of hundreds of genes in DC. Transcriptional regulation is performed by TFs, whose

activity can be modified by signals. These TFs are called stimulus-regulated transcription factor (SRTF) according to Glass and Natoli. Cofactors of SRTFs regulate RNA polymerase II (Pol-II) activity though the modification of histones and chromatin structure and other mechanisms. To what extent the activated SRTF can regulate a given gene in a cell type is determined by three determinants: (1) presence of specific DNA motifs, (2) the usability of the regulatory element and its relationship with the promoter, and (3) binding of other TFs.

The first determinant, which has effect on SRTF binding is the presence of a specific DNA motif. The SRTF DNA-binding domain of a molecule usually binds a 4–6 bp long part of the DNA. SRTFs bind as dimers or trimers, their entire recognition site consists of different combination of half-sites. The optimal sequence is usually called the canonical DNA sequence. The presence of canonical DNA is neither necessary nor a sufficient condition for the binding of a certain SRTF. A presence of canonical DNA sequence is not necessary because SRTFs can bind several, sequences that are slightly different from the canonical sequence. These can be specified by position weight matrix (PWM). The presence of canonical DNA sequence is not sufficient because the condensed chromatin structure and the methylated DNA can inhibit the binding of SRTFs even to the canonical DNA sequence.

The second determinant, which has effect on SRTF binding is usability of the regulatory element (which is determined by the combination of chromatin accessibility, histone modifications, DNA methylation, etc.) and the 3D structure of chromatin (enhancer-promoter relationship). Among these, chromatin accessibility and histone acetylations are the most commonly investigated factors, since the open, nucleosome-free chromatin favours SRTF binding, while acetylated histones (mainly H3 and H4 lysine side chains) show association with active regulatory regions. In different cell types, active or activatable regulatory elements can be globally identified with Assay for Transposase Accessible Chromatin sequencing (ATAC-seq). Genome-wide mapping of histone modifications is performed by sequencing following chromatin immunoprecipitation (ChIP-seq) and using an antibody specific for the given histone modification. In addition to histone modifications, ChIP-seq technology is also suitable for mapping promoter and enhancer regions of different cell types. Lineage determining TFs (LDTFs) are responsible for the formation of the cell type-specific enhancer repertoire. In case of promoters, other TFs can also induce activation. LDTFs influence the formation of the given cell type, the maintenance of the expression of tissue-/cell type-specific genes, and the accessibility of enhancers. SRTFs mostly (>70 %) bind to regions that are accessible or show specific enhancer characteristics, and only to a lesser extent to previously unopened (latent or *de novo*) regions. Usually several LDTFs together with collaborating TFs are involved in the formation of accessibility. Since the majority of active

and activatable enhancer elements are created by LDTFs, so they are cell type-dependent, the SRTF-dependent response to a given stimulus is also largely cell-specific.

The third determinant, which has effect on SRTF binding is binding of other TFs. Regulatory elements often contain several SRTF binding sites, thus enabling the binding of SRTFs. As a result of inflammatory signals for example following TLR activation, several SRTFs are activated at the same time and often bind close to each other.

1.5. IRF transcription factors

The interferon regulatory factor (IRF) family plays important role not only in I-IFN induction, but also in cell-intrinsic antiviral responses, inflammatory responses and differentiation of certain cell types. Loss of IRFs or changes in their function may lead to the absence of specific immune responses or oncogenesis, and may be associated with autoimmune diseases such as inflammatory bowel disease, systemic lupus erythematosus and rheumatoid arthritis. The IRF family contains 9 proteins in humans and mice. The members of IRF family differ in terms of tissue expression, activation, dimerization and complex formation, as well as regulated transcriptional programs.

1.5.1. Structure of IRF-proteins

The basic structural elements of IRF proteins are the N-terminal DNA binding domain (DBD) and the C-terminal association domain, IRF association domain (IAD). DNA sequence recognized by DBD is the 5’-GAAA-3’ tetranucleotide, of which there are two in the canonical IFN-stimulated response element (ISRE, 5’-GAAANNGAAA-3’). IAD is responsible for establishing interactions with other TFs. Some IRF-proteins, such as IRF1, IRF3 and IRF7 form homo- or heterodimers.

There are IRFs (IRF4, IRF8 and IRF9) that interact with TFs from other families and bind to DNA as dimers or trimers with them. Some TFs have been described as partners of IRF4 and IRF8, such as PU.1. IRF9 forms a heterotrimer with signal transducer and activator of transcription 1 (STAT1) and STAT2 TFs creating Interferon-stimulated gene factor 3 (ISGF3) complex. Posttranslational modifications regulate the activity of IRF proteins and the interaction between them.

1.5.2. DNA motifs recognized by IRFs

Here we review the DNA sequences and motifs that bind IRFs. Each sequence contains at least one IRF half-site (5’-GAAA-3’), but other than that, great variability is characteristic. Sequences where only one base is specified in each position are called canonical sequences.

ISRE and TISRE

Original consensus ISRE (12-15 bp, from some I-IFN-stimulated genes) were defined as YAGTTTC(A/T)YTTTYCC and AGTTTCNNTTTCNC/T. According to our current knowledge, these are binding sites of ISGF3. The 10 bp long ISRE (GAAANNGAAA) is the canonical ISRE sequence. IRFs may have preference for bases in 5' and 3' flanking ends on ISREs. The totality of the possible binding sites is often difficult to specify with letters, thus they are usually displayed with motif logo that is made based on PWM. The presence of 3-4 tandem repeats of 5'-GAAA-3' sequence has been documented in the regulatory elements of several interferon-stimulated gene, effector ISG. The three consecutive ISRE half-sites (tripartite ISRE, TISRE) have been identified as binding sites for IRF1, IRF2 and IRF8 proteins.

EICE, EIRE, IECS and AICE

A 10 bp long ETS-IRF composite element (EICE, 5'-GGAANNGAAA-3') have been identified as a binding site for IRF4 and IRF8 proteins in the promoter regions of genes that are essential for the functioning of macrophages and B-cells. IRF4 or IRF8 binds to one half of EICE (GAAA), while E26 transformation-specific or Erythroblast Transformation Specific (ETS) binds to the other half (GGAA). Similar to the EICE motif, the IRF4/8-PU.1 complex may also bind to a 11 bp log ETS-IRF response element (EIRE, 5'-GGAANNNGAAA-3'), which has a 3 bp spacer. IRF-ETS composite sequence (IECS, 5'-GAAANN(N)GGAA-3') is an alternative and less common element, in which the ETS and IRF binding sites are arranged in reverse order with 2 or 3 bp spacer. In T-cells (where PU.1 levels are low) IRF4 does not bind to EICE, but to a so-called AP-1-IRF composite element (AICE), which serves as a common binding site for IRF4 and AP-1 TFs. AICE was originally described as TGAnTCA/GAAA sequence.

1.5.3. IRF3, IRF5 and IRF9 and the transcriptional programs they regulate

1.5.3.1. IRF3 and antiviral cytokines

TLR3 and TLR4 signalling pathways lead to the activation of IRF3. In addition to TLR, RLR signalling also activates IRF3. Its activation is regulated by two related protein kinases, IkB kinase e (IKKe) and TANK-binding kinase 1 (TBK1), which phosphorylate the IRF3 protein on the Ser396 chain, promoting the protein dimerization and nucleus translocation. IRF3 forms a homodimer, or heterodimer with other IRFs (such as IRF3/IRF7).

Phosphorylated IRF3 translocated to the nucleus, where it activates, among others genes encoding antiviral cytokines.

1.5.3.2. IRF5 and the inflammatory program

IRF5 activation is regulated by TLR signalization through MYD88 through phosphorylation of Ser158 and Ser309 chains by TBK1. IRF5 protein also plays an important role in the inflammatory responses.

Genes encoding inflammatory cytokines form the bases of TLR-induced inflammatory programs in macrophages and DCs. These have multiple functions, from mediating systemic inflammation to local T-cell response. Genes encoding inflammatory cytokines are activated and regulated together with several other genes with different functions. The activation of these genes, which is activated by a specific inflammatory stimulus, is called inflammatory program.

1.5.3.3. IRF9 and the antiviral interferon-stimulated genes (ISGs)

I-IFNs bind to a heterodimeric receptor complex consisting of IFN-α receptor 1 (IFNAR1) and IFNAR2 subunits. I-IFN signalling leads to phosphorylation of STAT1 and STAT2 proteins, heterodimerization and interaction with IRF9 protein, which forms the ISGF3 complex. ISGF3 binds ISRE elements in the nucleus resulting in transcription of ISGs. Moreover, IFNs also induce non-canonical JAK/STAT signalling. In addition to ISGF3, complexes containing the IRF9 protein and only one STAT1 or STAT2 protein also regulate gene expression. Binding of ISGF3 to the ISRE leads to the recruitment of cofactors and ultimately to increased activation of RNA polymerase II on the promoters of regulated ISGs.

Proteins encoded by ISGs have various roles. Several effector ISGs play an important role in fighting the infections caused by viruses, bacteria or parasites by directly inhibiting the life cycle of pathogens.

2. OBJECTIVES

The experiments described in my thesis belong to two related topics, published in two scientific publications. First, we studied how different IRFs bind to different sites and regulate different transcription programs. Later the basic question of our second topic was the (epi)genetic features of effector ISGs, which enable I-IFNs to be induced in numerous cells. In order to answer these questions, we used global methods such as ChIP-seq, ATAC-seq and RNA-seq, as well as bioinformatic methods to analyse our data.

2.1. Identification and characterization of IRF3-, IRF5- and IRF9-specific binding sites

- Mapping binding sites of IRF3, IRF5 and IRF9 proteins in the whole genome by ChIP-seq.
- Identification of regions where there is a significant difference in the binding of IRF3, IRF5 and IRF9 (IRF3-, IRF5- and IRF9-specific binding sites).
- Characterization of IRF3, IRF5 and IRF9-specific binding sites based on the following features: ISRE motif, ISRE 5’ and 3’ flanking bases, binding of cofactors and a chromatin accessibility.
- Investigation of the relationship between IRF3-, IRF5- and IRF9-specific binding sites and transcription.

2.2. Identification and characterization of regulatory regions of effector ISGs

- Identification of I-IFN-induced effector ISGs and their possible regulatory regions using RNA-seq and IRF9 ChIP-seq.
- Identification of the DNA biding motifs of the regulator regions of effector ISG genes.
- Chromatin accessibility in the regulatory regions of effector ISG genes in DC and other immune cells.

3. Materials and methods

3.1. Cells and ligands

A wild-type $CD8^+$ DC line (Mutu1940) came from the laboratory of Prof. Hans Acha-Orbea (Unil, Lausanne, Switzerland). Cells were cultured at 37°C with 5% CO2 in medium containing 10% heat inactivated bovine serum, 100 U/ml penicillin, 100 μg/ml streptomycin and 50 μM β-mercaptoethanol.

Starting from and adequate number of cells (10-20 million cells/sample), the cells were treated with various TLR ligands and IFNβ ligand. In case of TLR9 activation 1mM CpG ligand, a synthetic CPG oligonucleotide was used. In case of TLR3 activations 5μg/ml Poly (I:C) ligand, an RNA analogue was used to treat the cells. In case of IFNα/β (IFNAR) receptor activation 100 U/ml IFNβ ligand was used to treat the cells. For ChIP-seq and ChIP-qPCR experiments $CD8^+$ cells were treated for 90 minutes with these ligands. In case of ChIP-qPCR experiments combination of these ligands were also used. For gene expression qPCR experiments cells were treated for 1.5, 3,6 and 12 hours.

3.2. ChIP-seq and ChIP-qPCR (chromatin immunoprecipitation coupled with high-throughput sequencing and chromatin immunoprecipitation coupled with real-time quantitative PCR)

Medium was aspirated, DCs were crosslinked with 2 mM DSG for 40 min and 1% formaldehyde for 10 min. Cross-linking was stopped by the addition of glycine to a final concentration of 125 mM for 10 min. Cells were washed and scraped. After centrifugation, the supernatant was aspirated and crosslinked cell pellet was frozen in liquid nitrogen and stored at -70°C until immunoprecipitation.

Cell lysis and nuclear lysis was performed with ChIP lysis buffer containing protease inhibitors. DNA was sheared with sonication into 100-2000 bp fragments. The following antibodies were used: IRF3 (D83B9, Cell Signaling), IRF5 (ab21689, Abcam), IRF9 (AF5629, R&D Systems), STAT2 (07-140, Merck Millipore) and STAT1 (sc-346, Santa Cruz Biotechnology). At this step, 50 μl of chromatin is stored at -20°C in the times the volume of absolute ethanol, later these samples are used as ‘input’ sample for ChIP-qPCR experiments.

Chromatin-antibody complexes were washed four times and eluted after being pulled down with magnetic beads. DNA fragments and the input samples were treated with RNase and Proteinase K, removing the RNAs and proteins. DNA was purified using a NucleoSpin Gel and

PCR Clean-up kit according to the manufacturer's instructions. ChIP-DNA was quantified using a Qubit fluorometer.

For ChIP-seq experiments indexed cDNA libraries were prepared from 1 to 10 ng ChIP-DNA using a TruSeq ChIP Sample Preparation Kit according to manufacturer's instructions. Libraries were sequenced on Illumina NextSeq 500 or HiSeq 2500 platforms at the Genomic Medicine and Bioinformatics Core Facility of the Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen. For ChIP-qPCR experiments primers were designed for amplifying the promoter regions of the nine paradigm genes and enrichment was determined by qPCR (quantitative polymerase chain reaction) in eight samples (UT, PIC, CPG, IFNB, PIC+CPG, PIC+IFNB, CPG+IFNB, PIC+CPG+IFNB) in triplicate normalized to input samples. SYBR Green dye was used.

The following temperature protocol was used: 1 minute at 94°C, 40 cycles of 12 seconds at 94°C, 30 seconds at 60°C.

3.3.ChIP-seq data analysis

The primary analysis of raw ChIP-seq reads was carried out using a ChIP-seq analysis pipeline developed at the Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen. Burrows- Wheeler Alignment Tool (BWA) was used to align the reads with the mm10 genome assembly. Model-based Analysis of ChIP-Seq 2 (MACS2) was used for predicting peaks (binding regions) with the following specific parameters: q value cut-off (q)=0.001 and subpeaks deconvolved within each peak (–call-summits). Artifacts were removed using the ENCODE blacklist. Consensus peak sets for each IRF contain the common peaks of the duplicates. By merging binding regions of IRF3, IRF5 and IRF9, a 'unified IRF3/IRF5/IRF9 cistrome' was generated. Coverage of predicted peaks (expressed as Reads Per Kilobase Million, RPKM) was calculated using bamtools, bedtools (coverageBed), and awk. For clustering, occupancy values were normalized by the median of values of each IRF. Integrative Genomics Viewer (IGV) was used for data browsing and creating representative snapshots. Genome coverage files (BedGraphs) were converted into tdf files using igvtools with the 'toTDF' option. Normalized tag counts for read distribution (RD) histograms and heat maps were generated by HOMER (Hypergeometric Optimization of Motif EnRichment), and then visualized by R or Java TreeView. Raw fastq files of ChIP-seq data can be found under the following GEO identifiers: GSE172376, GSE125340.

3.4. Motif discovery, PWM enrichment analysis and identification of DNA sequences

A *de novo* motif discovery was performed by findMotifsGenome.pl (HOMER). P-values were calculated by comparing the target region enrichments with those of background sets generated by HOMER.

Position weight matrices, PWMs were imported from HOMER database or defined by ourselves. PWMs for ISRE, EICE, AICE1, NFKB, TRE, CRE and GAS motifs were obtained from the HOMER database. PWMs for IECS, AICE2 and TISRE were not available in the database; therefore, PWMs were generated by re-analysing publicly available ChIP-seq datasets. Mapping of PWMs in the genome was performed by scanMotifGenomeWide.pl (HOMER).

For searching special DNA sequences in the clusters, lists of 6-mers and ISRE variants were generated. The DNA sequence of each binding region was obtained from the mouse genome (GRCm38/mm10) using bedtools (fastaFromBed). Using grep command, we searched for all 6-mers and ISRE variants and their reverse complements in all binding regions. The position of the 5'-GAANNGAAA-3' sequence was determined genome-wide using the R package Biostrings.

3.5. ATAC-seq experiments and data analysis

ATAC with high-throughput sequencing (ATAC-seq) was carried out in biological duplicates. Cells were scraped and counted to achieve 10 000–15 000/ml in ice-cold PBS. Nuclei were isolated with ATAC lysis buffer. Nuclei were used for tagmentation from two biological replicates using a Nextera DNA Library Preparation Kit (Illumina). After tagmentation, DNA was purified with a MinElute PCR Purification Kit (QIAGEN). Tagmented DNA was amplified with Kapa Hifi Hot Start Kit (Kapa Biosystems) using nine PCR cycles. Amplified libraries were purified again with MinElute PCR Purification Kit. Fragment distribution of libraries was assessed with Agilent Bioanalyzer and libraries were sequenced on an Illumina HiSeq 2500 platform.

Primary analysis of the ATAC-seq raw reads was carried out using an analysis command line pipeline previously described at ChIP-seq analysis. In addition to our own data, additional ATAC-seq raw data from the GEO database was downloaded and analysed. Bwa was used to align the reads with the mm10 genome assembly using default parameters and MACS2 was used for predicting ATAC-seq peaks (q-value $\leq$ 0.001). Artifacts were removed using the ENCODE blacklist. BedGraphs for visualization purposes were generated with makeUCSCfile.pl and then converted into tdf files using igvtools with the 'toTDF' option. IGV was used for data browsing and creating representative snapshots. For calculating the ratio of ATAC-positive regions a 'bedtools intersect' option was used. The GEO identifier of the ATAC-seq data is GSE125340.

3.6. Ternary diagrams and Random Forest analysis

An R software package, ‘ggtern’, was used to create the ternary diagrams. A given point on the ternary diagram was calculated based on three input variables (ChIP-seq signal or mRNA level). Random Forest machine learning method was applied in Python using a Random Forest Classifier from the scikit-learn package. Various sets of ChIP-seq, DNA motif or ATAC-seq values were used as input variables, while cluster names were used as ‘class labels’.

3.7 Gene sets

Genes belonging to various GO categories were downloaded from the Mouse Genome Informatics (MGI) database (http://www.informatics.jax.org). The list of ‘antiviral cytokines’ contains the common genes of ‘defence response to virus’ (GO:0051607) and ‘cytokine activity’ (GO:0005125) GO categories. The list of ‘inflammatory program’ contains those genes of ‘inflammatory response’ (GO:0006954), which were also listed in the ‘cytokine activity’ (GO:0005125) or ‘DNA-binding transcription factor activity’ (GO:0003700) categories. The list of ‘antiviral ISGs’ was selected based on review articles on ISGs. For the list of efector ISGs and inflammatory-related cytokines results of RNA-seq were also considered. Among the antiviral ISGs effector ISGs include those that were induced by IFNβ in DCs and were of effector function according to data found in the literature. Inflammatory-related cytokines (IR-cytokines) include those genes that were significantly induced by CpG in DCs and were found in both the inflammatory response (GO:0006954) and cytokine activity list (GO:0005125).

3.8. RNA isolation and RT-qPCR

RNA was isolated using Trizolate reagent (UD-GenoMed). After lysis chloroform was added for optimal separation of phases, and alcohol precipitation was used to recover RNA. After precipitation, RNA precipitate was centrifuged for 10 minutes at 4°C, 16000 g. The supernatant was removed, taking care not to touch the pellet. Pellet was washed with 70% ethanol. After removing the ethanol, RNA was dried at room temperature in a vacuum concentrator. RNA was dissolved in nuclease-free water (NFW, Affymetrix). RNA samples were stored at -20°C until measurements.

Reverse transcription was performed using a High Capacity cDNA Reverse Transcription Kit (Thermo Fischer Scientific) according to the manufacturer's instructions. The following temperature protocol was used for the PCR reaction: 10 minutes at 25°C, 120 minutes at 42°C, 5 minutes at 72°C. Applied Biosystems 2720 Thermal Cycler PCR Instrument was used. The resulting cDNA was stored on ice or at -20°C until QPCR measurements.

During the qPCR measurements, the expression of nine representative genes was determined by real-time quantitative PCR measurements in UT, PIC, CPG, IFNB samples in biological and technical triplicates. LightCycler 480 SYBR Green I Master and the LightCycler 480 Instrument (Roche) was used.

The following protocol was used: 1 minute at 94°C, 12 seconds at 94°C, 30 seconds at 60°C. Number of cycles: 40 (second and third steps). Rplp0 housekeeping gene was used to normalize the data.

3.9. RNA sequencing

For RNA sequencing, RNA was isolated using Trizolate reagent (UD-GenoMed). The quality of RNA was checked with qPCR and on an Agilent BioAnalyzer using a Eukaryotic Total RNA Nano Kit. Library preparation, sequencing and data analysis were performed by the staff of Genomic Medicine and Bioinformatics Core Facility of the Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Debrecen. RNA-seq libraries were prepared from total RNA using Ultra II RNA Sample Prep kit (New England BioLabs). Sequencing was executed on an Illumina NextSeq500 instrument using single-end 75-cycle sequencing. Sequence reads were aligned to the mm10 genome assembly using HISAT2. For subsequent analyses, BAM files were imported to the Strand NGS program (Strand Life Sciences Pvt., Bangalore, India). For identification of differentially expressed genes, moderated t tests with Benjamini-Hochberg false discovery rate $p < 0.05$ and filter on fold induction (fold change [FC] >2.0) were used.

4. Results

4.1. Clustering IRF3, IRF5 and IRF9 binding regions

Among the IRF transcription factors, IRF3, IRF5, IRF 8 and IRF9 were expressed at the highest level in our $CD8^+$ immortalized cell line.

To map IRF3, IRF5 and IRF9 binding regions, we performed ChIP-seq analyses after stimulating DCs for 90 minutes with CpG, pIC or IFNβ. These agonist activate receptors, signalling pathways, and the corresponding IRFs very efficiently (for example pIC→TLR3→IRF3, CpG→TLR9→IRF5 and IFN-β→IFNAR1/2→IRF9).

We continued working with those binding sites that could be identified in both replicates of a given sample. Using this method, we identified 25 203 (IRF3), 17 435 (IRF5) and 25 040 (IRF9) binding regions. If the occupancy values for IRF3, IRF5 and IRF9 on a given binding regions were similar (0.5 <ratio <2), the region was considered to be a common binding region for IRF3, IRF5 and IRF9, and referred to as a 'common cluster'. We also determined three additional clusters, in which the binding by one IRF was weaker than the other two IRFs, referred to as a 'shared clusters'.

4.2. Identification of DNA-motifs in the IRF clusters

We analysed DNA motifs in the clusters using three different strategies: (1) *de novo* motif discovery, (2) enrichment analysis of position weight matrices (PWMs) and (3) a search for specific DNA sequences. The three strategies have their own advantages and limitations and combining these approaches resulted in a more complex picture.

Using the first approach (*de novo* motif discovery), we found that the most enriched motifs in the IRF3- and IRF9-dominant clusters were DNA motifs that resembled ISRE, or a 'mixture' of ISRE and an Ets-IRF composite element (EICE). Consistent with previous IRF5 ChIP-seq studies, *de novo* motif discovery failed to detect an ISRE motif in the IRF5-dominant cluster. However, an ISRE half-site motif and a weak IRF-Ets composite sequence (IECS), were enriched in this cluster. In addition to these IRF-binding motifs, several other motifs were identified, including the NF-κB-binding motif (NFKB), the TPA response element (TRE), the cAMP response elements (CRE) and the gamma IFN activation site (GAS). These motifs are binding sites for TFs that are co-activated with IRFs in a pathway-specific manner. Upon TLR ligation, IRF3 or IRF5 are activated together with activating protein 1 (AP-1) members, which bind TRE or CRE, and members of the NF-κB family, which bind to κB sites. I-IFNs induce the formation of the ISGF3 complex together with STAT1 homodimers, which bind to GAS.

Our second approach compared the enrichment of the same PWMs in the clusters side-by-side. In *de novo* motif discovery similar motifs are combined, and direct comparison of specific motifs between clusters is not possible. This may be solved using enrichment analysis of PWMs. We obtained PWMs from the HOMER database or from reanalysed ChIP-seq data sets. In addition to ISRE, EICE, IECS and the motifs for co-activated TFs, we investigated three longer motifs, AICE1, AICE2 and TISRE motifs. We determined the threshold (cut-off) values for each motif systematically, based on a set of randomly selected size-matched genomic regions (random set).

We made the following three observations. First, ISRE was the most enriched motif in the IRF3- and IRF9-dominant cluster. In accordance with this, ISRE frequency was also high in IRF3-IRF9 shared cluster. Second, TISRE was especially frequent in the IRF3- and IRF9-dominant cluster, and IRF3-IRF9 shared cluster. Third, the enrichment of the EICE motif was similar to the ISREs' in the investigated clusters. These results do not imply necessarily that IRF3-dimers and ISGF3 complex bind EICE motif with a high frequency. More likely, IRF3-dimers and ISGF3 complex show a preference for accessible genomic regions that are primed by PU.1 and IRF8. These two TFs are lineage-determining TFs in this DC subtype, and together they bind EICE and IECS.

In further analyses of our second approach (PWM enrichment) we investigated the occurrence of NFKB, TRE, CRE and GAS motifs. These motifs are binding sites for transcription factors that are/may be activated together with certain IRFs. *De novo* motif discovery results and the enrichment analysis of PWMs agreed in most cases. The only discrepancy was observed in the IRF5-dominant cluster, where TRE and CRE motifs were enriched by *de novo* motif discovery, but not by PWM analysis. We found that NFKB motifs with high scores were polarized in the IRF3- and IRF5-dominant clusters, while regions containing GAS motifs with high scores were specially enriched in the IRF9-dominant cluster. TRE motif was found more frequently in IRF3-dominant, IRF3-IRF5 and IRF3-IRF9 shared clusters, while CRE in IRF3-dominant, and IRF3-IRF5 shared clusters.

In our third approach, we investigated the occurrence of specific DNA sequences in IRF clusters. First, we analysed ISRE half-sites. We investigated 6-mers instead of 4-mers because IRF binding is affected by 5′ and 3′ flanking bases of the 4bp long ISRE half-sites (5'-GAAA-3'). We generated a list of all 6-mers (n=535), which contained the canonical 5′-GAAA-3′ and two extra bases (5′-NNGAAA-3′, 5′-NGAAAN-3′ or 5′-GAAANN-3′). In the 5′-GAAA-3′ sequence, one mismatch was allowed. We determined the frequency of these 6-mers in the cluster. We identified a number of sequences (n=84), that occurred at least 1.5 times more frequently in the IRF3-, IRF5 and IRF9-dominant clusters, respectively, compared to the other two clusters. Most of these (n=68) showed the greatest enrichment in IRF5-dominant cluster. The enrichment of many 6-mers was

consistent with the results of the previous *in vitro* studies. Protein binding microarray (PBM) data indicated that sequences that contain 5′-GGAAAC-3′ are bound with higher affinity by IRF3 compared to IRF5, IRF6 or IRF9. Consistent with this observation, we found that GGAAAC is more enriched in the IRF3-dominant cluster. PBM experiments demonstrated that IRF5 is more tolerant of the ISRE half-site at position 3 and 4 than IRF7 or certain other IRFs. We detected enrichment of several 6-mers that differed at these positions from the canonical core sequence (for example, CGACAC and CGAGAC) in the IRF5-dominant cluster.

In PBM experiments, TGAAAC was bound with higher affinity by IRF9 compared to IRF3, IRF5 or IRF6. During our analysis, we found that this sequence and some other 6-mers, such as GAAACT, to be especially enriched in the IRF9-dominant clusters.

In further analyses of our third approach we searched for specific ISRE variants (n=37) that were selected based on previous studies or our 6-mer analysis. We found 20 ISRE variants that occurred at least 1.5 times more frequently in the IRF3-dominant clusters compared to the other two clusters. We identified 4 such ISRE variants in the IRF9-dominant cluster, however, we did not find any ISRE variants that occured more frequently in the IRF5-dominant cluster than in the other two clusters.

Consistent with the 6-mer ISRE results, we found that GGAAANNGAAA sequence occurred more frequently in the IRF3-dominant cluster relative to the other clusters. The ISREs with two extra AA at the 5′ position, also occurred more frequently in IRF3-dominant cluster. EMSA and protein crystallization experiments found that the sequences bound by the ISGF3 complex are longer than the 10 bp long consensus ISRE motif, due to the binding preference of IRF9 and STAT proteins. In our analyses, we found that 5'-GAAANNGAAACT-3' was enriched in the IRF9 dominant cluster.

4.3. Chromatin accessibility in IRF clusters

As reviewed in the Introduction, most of the binding of SRTFs occur in open, accessible enhancer regions previously marked (by LDTFs). Chromatin accessibility and SRTF binding influence each other: openness significantly determines binding, and binding further opens the region. IRF3 can promote nucleosome remodelling and/or the opening of previously inaccessible genomic regions in LPS-stimulated macrophages. Chromatin opening by ISGF3 complex has also recently been studied. To our knowledge, the ability of IRF5 to reshape the chromatin landscape has not been investigated in genome-wide studies. Chromatin accessibility may be an important determinant in the selection of IRF-specific enhancers. Therefore, we investigated accessible chromatin regions using ATAC-seq experiments in unstimulated (UT) DCs.

We calculated the percentages of accessible regions in the IRF cistromes and clusters. Regarding the cistromes, we found that the difference was moderate (~13%) between the lowest and highest frequencies (IRF3: 56.8% versus IRF5: 69.8%). Regarding the IRF-binding clusters, we detected much larger differences (~50%) between the lowest and highest frequencies (IRF3-dominant: 23.8% vs. Common: 73.8%). Among specific clusters, we found the most accessible chromatin regions (62.4%) in the IRF5-dominant clusters. The overlap between IRF binding regions, the accessible genomic regions and the canonical ISRE (5'-GAAANNGAAA-3') sequence was also determined.

We found that only a minority of the canonical ISRE sequences (<1%) were occupied by IRF3 upon activation *in vivo*. In contrast, ~35% of the regions, which were accessible before stimulation, were occupied by IRF3 after stimulation. A certain gene may be associated with multiple IRF binding sites. The diversity of marked (open) binding regions is exemplified by the cis-regulatory elements of Ccl5 gene. To conclude, we found that ATAC-negative regions were especially enriched in the IRF3- and IRF9-dominant regions suggesting that the 'chromatin barrier' contributed to IRF3- and IRF9-specific, but not to IRF5-specific, enhancer selection, by inhibiting the binding of other IRFs. Our observations are consistent with previous findings related to other TFs indicating that chromatin features play active roles in shaping the selective transcriptional responses.

4.4. Machine learning predictions of IRF dominance

One application of machine learning methods is prediction of class labels based on input data. Using Random Forest, we investigated how accurately the IRF-dominance could be predicted using features of binding sites. We performed pairwise comparisons using the cluster label (IRF3-, IRF5- or IRF9-dominant) of each binding region as class label. The motif scores for ten PWMs, ATAC-seq signals, and occurrence data concerning variants of ISRE and 6-mers, or a combination of these was used as input data. The accuracy of the prediction was calculated based on the ratio of correct vs. all predictions. A prediction was correct, when original and predicted class labels agreed. There was a significant difference depending on which features were used for the prediction.

Overall, the results show that considering all features, class labels were predictable with high accuracy (~80-85%) when comparing IRF5-dominant cluster with the other two dominant clusters. The class label prediction was less accurate (~65%) for the comparison between IRF3-dominant and IRF9-dominant clusters, which means that IRF5-dominant cluster significantly differed from the other two clusters. Based on the features of binding site, IRF3- and IRF9-dominant clusters were less separated from each other.

4.5. Genes encoding antiviral cytokines, inflammatory cytokines, and antiviral ISGs are often associated with IRF3-, IRF5- and IRF9-dominant regions

We analysed the enrichment of IRF clusters in the extended promoter regions of the genes of three transcriptional programs. Many of the antiviral cytokines, inflammatory program genes (cytokines and transcription factors) and antiviral ISG-s have been identified as target genes of the IRF3, IRF5 and IRF9 proteins. We compiled three gene sets based on Gene Ontology (GO) annotations and review articles.

The gene lists for antiviral cytokines, inflammatory program, and antiviral ISGs contained 25, 119 and 63 genes, respectively. IRF binding regions were determined and counted separately for each gene in the 20-kb regions around the transcription start sites (TSS ± 10 kb). Type of the binding site was examined and relative enrichment (normalized for cluster size) was calculated. We found that in the TSS ± 10 kb region of antiviral cytokines, the IRF3-dominant cluster was especially enriched (and to a lesser extent the IRF3-IRF5 shared cluster). In the TSS ± 10 kb region of genes involved in the inflammatory program, the IRF5-dominant and IRF3-IRF5 common clusters were enriched, while in the TSS ± 10 kb regions of antiviral ISGs, the IRF9-dominant and IRF3-IRF9 shared clusters occurred more frequently. We examined the characteristic DNA motifs and chromatin accessibility of the identified 28 IRF3-, 62 IRF5- and 45 IRF9-dominant binding sites. The analysis showed that the entire cluster and its subsets had similar patterns in most cases. The IRF5-dominant binding sites had particularly striking high NFKB motif score and ATAC-seq signal, and a low ISRE motif score.

Representative genes were selected to examine the relationship between IRF3, IRF5 and IRF9 binding and gene expression. We selected antiviral cytokines, inflammatory cytokines and antiviral ISG-s, which were associated with IRF3-, IRF5- or IRF9-dominant peaks in their promoter regions. Using RT-qPCR technique, we investigated their mRNA levels in the presence of different ligands, depending on time (1.5-12 hours).

We found that selected antiviral cytokines, inflammatory cytokines and antiviral ISGs were regulated most efficiently by pIC, CpG and IFN-B, respectively. It is important to emphasize that since genes could be regulated by several enhancers and TLR signalling pathways could activate several TFs simultaneously, the effect of the given ligand is not necessarily mediated exclusively by these regions and the examined IRF proteins. Our results show that regions predominantly regulated by IRF3, IRF5 or IRF9 proteins are important binding sites for specific transcriptional responses. Using ChIP-qPCR method we investigated whether co-operation or inhibition between TLR and IFN pathways could be detected in case of combined ligand treatments. In these experiments, the DCs were stimulated with either single agonists or a combination of ligands. We found that IRF3 binding to the IRF3-dominant regions of antiviral cytokines was inhibited by all

combined treatments. The IRF5 binding to the IRF5-dominant regions of inflammatory genes was increased when the pIC + CpG combination was used. Since IRF5 protein is activated by both TLR3 and TLR9 agonists, the increased IRF5 binding may be caused by cooperation between the two pathways for IRF5 activation. Finally, the binding of IRF9 to IRF9-dominant regions of antiviral ISGs was similar when IFN-β was used alone or in combination with CPG, suggesting that the TLR9 pathway cannot inhibit IRF9 activation or binding.

In contrast, the TLR3 pathway inhibited IRF9 activation or binding to IRF9-dominant regions. Collectively, these results suggest that competition or co-operation between IRFs and other SRTFs could markedly influence binding of IRF3, IRF5 and IRF9 proteins.

4.6. Identification of possible regulatory elements of effector ISGs

Using RNA-seq data, we generated a list of ISGs (n=1936), showing at least 2-fold induction after 6 hours of IFNβ treatment. We identified 37 effector ISGs selecting genes from this list described as antiviral effectors in previous studies.

The regulatory elements of the effector ISGs were identified by mapping the binding sites of IRF9. Active gene regulatory elements or that could be activated may be identified using several methods, but each method has its limitations. One possible method is to map activated SRTFs. The disadvantage of this method (if not combined with another method) is that it does not provide information whether the binding comes with the activity of enhancer. In case of certain pathways, another problem could be that one receptor activates several SRTFs (for example NF-κB, AP-1 and IRF factors in case of TLRs), whose binding sites only partially overlap. In our case, the latter was not a significant problem, since IRF9 is an essential component of IFN-induced transcriptional programs.

Assigning regulatory elements to regulated genes remains a significant technical challenge. Since most of the enhancer-promoter interactions occur within a distance of ∼50 kb of the enhancer, and the binding regions mostly regulate the expression of the nearest gene, it is possible to predict the regulated genes based on proximity. It is important to note, that assigning enhancer-promoter pairs in this case is only estimated or assumptions and may lead to false assignments in some cases. Nevertheless, it is a nonexperimental approach, performed with simple bioinformatic methods resulting in correct assignments in notable part of the enhancer-promoter pair assignments, thus it is a commonly used approach, also used in our study.

IRF9 ChIP-seq data was used for our analysis. Additional ChIP-seq experiments were also performed using antibodies against STAT1 and STAT2. A total of 27546 IRF9 binding regions were identified in IFNβ-stimulated DCs (consensus set: common regions of replicas). We identified 9289 regions in which binding of all ISGF3 subunits (IRF9, STAT1 and STAT2) were

detected in IFNβ-stimulated DCs. We found 78 IRF9 binding regions for which an effector ISG was assigned as the closest gene by HOMER algorithm. We found that 34 of the 78 binding regions were located within 1 kb of the TSS (TSS-proximal IRF9 binding regions). The other IRF9 binding regions were located within a greater distance, a maximum of 33 kb. We determined the binding of ISGF3 subunits and found that intensities of STAT1 and STAT2 signals in these regions were correlated with IRF9 signal intensities.

Notably, the IRF9 signals were higher at TSS-proximal regions than at TSS-distal regions. A majority of effector ISGs (n=30) were associated with an IRF9 binding site located <1kb from TSS. Effector ISGs were classified based on the number of associated IRF9 peaks. As previously mentioned, there were 3 genes without any associated IRF9 peaks. Most effectors ISGs (n=25) were associated with one or two IRF9 peaks, while 9 effector ISGs with three or more IRF9 peaks.

4.7. ISRE and TISRE in the regulatory regions of effector ISGs

We investigated the presence of ISRE sequence (5′-GAAANNGAAA-3′) in IRF9 binding regions belonging to ISGs. ISRE sequence was identified in 43 of 78 IRF9 binding sites associated with ISGs. This frequency (55.1%) was higher than the frequencies detected with binding regions of other ISGs. Using the HOMER program, we identified the sequences in each IRF9 binding sites that were the most similar to ISRE motif. We also determined the motif score, which reflects the similarity of the identified sequence to the motif matrix. We found that motif score was higher in IRF9 binding sites located in effector ISGs promoter regions than in the IRF9 binding sites of effector ISGs. Calculating the frequency of TISRE, we found that it was more than three times more common in IRF9 binding sites associated with effector ISGs compared to other ISGs (29.5%, in 23 of 78 binding sites). The following 22 effector ISGs contained the TISRE motif in the associated IRF9 binding regions: *Bst2*, *Ddx60*, *Eif2ak2*, *Gbp2*, *Gbp3*, *Ifit1*, *Ifit2*, *Ifit3*, *Isg15*, *Isg20*, *Mx1*, *Mx2*, *Oas1a*, *Oas1c*, *Oas1g*, *Oas3*, *Oasl2*, *Rnasel*, *Samhd1*, *Trim21*, *Trim5* and *Zc3hav1*. Most TISREs differed by only 1-2 bases from the canonical 5’-GAAANNGAAANNGAAA-3 ’sequence.

4.8. Chromatin accessibility in regulatory elements of effector ISGs

We focused on three questions related to the accessibility of the regulatory regions of effector ISGs: (1) To what extent are TSS-proximal and TSS-distal IRF9 binding regions of effector ISGs accessible before simulation in DCs? (2) To what extent are TSS-proximal and TSS-distal IRF9 binding regions of effector ISGs accessible in other cells? (3) To what extent are the promoter regions of different gene sets accessible in DCs and other cells? In order to answer these questions, we used our ATAC-seq data and publicly available ImmGen data.

To answer our first question, we determined the the frequency of accessible regions in various binding site sets. We found that large proportions of the entire IRF9 binding sites (ATAC-seq positive) were accessible before stimulation (69.0%), suggesting that similarly to other SRTFs, ISGF3 binds to pre-existing accessible (either active or primed inactive) regions to a major extent. TSS-proximal IRF9 peaks associated with effector ISGs were more frequently accessible (79.4%) than the TSS-distal IRF9 peak set (65.9%). Comparing ATAC-seq signal intensities of TSS-proximal and TSS-distal IRF9 peaks we found no difference.

To answer our second question, we assessed the accessibility of IRF9 binding sites associated with effector ISGs using ImmGen ATAC-seq data. We calculated the ATAC signals of 34 TSS-proximal and 44 TSS-distal IRF9 binding regions in 92 different immune cell types. We used the coefficient of variation (CV) values to quantify cell-to-cell variances. We found that ATAC signals at the TSS-proximal IRF9 peaks of a given effector ISG were typically less variable (resulting in lower CV values) than TSS-distal IRF9 peaks. These data suggest that many TSS-distal regulatory elements are primed in a cell-specific manner. Binding of ISGF3 to these site show cell-specific binding patterns and targeted gene activation.

To answer our third question, we analysed accessibility of promoter regions of DCs and various cell types. In these analyses two additional gene lists, housekeeping genes (HK) (n=31) and inflammation-related (IR) (n=36) genes were analysed as references. We calculated the median of ATAC signals in DCs and ATAC-seq signal CVs in the 92 studied cell types. We found that the median values of ATAC signals were similar to the values of ‘other ISGs’ set and higher than the values of the highest FC ISGs and IR-cytokines. HK genes were more homogenous (lower CV values with smaller SD) than any other set with respect to CVs of ATAC signals. CVs of ATAC signals were higher in IR-cytokines than is other sets.

5. Discussion

In our first study, the investigated IRFs, namely IRF3, IRF5 and IRF9 play distinct but partially overlapping roles in the regulation of antiviral cytokines, inflammatory responses, and cell-intrinsic antiviral immunity. Our result showed that DNA motifs recognized by IRFs and additional features, for example binding of cofactors and chromatin accessibility, are involved in mediating IRF-specificity. It is important to mention that the regions of IRF-dominant clusters were diverse, and not all regions shared the prototypic pattern of features. The diversity of the regions and the fact that none of the features occurred exclusively in one cluster indicate that a series of features in a combinatorial fashion, rather than exclusively highly polarized DNA sequences, mediate IRF-specific binding. We identified and compared the IRF3-, IRF5- and IRF9-dominant regions. Among the three dominant clusters, IRF5-dominant cluster showed a fundamental difference from IRF3 and IRF9-dominant clusters. IRF5 protein can occupy genomic regions through three potential mechanisms: as a homo- or heterodimer, as a monomer or by indirectly binding to ISRE. Our result showed that the ISRE motif was enriched in IRF3 and IRF5 in shared and common clusters indicate that some of the IRF5 proteins bind to the canonical ISRE motif as a dimer. Observations that ISRE half-sites were identified by *de novo* motif discovery in the IRF5 cistrome, and the IRF5 logo from PBM study showed similarity to ISRE half site motif (GAAA), do not necessarily mean that IRF5 protein binds dominantly or exclusively as monomer to DNA. According to Andrilenas et al. the shorted logo represents a dimeric site, resulting in stronger binding of IRF5 protein to ISRE half-site. Thus, based on PBM experiments IRF5 binds as a dimer and prefers asymmetric ISRE half-sites. According to the third mechanism IRF5 does not bind DNA directly, its binding site is determined by other TFs. This mechanism has been shown for IRF3 protein, which functions as signal-specific cofactor in transcriptional activation of NF-κB-dependent genes without binding to ISRE motif. Further experiments are needed to confirm IRF5-mediated gene regulation by indirect binding of this protein.

In our study we found IRF5-dominant clusters to have three characteristic features. First, most of the IRF5-dominant regions do not contain ISRE motifs. A recent study showed that bases other than C at the 5’ and 3’ ends of ISRE half-site (5-CGAAAC-3) could prevent binding of IRF5 to the virus response element of IFN promoter. Our analysis demonstrated that that not only ISRE variants containing bases other that C at these positions (such as GAAANGGAAA), but all ISRE variants were underrepresented in IRF5-dominant cluster. Next, we found that special ISRE half-sites, such as 5’-GAGA-3’ and 5’-GACA-3’ were enriched in IRF5-dominant cluster compared to other clusters. Further studies are needed to confirm the suggestion and molecular mechanism according to which IRF5 binds to special 6-mers and whether IRF5 protein as a monomer could

bind these ISRE half-sites. Finally, we found that NFKB motif was enriched in IRF5-dominant cluster, which was consistent with the observation that IRF5 binding is largely based on interactions with NF-κB.

Our results demonstrated that IRF3- and IRF9-dominant regions were more similar to each other, and IRF3- and IRF9-dominance was less predictable. There could be several reasons for the lower accuracy in prediction by machine learning. Clustering could be inaccurate due to differences in ChIP efficacy or other reasons. Accurate prediction may also require features that were not investigated. Such missing features could include the number of motifs in a certain region, distance between different motifs, the occupancy values of TFs, the presence of additional motifs or other variants of ISRE or 6-mers. In the future, a more complex analysis may reveal additional features needed for a more accurate prediction. Our analysis identified patterns, which favour either IRF3-dominant or IRF9-dominant binding.

In the IRF3-dominant cluster we found that many of the binding regions had low chromatin accessibility prior to stimulation. The high frequency of NFKB, AICE, TRE and CRE motifs in the cluster indicate that IRF3 requires other cofactors to open the chromatin. An alternative solution may be tandem binding of three or four IRF3 or other IRFs to the TISRE motif, which are also enriched in the cluster and may be sufficient to remove the nucleosome. IRF3-dominant regions were not bound by IRF9 with high affinity, probably because NF-κB and AP-1 are not induced in the I-IFN pathway. Furthermore, we found that many IRF9-dominant clusters showed enriched ISRE and GAS motifs. Although the consensus sequence of 5-WBVGGAAANNGAAACT-3 and its variants were enriched in the IRF9-dominant cluster, but their frequencies were low. The enrichment of the GAS-motif indicates that the STAT1 homodimer activated by I-IFN signalling are important determinants for IRF9-dominant binding.

In our second study we investigated regulatory elements of effector ISGs. We found that the promoter regions of most effector ISGs (81.0%, 30 of 37 genes) have an ISGF3 binding regions that contains ISRE motif and is accessible in several cell types. Promoter regions may contain binding sites not only for general TFs, but also for promoter-specific TFs (such as SP-1, YY-1 and NFY), LDTFs and other TFs. In addition to DNA sequences, the accessibility of genomic regions is crucial for SRTF binding. Promoters that are made accessible by promoter-specific TFs do not depend on LDTFs, so they can be activated regardless of cell type. IRF9 plays an important role in maintaining accessibility of the promoter regions of ISGs, because in many promoters (for example promoters of Mx1, Mx2, Oas1a and Oas2), where IRF9 was bound in resting wild-type (WT) macrophages, the chromatin accessibility was reduced in *Irf9*$^{-/-}$ cells.

Only a smaller set of effector ISGs (23.0%, 9 of 39 genes) was associated with two or more TSS-distal elements. TSS-proximal and TSS-distal IRF9 binding sites differed in some aspects.

The scores and frequency of ISRE motifs were lower in TSS-distal IRF9 binding sites than in TSS-proximal regions. The ‘weaker’ ISRE motifs coupled with lower IRF9 binding. Furthermore, their accessibility was more cell type-specific than the openness of TSS-proximal peaks. Since ISGF3 binds mainly to pre-existing accessible regulatory regions it is likely that ISGF3 occupies different TSS-distal regulatory regions in different cell types.

6. Summary

In our first study, we determined the features mediating IRF-specific enhancer selection. To identify regions occupied predominantly by IRF3, IRF5 or IRF9, we performed ChIP-seq experiments in activated murine DCs. The identified regions were analysed with respect to the enrichment of DNA motifs, ISRE and ISRE half-site variants, and chromatin accessibility. Using a machine learning method, we investigated the predictability of IRF-dominance. We found that IRF5-dominant regions differed fundamentally from the IRF3- and IRF9-dominant regions: ISREs were rare, while the NFKB motif and special ISRE half-sites, such as 5′-GAGA-3′ and 5′-GACA-3′, were enriched. IRF3- and IRF9-dominant regions were characterized by enriched ISRE motif and lower frequency of accessible chromatin. The IRF3- or IRF9-dominant regions were similar to each other, but for example some special ISRE variants, the TISRE and NFKB motifs were more frequent in case of IRF3, while GAS motifs and certain ISRE variants in case of IRF9.

In our second study, we used a multi-omics approach to identify the (epi)genetic features that permit robust and widespread transcriptional regulation of IFN-inducible antiviral effectors. We determined the location of regulatory elements, the DNA motifs, the occupancy of ISGF3 subunits (IRF9-STAT1-STAT2) and other transcription factors, and the chromatin accessibility in murine DCs. The ISRE and TISRE occurred more frequently in the regulatory elements of effector ISGs than in any other tested ISG subsets. Chromatin accessibility at their promoter regions was similar to most other ISGs but higher than at the promoters of inflammation-related cytokines, which were used as a reference gene set. Most effector ISGs (81.1%) had at least one ISGF3 binding site proximal to TSS, and only a subset of effector ISGs (24.3%) was associated with three or more IRF9 binding regions. The IRF9 signals were typically higher and ISRE motifs were ‘stronger’ in TSS-proximal versus TSS-distal regulatory regions. Moreover, most TSS-proximal regulatory regions were accessible before stimulation in multiple cell types. Our results indicate that ‘strong’ ISRE motifs and universally accessible promoter regions that permit robust, widespread induction are characteristic features of effector ISGs.

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/458/2022.PL
Subject: PhD Publication List

Candidate: Mária Nagyné Csumita
Doctoral School: Doctoral School of Molecular Cellular and Immune Biology
MTMT ID: 10057535

List of publications related to the dissertation

1. Göczi, L., **Csumita, M.**, Horváth, A., Nagy, G., Póliska, S., Pigni, M., Thelemann, C., Dániel, B., Mianesaz, H., Varga, T., Sen, K., Raghav, S. K., Schoggins, J. W., Nagy, L., Acha-Orbea, H., Meissner, F., Reith, W., Széles, L.: A Multi-Omics Approach Reveals Features That Permit Robust and Widespread Regulation of IFN-Inducible Antiviral Effectors.
   J. Immun. 209 (9), 1-12, 2022.
   DOI: http://dx.doi.org/10.4049/jimmunol.2200363
   IF: 5.426 (2021)
2. **Csumita, M.**, Csermely, A., Horváth, A., Nagy, G., Monori, F., Göczi, L., Orbea, H. A., Reith, W., Széles, L.: Specific enhancer selection by IRF3, IRF5 and IRF9 is determined by ISRE half-sites, 5' and 3' flanking bases, collaborating transcription factors and the chromatin environment in a combinatorial fashion.
   Nucleic Acids Res. 48 (2), 589-604, 2020.
   DOI: http://dx.doi.org/10.1093/nar/gkz1112
   IF: 16.971

DEBRECENI EGYETEM * Egyetemi és Nemzeti Könyvtár *

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

List of other publications

3. Póliska, S., Besenyei, T., Végh, E., Hamar, A. B., Karancsiné Pusztai, A., Váncsa, A., Bodnár, N., Szamosi, S., **Csumita, M.**, Kerekes, G., Szabó, Z., Nagy, Z., Szűcs, G., Szántó, S., Zahuczky, G., Nagy, L., Szekanecz, Z.: Gene expression analysis of vascular pathophysiology related to anti-TNF treatment in rheumatoid arthritis.
Arthritis Res. Ther. 21 (1), 94, 2019.
IF: 4.103

Total IF of journals (all publications): 26,5
Total IF of journals (publications related to the dissertation): 22,397

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

24 October, 2022

DEBRECENI EGYETEM * Egyetemi és Nemzeti Könyvtár *