TUM SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY

TECHNICAL UNIVERSITY OF MUNICH

Dissertation

Data science approaches to decipher immune processes

Barbara Höllbacher

September 2023

HELMHOLTZ
MUNICH

MUDS
MUNICH SCHOOL FOR
DATA SCIENCE
HELMHOLTZ | TUM | LMU

TUM SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY

TECHNICAL UNIVERSITY OF MUNICH

Data science approaches to decipher immune processes

Barbara Höllbacher

Complete reprint of the dissertation approved by the TUM School of Computation, Information and Technology of the Technical University of Munich for the award of the

Doktorin der Naturwissenschaften (Dr. rer. nat.)

Chair: Prof. Dr. Niki Kilbertus

Examiners:

1. TUM Junior Fellow Dr. Matthias Heinig
2. Prof. Dr. Julien Gagneur

The dissertation was submitted to the Technical University of Munich on 27.09.2023 and accepted by the TUM School of Computation, Information and Technology on 09.02.2024.

Acknowledgments

This thesis would not have been possible without the many wonderful people that supported me throughout this journey.

Foremost, I want to acknowledge my supervisor **Dr. Matthias Heinig**. Thank you for providing me with the environment I needed to learn many new things and encouraging me to tackle challenging tasks. You gave me the freedom to pursue my passion projects and took time for me when I needed your input.

I want to thank **Dr. Markus List**, for being an approachable and empathetic member of my thesis advisory committee who always had my back.

Thank you **Dr. Franziska Freulich** and **Benjamin Strickland** for your useful input.

I want to acknowledge my collaborators **Dr. Talyn Chu** and **Dr. Dietmar Zehn**. It was a pleasure to work on these exciting projects together.

From the Helmholtz Munich Bioinformatics Core facility I want to thank **Dr. Thomas Walzthöni** and **Xavier Pastor Hostench** for the helpful discussions.

My position was funded through the **Munich School for Data Science** and their administrative team of **Mara Kieke** and **Dr. Julia Schlehe** is truly wonderful.

I would not be where I am today, without my previous supervisors **Dr. Daniel Campbell** and **Dr. Iris Gratz**. Thank you for taking the time to teach me new things and introducing me to the fascinating world of immunology. You taught me the value of collaborations and your enthusiasm for scientific research served as a source of inspiration that motivated me to pursue a PhD.

During my PhD I was surrounded by an awesome group of colleagues. I would like especially thank **Corinna** and **Katharina** who took the time to proofread parts of this thesis and give me feedback. Katharina offered emotional support and always took time for cathartic dance evenings: I am glad to call you a close friend.

I want to express my sincere gratitude to my parents, **Marianne and Hermann Höllbacher**, for their unwavering support, love, and encouragement throughout my PhD. Your financial assistance gave me freedom to chart my own path and pursue my education. Thank you for always providing me with a safe haven and believing in my abilities.

My social support here has been the foundation
that made me endure and obtain education.
I've gained many skills and have mentally grown,
learned I'm independent, but I'm not alone.
- Barbara Höllbacher

Abstract

The immune system is complex and our knowledge gaps in the underlying regulatory mechanisms are complicating the treatment of many diseases. Large amounts of sequencing data help to bridge this gap but require new multi-OMICs data analysis strategies to draw meaningful biological conclusions from them. Our goal for this thesis is to devise project-specific data analysis strategies in a biologically informed way so that we can gain new insights into immunological mechanisms on the innate and adaptive immune system.

Within the innate immune system we investigate the transcriptional changes triggered by glucocorticoids. Glucocorticoids are widely used anti-inflammatory drugs, but their long-term treatment leads to severe side-effects. We look into macrophages and the gene regulation mediated through the glucocorticoid receptor (GR), to understand how it can simultaneously activate some target genes and repress others. We apply a combination of standard motif enrichment tools and a custom machine learning workflow to identify sequence determinants that drive gene expression changes. We find that while the NR3C1 motif is associated with GR-mediated gene activation, gene repression is more complex and involves factors of the activator protein 1 (AP-1), nuclear factor-kappa B (NF-κB) and signal transducer and activator of transcription (STAT) families. Taken together, our computational results and wet-lab experiments indicate that GR competes with STATs for DNA binding, which leads to a suppression of STAT target genes. While further validations are needed to confirm this conclusion, our findings improve our understanding of the immunosuppressive action of glucocorticoids and lay the necessary groundwork to engineer therapies with less side-effects.

Within the adaptive immune system we investigate CD8 T cell progenitor populations, specifically looking into mechanisms of exhaustion, which describes a hypofunctional T cell state that limits the effectiveness of cancer immunotherapies. We integrate bulk and single-cell sequencing datasets from public and private sources into exploratory analyses that are followed up by flow cytometry based validations. We find that progenitors of exhausted T cells are formed in acute infection and, to a small degree, maintained after the infection is resolved. This shows that a diverse progenitor repertoire is preemptively formed irrespective of the outcome of an infection and environmental factors shape which populations subsequently get maintained in high numbers. Correspondingly, we followed up on the environmental factors needed by non-exhausted cells and found that interleukin-2 treatment successfully maintains them in a chronic environment. This discovery not only sheds new light on the mode of action of existing compounds but more importantly, it identifies a novel key target population of immunotherapy approaches which can be expanded to circumvent exhaustion altogether.

Kurzfassung

Unser Ziel für diese Arbeit ist es, projektspezifische Datenanalysestrategien auf biologisch informierte Weise zu entwickeln, damit wir neue Erkenntnisse über immunologische Mechanismen im angeborenen und adaptiven Immunsystem gewinnen können.

Im angeborenen Immunsystems untersuchen wir die durch Glukokortikoide ausgelösten transkriptionellen Veränderungen. Glukokortikoide sind weit verbreitete entzündungshemmende Medikamente, aber ihre langfristige Anwendung führt zu schwerwiegenden Nebenwirkungen. Wir untersuchen Makrophagen und die durch den Glukokortikoidrezeptor (GR) vermittelte Generegulierung, um zu verstehen, wie er gleichzeitig einige Zielgene aktivieren und andere reprimieren kann. Wir wenden eine Kombination aus Standardwerkzeugen zur Testung von Motivanreicherung und einem massgeschneiderten Machine Learning workflow an, um Sequenzdeterminanten zu identifizieren, die Genexpressionsveränderungen steuern. Wir stellen fest, dass das NR3C1-Motiv mit der GR-vermittelten Genaktivierung assoziiert ist, während die Genrepression komplexer ist und Faktoren der activator protein 1 (AP-1), nuclear factor-kappa B (NF-κB) und signal transducer and activator of transcription (STAT) Familien einschliesst. Zusammenfassend deuten unsere Berechnungsergebnisse und Experimente im Labor darauf hin, dass GR mit STATs um die DNA-Bindung konkurriert, was zu einer Unterdrückung von STAT-Zielgenen führt. Während weitere Validierungen erforderlich sind, um diese Schlussfolgerung zu bestätigen, verbessern unsere Ergebnisse das Verständnis für die immunsuppressive Wirkung von Glukokortikoiden und legen die notwendigen Grundlagen für die Entwicklung von Therapien mit verminderten Nebenwirkungen.

Im adaptiven Immunsystems untersuchen wir CD8 T Zell Vorläuferpopulationen und konzentrieren uns speziell auf die Mechanismen der Erschöpfung, welche einen hypofunktionalen Zustand von T Zellen beschreibt, der die Wirksamkeit von Krebsimmuntherapien einschränkt. Wir integrieren Datensätze von Massen- und Einzelzell-Sequenzierungen aus öffentlichen und privaten Quellen in explorative Analysen, die durch Durchflusszytometrie-basierte Validierungen ergänzt werden. Wir stellen fest, dass Vorläufer von erschöpften T Zellen während einer akuten Infektion entstehen und in geringem Ausmass auch nach erfolgreicher Eliminierung der Infektion aufrechterhalten werden. Dies zeigt, dass eine vielfältiges Vorläuferrepertoir, unabhängig vom Ausgang einer Infektion, vorsorglich gebildet wird und Umweltfaktoren bestimmen, welche Populationen anschliessend in hoher Anzahl aufrechterhalten werden. Gleichermassen sind wir den Umweltfaktoren nachgegangen, die von nicht-erschöpften Zellen benötigt werden, und stellten fest, dass eine Interleukin-2 Behandlung sie erfolgreich in einer chronischen Umgebung aufrechterhält. Diese Entdeckung wirft nicht nur ein neues Licht auf die Wirkungsweise bestehender Präparate, sondern identifiziert vor allem eine neue zentrale Zellpopulation für immuntherapeutische Ansätze, die expandiert werden kann, um Erschöpfung gänzlich zu umgehen.

Preface

1. Structure and contributions

During the course of my PhD I had the privilege to work on a number of different scientific projects and publications. Several of them have been published already, one project is still under active investigation and one manuscript is waiting for submission. This work would not have been possible without my collaborators. However, in the context of this thesis, I will focus on the portion of the projects that I was involved in and will omit the work of others unless needed for conceptual understanding.

We start the thesis by introducing the biological background in Chapter 1 that is necessary to understand the projects discussed in this thesis. This includes a general overview of the immune system followed by an in depth description of macrophages, immune receptors and CD8 T cell exhaustion. Furthermore, it includes a section regarding processing workflows of next-generation sequencing techniques whose content is similar to a review we published on this topic:

Höllbacher, B., Balázs, K., Heinig, M., & Uhlenhaut, N. H. (2020). Seq-ing answers: Current data integration approaches to uncover mechanisms of transcriptional regulation. Computational and Structural Biotechnology Journal, 18, 1330–1341. https://doi.org/10.1016/j.csbj.2020.05.018 [1]

This is followed by Chapter 2 on technical background, which covers common concepts in the field of machine learning and mathematical methods used within the projects.

After this general introduction we start by presenting our research on gene regulation in macrophages in 3. This chapter is similar to the corresponding publication:

Höllbacher, B., Strickland, B., Greulich, F., Uhlenhaut, N. H., & Heinig, M. (2023). Machine learning reveals STAT motifs as predictors for GR-mediated gene repression. Computational and Structural Biotechnology Journal, 21, 1697–1710. https://doi.org/10.1016/j.csbj.2023.02.015 [2]

Figure 3.6C and Supplemental Figure A.6 on western blots were made by Benjamin Strickland. All other visualizations and analyses shown in that section were performed by me.

Next, we present our work in the field of CD8 T cell exhaustion in Chapter 4. This chapter includes two projects that are related in their subject matter but represent to independent projects. The first part of the chapter (Section 4.1) covers some of the aspects discussed in the corresponding manuscript:

Wu, M.*, **Höllbacher, B.***, Wurmser, C., Berner, J., Donhauser, L., Bongers, L., Toppeta, F., Strobl, P., Heinig, M., Chu, T., & Zehn, D. (2023). Precursors of exhausted T-cells are preemptively formed regardless of the outcome of infection. Manuscript under submission. * equal first author contribution [3]

With the exception of Supplemental Figure B.1 showing the FACS gating strategy for sequencing, all included visualizations and analyses shown in that section were performed by me.

The second part of the chapter (Section 4.2) covers still ongoing work that is topically related but independent. While this project includes a lot of wet lab experiments, I will focus on the computational analyses and with the exception of Figure 4.9 all included visualizations and analyses were performed by myself.

Another projects that I contributed to during my PhD is:

Schmid, K. T., **Höllbacher, B.**, Cruceanu, C., Böttcher, A., Lickert, H., Binder, E. B., Theis, F. J., & Heinig, M. (2021). ScPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. Nature Communications, 12(1), 6625. `https://doi.org/10.1038/s41467-021-26779-7` [4]

Since I merely had a supporting role in that project and it was not the main focus of my PhD work, it is not discussed in this thesis.

The thesis concludes with a discussion of the individual projects as well as joint conclusions and outlook in Chapter 5.

Contents

1. Introduction

1.1. Immune system

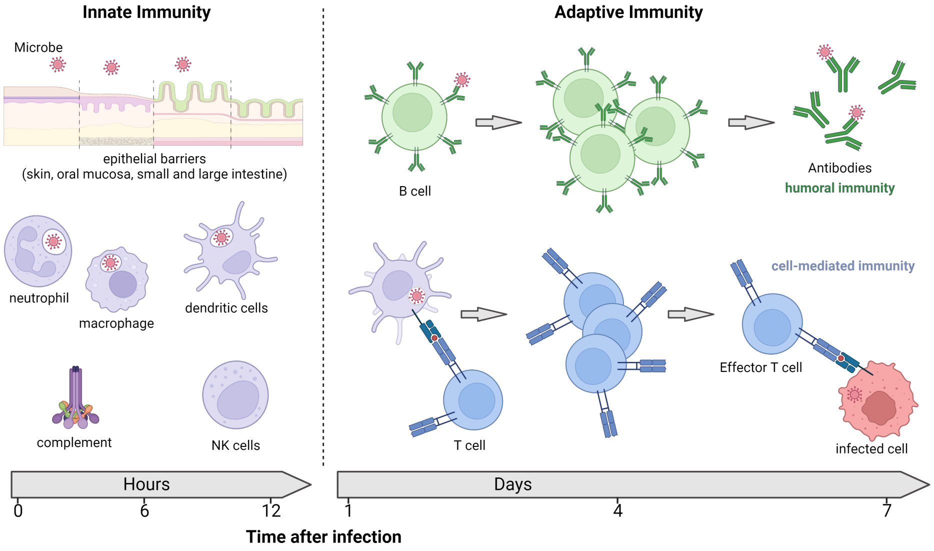

Figure 1.1.: **Innate and adaptive immunity.** The innate immune system provides a fast acting first line of defense which controls pathogens until the adaptive immune system is activated. Epithelial tissues form a physical barrier against pathogen entry. B cells and T cells specific for the antigen must undergo clonal expansion and differentiation into effector cells. Inspired by [5]. Schematic created with BioRender.

The role of the immune system is to protect the body from external hazards as well as malignant changes that come from within the body itself. In other words, the immune system fights off harmful infections by bacteria, viruses and parasites while concurrently eliminating mutated cells that might otherwise turn into tumors. In the fight against infections, the fast acting branch called innate immunity restrains infections while the slower but more specialized adaptive immune response gets mounted (see Figure 1.1).

Cells from the innate immune system are equipped with germ-line encoded molecular pattern-recognition receptors such as Toll-like receptors, whose specificity is genetically determined [6]. Among the innate immune cells with pattern recognition receptors are macrophages that digest debris such as dying cells and bacteria [7].

While the innate immune system is good at keeping most infections at bay by using conserved patterns, some pathogens escape these mechanisms and require the adaptive immune system to successfully clear the infection [8]. The adaptive immune system takes a couple of days to be fully activated and is made up by lymphocyte populations. The two interconnected parts of the adaptive immune system, the humoral (antibody-mediated) and cell-mediated immune system, are formed by B lymphocytes and T lymphocytes, respectively.

1.1.1. Macrophages

Tissue-resident macrophages are first created during embryonic development, when fetal-derived macrophages form specialized populations in various organs such as the liver, the skin and the central nervous system. In the adult organism, monocyte-derived macrophages can replenish those populations, by developing from hematopoietic stem cells in the bone marrow and differentiating into macrophages within their target tissue [7]. In fact, this process of generating macrophages from hematopoietic stem cells is leveraged in experimental models working with bone-marrow derived macrophages (BMDMs).

Macrophages express the pattern recognition receptor Toll-like receptor 4 which binds Lipopolysaccharide (LPS) and leads to upregulation of B7 molecules [9], a costimulatory signal promoting T cell activation. Furthermore, the signalling downstream of the pattern recognition receptors triggers the activation of inflammatory transcription factors (TFs) such as nuclear factor-kappa B (NF-κB), activator protein 1 (AP-1) and IRFs [10] and the production of pro-inflammatory cytokines.

In cases such as severe COVID-19 [11], autoimmune diseases [12] or asthma [13], excessive inflammation causes harm and requires treatment with immunosuppressive glucocorticoids such as the synthetic compound Dexamethasone (Dex). Glucocorticoids have long been shown to act on macrophages, which are among the most effective producers of inflammatory cytokines [14, 15], putting them at the center of anti-inflammatory treatments. While treatment leads to the clinically beneficial upregulation of anti-inflammatory genes and downregulation of inflammatory genes, long term Dex treatment comes hand in hand with side-effects affecting the hormone system and energy metabolism [16]. An incomplete understanding of the underlying molecular mechanisms have hindered the efforts of pharmaceutical companies to develop drugs with fewer side effects [17].

What has been established so far is that glucocorticoids bind to the glucocorticoidreceptor (GR), which is encoded by the gene *Nr3c1*. Upon binding its ligand, GR dimerizes and translocates into the nucleus where it exerts both, gene repression and activation [18]. For the case of gene activation, it is widely accepted that GR recognizes and binds to genomic sequences, coined GR response elements (GREs), and acts together with a plethora of cofactors to mediate gene transcription [19, 20, 21, 22]. For the case of gene repression there currently exist a number of contradictory explanations. While some research suggests that GR-mediated gene repression is accomplished through DNA-binding independent tethering of GR to the inflammatory TFs AP-1 and NF-κB [23], other evidence shows that direct DNA-binding is required for GR-mediated suppression [24].

The concept that the transcriptional outcome is encoded in the DNA sequences was initially suggested when scientists observed that in addition to the classical activating GREs, there exist GR-bound DNA sequences linked to gene repression, which were named negative GREs (nGREs) [25]. The existence of nGREs could not be validated by some groups [26, 27] but the idea of sequence-encoded repression was substantiated by luciferase reporter asssays [27].

Recent investigations have revealed that the repression mediated by GR is more intricate than originally perceived, involving additional factors such as epigenetic elements like chromatin structure and phase separation [28, 29]. Furthermore, changes in accessibility,induced by interactions with chromatin remodelers, also contribute to this repression [30]. The goal of our project described in chapter 3 is to combine all these epigenetic components to get a more complete view on GR-mediated gene regulation and identify novel co-regulators involved in repression. This knowledge could pave the road to developing therapeutics with fewer side effects.

1.1.2. Lymphocytes

Lymphocytes are the cells of the adaptive immune system and encompasses B and T lymphocytes, which mature in the bone marrow and the thymus, respectively. In adults, generation of novel T cells in the thymus decreases and the T cell population gets maintained through proliferation of already matured cells [31]. B lymphocytes are responsible for the humoral immunity, by producing a soluble version of its B cell receptor (BCR), referred to as antibodies or immunoglobulins, that bind extracellular antigens. T lymphocytes constitute the cell-mediated immunity and recognize antigens presented by other cells using their T cell receptor (TCR).

Antigen receptors

A functional antigen receptor requires successful rearrangement of two genetic loci; the α and β (or γ and δ) chains in T cells or the heavy and light chain of immunoglobulins in B cells. Throughout the lymphocyte maturation process several checkpoints ensure that the T cell contains a functional antigen receptor. Potentially useful clones are preserved through a process of positive selection, while clones that fail to produce functional antigen receptors are sent into apoptosis [32, 5]. Another crucial checkpoint during lymphocyte development is negative selection of clones that react strongly to molecules naturally occurring within the organism (i.e. self antigens). In the case of B cells, self reactivity can be undone through additional receptor editing [5]. In the case of T cells, self reactive clones are either sent into apoptosis or differentiate into regulatory T cells [32]. This special subset of T cells is crucial for preventing autoimmune disorders and harmful inflammation by acting on other immune cells and reducing their effector function.

The majority of T cells has a TCR composed of α and β chain. The exception is a small population of $\gamma\delta$ T cells, which is especially important in epithelial tissues and are situated at the intersection of innate and adaptive immune system [33]. The $\alpha\beta$ T cells consist of two major subsets; cytotoxic CD8 T cells and CD4 T cells. CD4 T cells exert their function by releasing small signalling proteins, so called cytokines, and acting on other immune and stromal cells. They can be further subdivided into specialized populations such as follicular helper T cells, conventional helper T cells (including Th1, Th2 and Th17 cells) and the previously mentioned regulatory T cells [34].

Both, BCRs and the $\alpha\beta$ TCRs, are highly variable antigen receptors. They get generated through a process called V(D)J recombination, which describes somatic rearrangement of gene segments of the antigen receptor genes. Variability in the receptor stems from different gene segment combinations that can be chosen, as well as the process of non-homologous end joining that repairs the double stranded breaks introduced during the recombination events [35]. In its native structure, the antigen specificity is driven through 3 hypervariable loops of each chain coming together to form the antigen-binding site. These loops are commonly referred to as complementary-determining regions (CDRs), with CDR3 sitting at the center of the antigen-binding site [32]. CDR3 spans across the V(D)J gene segment junctions which is why it is the most diverse CDR. The mechanisms of V(D)J recombination allow for a stunning theoretical TCR diversity in the order of 10^{15} [36].

In fact, the chance of the the same TCR being created independently more than once within the same individual is so low, that the TCR can be used as a unique identifier of the cells [37]. For this purpose, the TCR sequence is used to define T cells originating from the same clone, i.e. clonotypes, which can be leveraged to track cell progeny. In this vein, combining single-cell TCR sequencing (TCR-seq) (scTCR-seq) with single-cell transcriptomics analyses can provide insights about the developmental relationship of T cells [38]. In subsection 4.2.5 it lets us identify what progenitor population effector cells developed from based on their common clonal origin and allows us to investigate the link between TCR activation strength and gene expression in subsection 4.1.3.

CD8 T cells

Cytotoxic T cells, also referred to as CD8 T cells in accordance with the expression of the CD8 molecule on their cell surface, are specialized in recognizing intracellular bacteria and viruses (e.g. *lymphocytic choriomeningitis virus* (LCMV)) as well as malignant transformations within the host proteins which are a signal for tumor development. After successfully completing somatic rearrangement of their TCR, mature, naive CD8 T cells leave the thymus and patrol the body in search for their cognate antigen [39].

Meanwhile, antigen presenting cells (APCs) take up and process pathogens in the periphery. Antigen gets loaded onto their MHC molecules and the APCs are transported to the lymph nodes [40]. This coordinated effort leads to a near complete recruitment of antigen-specific T cells [41]. Engagement of the TCR with its cognate antigen loaded onto the MHC, together with costimulation through CD28 on the T cell with B7 molecules on the APCs and cytokines are the 3 signals needed to fully activate a naive CD8 T cell [42]. Activation leads to the generation of long-lived progenitors, that express the T cell factor 1 (TCF1, encoded by *Tcf7*) [43] and provide a pool of self-renewing cells with proliferative potential as well as terminally differentiated effector cells.

Effector CD8 T cells can migrate to the site of infection and perform effector function through secretion of effector cytokines (interferon (IFN)-γ and tumor necrosis factor (TNF)α) as well as cytolytic granzymes and perforins [42]. Antigen-specific cells expand 10^4 to 10^5

fold during the week after activation, but only a fraction of them will survive, once the virus is cleared [44]. These long-lived memory cells provide long-term protection and a fast recall response in case of reinfection [44]. Throughout chapter 4 we refer to long-lived cells during early phases of infection as progenitor populations, rather than memory populations, since at that point the antigen is not cleared yet.

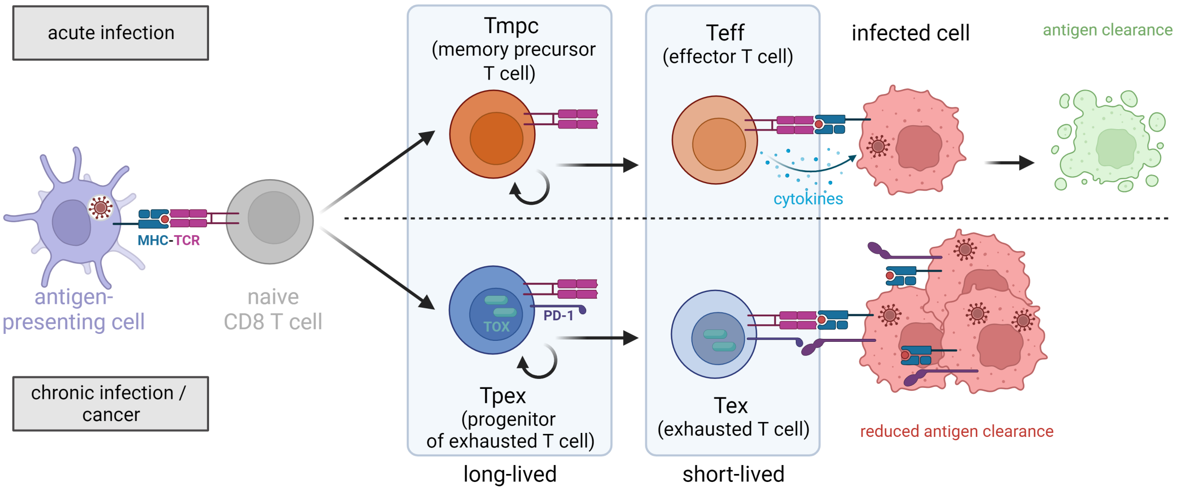

Figure 1.2.: **Current paradigm of T cell exhaustion.** Acute infection (e.g. with LCMV Armstrong) leads to a strong effector response and antigen clearance. Chronic infection (e.g. with LCMV clone13) leads to hypofunctional PD-1hi TOX+ CD8 T cells and reduced antigen clearance. Schematic created with BioRender.

CD8 T cells constantly patrol the body to fight off infections. Those that are successfully cleared are also referred to as acute infections, which are characterized by a strong antiviral response and robust CD8 effector function [44]. Opposed to this are chronic infections, (e.g. HIV, hepatitis C) where the antigen cannot get cleared and persists, leading to chronic antigen exposure. These two infection types are experimentally frequently studied through two strains of the virus LCMV. The Armstrong strain of LCMV leads to an acute infection, whereas the clone13 strain of LCMV leads to a chronic infection [45]. These two infection types are frequently used in combination with the so called P14 mouse model that contains a transgenic T cell receptor recognizing the gp33 epitope of LCMV. This system allows the controlled transfer of antigen-specific CD8 T cells into a host where we can recover them with an allelic marker of the gene *Cd45* and investigate their response to infection. In fact, we leverage the P14 system in multiple experiments within section 4.1 and section 4.2.

Chronic infections, as well as cancer, have been linked to hypofunctional effector T cells, so called exhausted T cells (Tex). They are marked by reduced effector function, reduced proliferative potential [46] and increased expression of inhibitory receptors such as *Pdcd1* (encoding for PD-1), *Ctla4*, *Lag3*, *Tigit* and *Havcr2* (encoding for TIM-3) [47, 46, 48, 49]. Transcription factor genes associated with exhaustion include *Tox*, *Tox2*, *Ikzf2* (encoding for

HELIOS) and *Nr4a2* [50]. Throughout section 4.1 and section 4.2 we use a combination of these markers to gauge the transcriptional exhaustion phenotype of cells. Furthermore, despite the many different nomenclatures present in the literature, we will refer to the progenitors of these exhausted T cells as Tpex and non-exhausted progenitors as memory precursor T cells (Tmpc) throughout this thesis (see also Figure 1.2).

The idea to target inhibitory molecules and thereby reactivate hypofunctional CD8 T cells and increase tumor control, led to the emergence of Immune Checkpoint Blockade. Reactivation of exhausted T cells by inhibiting interactions of PD-1 with its ligand PD-L1 has shown promising results in cancer immunotherapy [51]. Notably, the effect of PD-1/PD-L1 blockade is exerted by acting on the Tpex population[52, 53, 54]. Immune Checkpoint Blockade shows great effectiveness in some patients, while others do not respond or may experience immune-related adverse events that can affect any organ within the body [55].

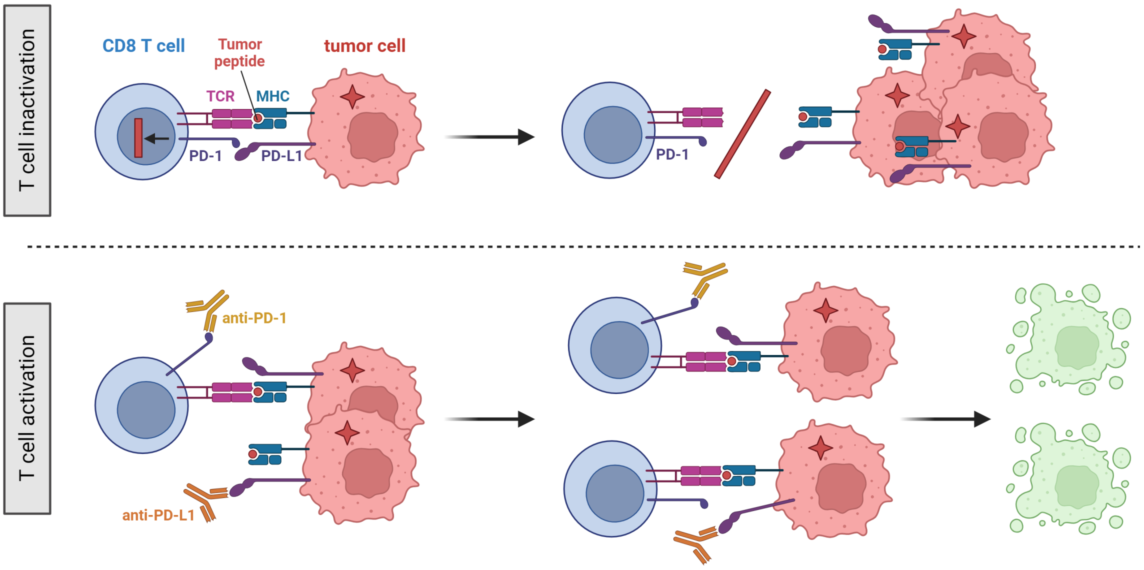

Figure 1.3.: **Immune Checkpoint Blockade.** Blocking antibodies suppresses the interaction of inhibitory receptors with their ligands, leading to a reactivation of CD8 T cells and increased antigen clearance. Schematic created with BioRender.

Alternative therapeutic approaches involve treatment with immunocytokines, either as monotherapy or in combination with checkpoint inhibition. Considering that both differentiation and maintenance of T cells is in large part guided by cytokines, it comes as no surprise that they would be used as therapeutic access point. Interleukin (IL)-2 was the first substance of cancer immunotherapy [56]. The IL-2 receptor can be either a low-affinity dimeric or a high-affinity trimeric receptor, composed of IL-2Rβ, the common γ-chain and either with or without the IL-2Rα chain (also known as CD25). IL-2 acts downstream through multiple pathways including the phosphoinositide 3-kinase (PI3K)–AKT pathway, the Janus kinase (JAK)–signal transducer and activator of transcription (STAT) pathway and the mitogen-activated protein kinase (MAPK) pathway [57]. IL-2 has a pro-inflammatory role by acting on

CD4 and CD8 T cells, but on the flip side also maintains regulatory T cells which express the high affinity receptor. This gives regulatory T cells an advantage when competing for low levels of endogenous IL-2 [57].

There exist multiple version of IL-2 that have been modified to either shift IL-2 binding further towards the trimeric receptor to offer a potential treatment for autoimmune diseases [58] or towards the dimeric receptor to increase its pro-inflammatory action. IL2v is a variant of IL-2 that has mutations at the interface with CD25 while leaving the binding interface with the β and γ chains intact [59]. This reduces its effect on regulatory T cells which have high levels of CD25 and diminishes the disadvantage of CD8 memory populations which have high levels of the dimeric receptor. Another way to target cytokines to a population of interest is by fusing it to an antibody. An example for this is the compound FAP-IL2v where IL2v is fused to an antibody against fibroblast activation protein (FAP) which is highly expressed in cancer-associated fibroblasts thereby targeting it to the tumor environment [60]. A way to combine the benefits of IL2v with checkpoint inhibition is by fusing it to PD-1, which shows promising results in mouse experiments [61]. We investigate data from multiple of these treatment options within subsection 4.2.5.

In recent years, we have learned that the thymocyte selection-associated high mobility group box (TOX) protein is a central transcription factor in exhausted T cells [49, 62] which made knocking it out seem like a promising therapeutic avenue. However it has since been shown that TOX is required for T cell maintenance in an environment of chronic antigen exposure and that numbers of progenitor T cells without functional TOX rapidly decline after the infection [48]. Stable conversion of exhausted T cells into fully functional memory cells remains a challenge [63] and understanding the early developmental processes leading to the formation of exhausted T cells in the first place would bring great clinical benefit to the field of chimeric antigen receptors (CAR) T cells. This novel immunotherapy is currently of limited use in the treatment of solids cancers since CAR T cells become exhausted in the tumor microenvironment [64].

1.2. Transcriptional regulation

All cells within an organism contain the same genetic material and yet the individual cell types have vastly different roles. Their cell identity get defined through tissue and cell type specific gene expression conferred by transcriptional regulators. These regulators include transcription factors, which are proteins that bind to the DNA at specific sequence motifs, promoter and enhancer regions within the DNA, histone modifications and DNA methylation as well as the chromatin accessibility and 3D structure (Figure 1.4).

The main flow of information is from DNA, which can be replicated or transcribed to RNA and from RNA which can be translated to protein [65]. There are some special circumstances such as reverse transcription in RNA viruses, where RNA can be reverted back to DNA. Nonetheless, in accordance with the central dogma of molecular biology, once sequential information has been translated into protein, the information cannot get out again [66]. Importantly, each step in this flow of information is regulated.

Various genome analysis technologies have been developed to investigate individual parts of this transcriptional regulation machinery. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) leverages antibody-mediated pull-down of DNA sequences to investigate TF binding locations and histone modifications [67]. Ribonucleic acid sequencing (RNA-seq) measures gene expression at the level of RNA transcripts [68] [69] and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) determines genome-wide chromatin accessibility through the transposase Tn5 which preferentially inserts at open chromatin sites [70]. There are also methods to assess the 3D interactions within the chromatin at regions of interest [71] or on a genome wide [72] scale.

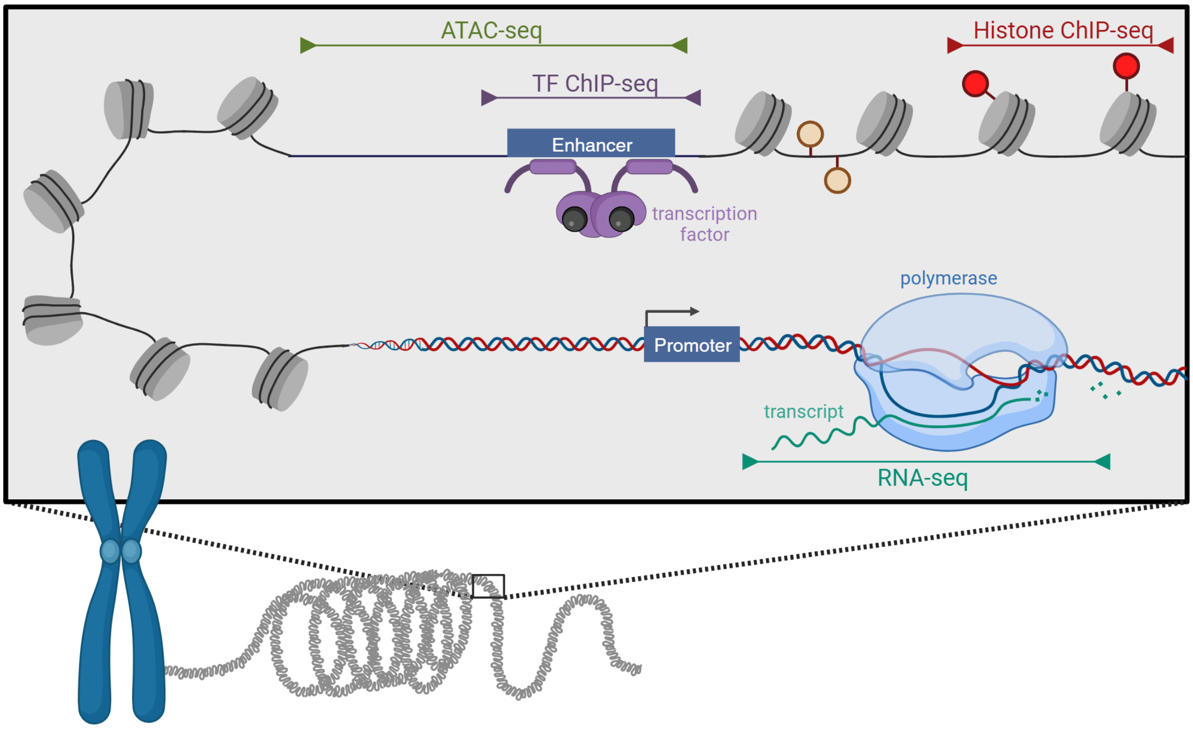

Figure 1.4.: **Contributors to gene regulation.** Cis-regulatory elements (enhancers or promoters), trans-regulatory elements (transcription factors) as well as epigenetic modifications and 3D chromatin structure are known to influence gene expression. Various sequencing methods can give insight into open chromatin regions (ATAC-seq), TF binding (TF ChIP-seq), histone modifications (Histone ChIP-seq) and gene expression levels (RNA-seq). TF = transcription factor. Schematic created with BioRender.

In the last decade, advancements in the next-generation sequencing (NGS) field lead to commercialization of single-cell RNA-seq (scRNA-seq) which allow transcriptome sequencing at single-cell resolution [73]. scRNA-seq is available in combination with assessment of cell surface protein levels, marketed as Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-seq) [74] or in combination with TCR-seq [37]. Simultaneously capturing these layers of molecular information allows to link the transcriptomic phenotype of cells to their TCR sequence. Furthermore, single-cell ATAC-seq (scATAC-seq), which measures accessibility at single-cell resolution, can also be performed either as multiome assay in combination with scRNA-seq or by itself.

Depending on the experimental question, researchers will assess samples in steady-state e.g. across multiple tissues or cell-types or investigate changes in response to a treatment. Common experimental approaches include the overexpression or knock-out of a gene of interest to see its effect on TF binding, histone modifications or gene expression levels. In other cases it is of interest to investigate treatment conditions that lead to changes in TF availability and subsequent changes in transcript levels.

1.2.1. ChIP-seq data processing

ChIP-seq is used to locate and quantify genome-wide interactions of DNA with a protein of interest, which can either be a transcription factor or modifications to histone tails. These histone modifications influence nucleosome positioning and gene regulation [75, 76]. Possible modifications to the histone tail are acetylation, methylation, phosphorylation and ubiquitination, some of which have an activating effect on gene expression while others lead to repression [77]. The concept that the combination of histone modifications form a crucial regulatory mechanism is also referred to as the histone code [78].

An antibody specific for the protein of interest is added to the sample and will bind to its target. Sequence fragments of the regions it binds to can be enriched by cross-linking of protein with DNA, followed by fragmentation and antibody-mediated pulldown. After that, the crosslinking can be reversed and the pulled-down fragments get turned into a sequencing library. After sequencing the reads, various quality control metrics are used to assess read quality. Commonly used metrics include the quality of the base calls, duplication rates, GC content and adapter content all of which are returned as part of the popular tool FastQC. Low quality bases and adapter sequences are trimmed from reads with tools like Trimmomatic [79] or Cutadapt [80] before mapping it to the reference genome with aligners such as bowtie2 [81].

Local enrichment of reads along the reference genome, which is indicative of TF binding / histone modifications, are referred to as peaks. The most commonly used peak calling algorithm is MACS2 [82] and uses dynamic Poisson distributions to determine fold enrichment over the background signal. By default MACS2 is set up for the narrow peak shape resulting from TF binding but it also offers parameters to accommodate the broad peak shape resulting from many histone modifications.

To ensure reproducibility of the experimental findings, it is recommended to sequence a pair of biological replicates, However, submitting more than two replicates rarely warrants the increased cost [83]. The concept of irreproducible discovery rate (IDR) assesses the agreement between both replicates by ranking the peaks and comparing them between samples. A predefined significance level α is then used as threshold to determine the number of reproducible peaks [84].

Biological interpretation of TF binding peaks requires functional annotation to its putative target genes. Some researchers choose to do this manually through visual inspection of region of interest with tools such as the University of California, Santa Cruz Genome Browser [85] or the Integrative Genomics Viewer [86]. Alternatively, tools such as the Bioconductor package ChIPseeker use linear distance to systematically annotate genome-wide peaks to target genes [87]. TFs can bind in promoter regions of genes but frequently bind to enhancers that can be located several kilobases or even megabases from the gene they regulate which can make correct functional annotation based on linear distance challenging. Hi-C [72] and Promoter capture Hi-C [88] tackle this shortcoming by assessing loop formation between distal genomic regions which allows to integrate 3D structure of chromatin in the process of peak annotation.

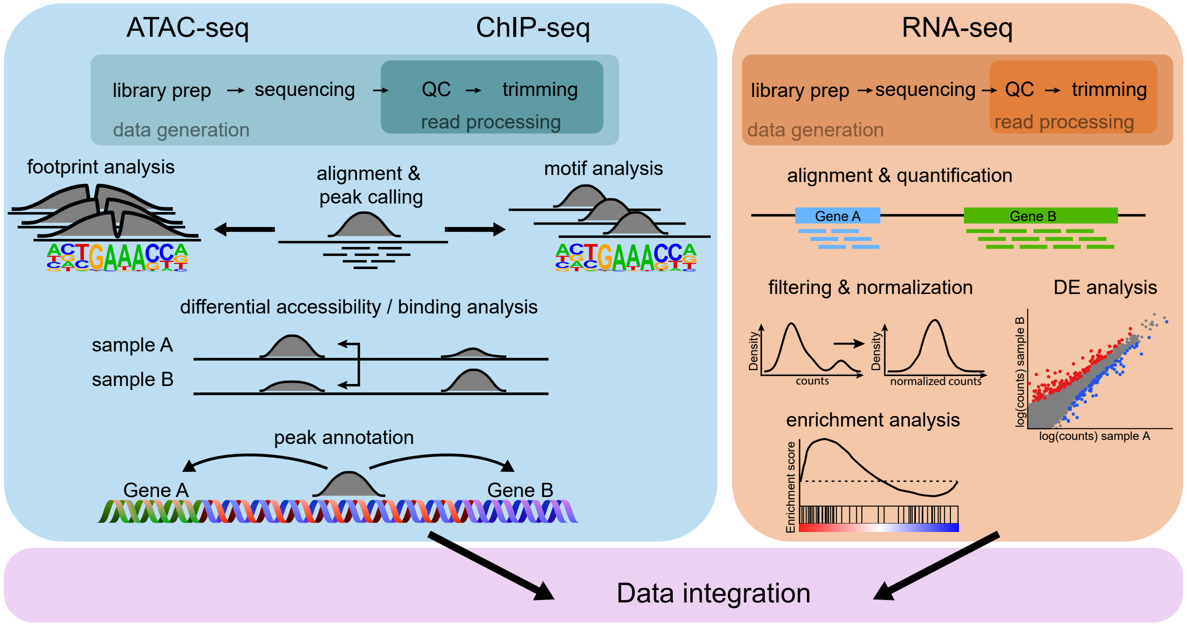

Figure 1.5.: **Standard processing workflow of ATAC-seq, ChIP-seq and RNA-seq.** In all cases, the quality of the sequenced reads is checked before performing the alignment. For ATAC-seq and ChIP-seq data analysis continues with peak calling, followed by differential accessibility and differential binding analysis, respectively. In ATAC-seq, accessible regions can be searched for footprints which are then matched to motifs. In ChIP-seq checking for motif enrichment within the peak regions and peak annotation are crucial steps. For RNAseq, the aligned reads are quantified at gene level, the raw counts are then filtered and normalized to enable further comparisons. The differential expression analysis provides a list of significant genes, from which biological meaning may be retrieved. QC: Quality control, DE: differential expression. Figure based on a schematic originally drafted for [1].

TF ChIP-seq has the potential to reveal regulatory factors, by looking for motifs (sequences typically 8–16 base pairs long) that occur more frequently than expected within the set of input sequences [89]. On one hand this can help identify the consensus motif of the TF targeted by the antibody used for the pulldown, on the other hand it can also find binding sites of co-factors that are enriched due to its interaction with the TF of interest Figure 1.5. The widely used tool MEME-ChIP [90] employs expectation maximization, and allows to either perform *de novo* motif discovery or test for the enrichment of already known motifs, deposited as position weight matrices in motif databases such as JASPAR [91]. Unfortunately, based on ChIP-seq data alone, it is hard to distinguish between direct interactions of the targeted TF with DNA and indirect interactions, where the TF is tethered to another DNA-binding factor.

1.2.2. ATAC-seq data processing

In eukaryotic organisms, DNA is compacted together with proteins into a complex named chromatin. The elemental subunit of chromatin are nucleosomes which consist of DNA tightly wound around an octamer of histones [92]. Depending on how tightly compacted the DNA is, the accessibility of encoded genes to TFs, polymerase and other regulatory proteins changes which in turn influences the genes' expression. This accessibility is assessed in ATAC-seq by using the enzyme transposase enzyme Tn5, which cuts at accessible chromatin regions and simultaneously inserts an adapter sequence. Its predecessor DNase-seq uses the enzyme DNase I [93], but contrary to the transposase it does not simultaneously insert sequencing adapters and hence includes more steps during library preparation.

Removal of low quality base pairs and trimming adapter sequences in ATAC-seq is performed analogous to ChIP-seq processing. For downstream analyses it is important to consider the sequence bias [94] involved in Tn5 insertion and account for it.

Besides peak calling, popular downstream analyses include so called footprint analysis. The idea behind this analysis is that TFs occupying the DNA prevent Tn5 from cutting. This results in the TF leaving a footprint, seen as sudden drop of read coverage within high-read coverage, nucleosome free areas [95]. The bound TF can be identifyed by matching these protected sequences, known as footprints, with established TF binding motifs [96]. HINT-ATAC identifies footprints in ATAC-seq data with hidden Markov models, while accounting for transposase-specific biases [95].

1.2.3. RNA-seq data processing

Next-generation sequencing was a breakthrough for transcriptomics studies. Information that could until then only been gathered in a targeted fashion using real-time quantitative reverse transcriptase polymerase chain reaction or microarrays, could now be collected on a genome-wide level without requiring prior knowledge.

It became possible to quantify different kinds of transcripts such as messenger RNA (mRNA), microRNA and noncoding RNAs, perform de-novo transcript assembly or perform isoform analyses [97, 98]. Variations of RNA-seq include 4-thiouridine labelling followed by sequencing (4sU-seq), a method where newly transcribed transcripts are labelled to quantify the levels of mRNA synthesis [99] and Cap-analysis gene expression followed by sequencing (CAGE-seq) which determines site of transcription initiation through sequencing of 5' end of capped transcripts [100].

Most commonly, researchers are interested in the levels of transcripts as proxy for protein levels and as such are mainly interested in quantifying mRNA. In that case, protocols to prepare sequencing libraries include steps to enrich for RNA molecules with poly-A tails or directly deplete ribosomal RNA [101]. Ribosomal RNA takes up the majority of transcripts within the cell and would reduce the sequencing depth available for other transcript types [68].

Trimming of low-quality bases and adapter sequences is performed as for ChIP-seq analysis. When aligning reads to a reference, they can either be mapped to a transcriptomic reference, possibly including multiple isoforms per gene, or to a genomic reference. Especially when mapping to a genomic reference, it is paramount that the alignment algorithm is splice aware, meaning the aligner is able to consider that the read might be spanning non-coding regions of the genome that got removed from the mature transcript through a process called splicing. Popular splice-aware alignment tools include STAR [102], TopHat2 [103] and Bowtie2 [81]. Once aligned, the number of reads overlapping the features of interest can be quantified with transcript-based or exon-based approaches such as Salmon [104], kallisto [105] or featureCounts [106], respectively. The output of the quantification step is a matrix with rows representing features, columns representing samples and the values in the cells representing read counts.

Depending on the experimental question at hand, different types of normalization methods can be used to correct for technical noise. If the goal is to compare different features within the same sample (e.g. expression level of gene A compared to gene B) it is crucial to correct for GC-content [107] and gene length (as longer genes have a bigger sequence that reads can map to). For comparisons between samples (e.g. to compare expression levels of the same genes in response to a treatment or perturbation) it is indispensable to perform between sample normalization. The most straight-forward way to approach this is by adjusting for the total number of reads in the sample which is also referred to as library size. The assumption in this approach is, that all samples have similar amounts of mRNA, which depending on the treatments and the physiological impact on the cells, may or may not be true. More elegant and popular tools to accomplish between sample normalization are edgeR's trimmed mean of M-values [108] and DESeq's Mean Ratio Normalization [109]. The former uses trimmed log expression values to calculate scaling factors that are not skewed by outlier genes, while the latter computes scaling factors through the median ratio of gene counts relative to the geometric mean per gene.

An intuitive way to explore the factors that contribute to variation in the data is to perform exploratory analyses with principal Component Analysis (PCA). To systematically test differential expression between conditions, the most widely-used methods fit a gene-wise Generalized Linear Model (GLM) based on a design matrix [109, 110, 111]. Approaches differ in how they account for dispersion and how flexible they are in terms of the experimental design. DESeq2 can test for complex design with multiple fixed effects, whereas limma additionally allows for the inclusion of a random effect (see also subsubsection 2.2.2).

1.2.4. Single cell sequencing data processing

While bulk sequencing methods are still widely used and lead to valuable biological insights, they discount the considerable cell heterogeneity present within biological samples. In cases where markers for celltypes contributing to this heterogeneity are known beforehand, this problem can be alleviated by enriching for the cells of interest before sequencing through a methods called Fluorescence-activated Cell Sorting (FACS). Often times the research goal is to investigate cell heterogeneity in an unbiased way and discover novel cell populations linked to a certain phenotype.

In this case, single-cell sequencing technologies are the method of choice and allow to investigate cells with unprecedented resolution and throughput. 10x is the leading company for microfluidic approaches to single-cell sequencing and works in combination with Illumina sequencers. They offer an array of assay types including scRNA-seq, scATAC-seq, CITE-seq and TCR-seq that can be used separately or in certain combinations. Recent additions to their portfolio include two different spatial transcriptomics solutions [112] (on the chromium and visium platform). The tool Cellranger is part of 10x' Chromium Single Cell Software Suite and represents a convenient analysis pipeline that performs sample demultiplexing, barcode processing, and feature quantification.

Cellranger includes the mkfastq command to convert raw sequence BCL files to FASTQ files which serve as main input to the various downstream pipelines, which have to be selected based on the chosen assay type. For gene expression assays the main outputs besides QC and preliminary clustering results, are a barcodes, features and matrix file, jointly referred to as count matrix, that can be used to for downstream processing with R or Python packages. The most common toolkit in Python is scanpy [113], while those in R are Seurat (for scRNA-seq) [114] and Signac (for scATAC-seq) [115].

The reason that Cellranger technically returns a barcode by feature matrix, rather than a cell by feature matrix is that, while each barcode would ideally label an individual cell, it is possible that a barcode labelled a cell doublet or an empty droplet instead. Still, empty droplets can actually be useful by giving insights into the ambient RNA of a sample. These cell-free mRNA molecules make there way into droplets by ways of background contamination from the solution that dissociated cells were contained in. Ambient RNA is an unwanted contaminant in the gene expression profile of sequenced cells and can be estimated from empty droplets with the tool SoupX [116].

When cells are dying, the integrity of the cell membrane can be lost and cytosolic mRNA can leak out, while RNA within the mitochondria stays behind. Those low quality cells can be removed by filtering out observations with a low number of counts and genes per barcode as well as those with a high fraction of mitochondrial reads [117]. On the other hand, observations with a high number of counts per barcode might be doublets. However, a superior way to remove cell doublets, rather than just excluding observations with high reads counts, is via specialized tools such as scDblFiner [118]. Cutoffs for QC metrics can be chosen manually after inspecting the distribution of QC metrics for the data at hand or can be performed with automated thresholds using median absolute deviations (MAD) [119].

Variance is not stable across expression levels; counts for highly expressed genes vary more than those of lowly expressed genes. There are different approaches to tackle heteroskedasticity [120], with the method SCTransform attaining variance stabilization through Pearson residuals [121].

After normalization of the gene counts, computation time and memory consumption can be reduced up by honing in on informative features. There exist multiple approaches to do this, with some workflows selecting the most variables genes, while other might choose to use deviance for feature selection [122]. From there, features can further be reduced by dimensionality reduction tools such as PCA [123]. Frequently, the principal components explaining most of the variance in the data are then used as input to compute a neighborhood graph which is in turn used to generate a UMAP embedding [124].

A 2D UMAP representation can help to visually assess cell similarity. Additionally, the same neighborhood graph can be used as input to the Leiden clustering algorithm [125] in order to systematically group the individual cells into clusters. These clusters are then annotated to cell types or states and frequently used as the grouping in differential gene expression analysis, with the goal to identify markers. By default Scanpy performs differential gene expression analysis based on t-tests combined with Benjamini-Hochberg correction for multiple testing, while Seurat opts for significance testing with the non-parametric Wilcoxon rank sum test.

1.3. Predicting gene expression

Transcriptional regulation involves a multitude of mechanisms on the genetic and epigenetic level. Promoter and enhancer regions, transcription factors, histone modifications as well as chromatin accessibility and structure are just some of the cogs in this complex machinery. As such it is paramount to integrate multiple assays, each providing an insight into a specific part of the process, to create a complete picture and gain understanding without missing certain aspects. Depending on the specific question at hand, various methods for integrating data can be employed. In light of the projects covered in this thesis, we will discuss existing methods for integrating data to predict gene expression.

In subsection 1.2.2 we mentioned tools that identify transcription factor binding sites by predicting footprints from accessibility data. Other groups expanded on this idea by deriving TF binding scores from open chromatin and use them to generate scores for regression models that predict celltype specific gene expression [126]. This approach created a link between accessibility, TF binding and gene expression using linear models.

The availability of extensive training data and advancements in high-performance computing, particularly the use of graphical processing units (GPUs) fueled the comeback of neural networks in genomic data analysis [127] and the development of new architectures such as convolutional neural networks. Neural networks are applied to a multitude of genomic tasks including the predictions of transcription factor binding [128], single-base resolution read coverage tracks [129] or mRNA levels [130, 131]. Xpresso [131] focuses on predicting steady-state median mRNA levels from promoter sequences in a tissue agnostic fashion, whereas ExPecto [130] uses a neural network to extract regulatory features which are passed to tissue-specific regularized linear models to predict gene expression.

Transformer architectures [132] revolutionized natural language processing and have recently found their way into genomics [133, 134]. Enformer [134] takes tissue-specificity into account by using a multi-task setting to predict thousands of epigenetic tracks. Gene expression is quantified by Cap Analysis Gene Expression (CAGE) assays which measures read counts at the transcription start sites. The authors looked into the performance regarding prediction of expression changes but had to conclude that it is difficult to predict fold-changes of highly correlated samples.

While neural networks are an exciting new avenue in genomic analysis, the application of existing models to a specific research question can be hampered by a variety of factors. Namely, in the context of this thesis, most models are not suitable for investigating treatment-induced transcriptional changes. They either investigate gene expression in a celltype agnostic fashion to find general patterns of transcriptional regulation or inquire the more pronounced cell type specific differences in steady state mRNA levels. The Enformer model allows conclusions about gene expression change, albeit with mediocre performance.

Another hurdle when trying to use the publicly available Enformer model, comes with projects focused on specific cell populations. Unfortunately, using the pretrained weights to predict gene expression requires that the celltypes of interest were part of the set used for model training. This prerequisite is hard to meet in the field of immunology, where projects include a myriad of highly specialized cell subsets and possible treatment conditions. Retraining transformer models is not only hugely expensive [133] but also requires data of the celltype of interest in the form of CAGE assays. Taken together, since predicting gene expression changes is still an open challenge, we decided to develop our own workflow to investigate perturbation induced changes in a biologically informed way that will identify validation targets.

1.4. Scope of this thesis

The immune system is a complex system and tight regulation is needed to hold the balance between pro- and anti-inflammatory processes and prevent pathologies.

An excessive immune response can lead to a variety of issues ranging from allergies and autoimmune disorders to cytokine storm, a life threatening condition with elevated levels of pro-inflammatory cytokines linked to mortality in severe COVID-19 cases [135]. Glucocorticoids are the mainstay of anti-inflammatory treatments but long-term administration can lead to unfavourable side-effects on the hormone and energy metabolism causing disorders such as diabetes [16]. On the molecular level it is unclear how signalling through the glucocorticoid receptor, can regulate gene expression in a fashion that leads to activation of some genes but repression of others.

In order to further the understanding of glucocorticoids' mode of action in macrophages we set out to:

- gather and integrate macrophage specific sequencing data from multiple data modalities
- build a machine learning model that links genomic sequence to transcriptional activation versus repression
- interpret the model to find and validate sequence determinants predicting the transcriptional response to glucocorticoid treatment

A reduced immune response of CD8 T cells is seen in the case of a phenomenon called CD8 T cell exhaustion. While this mechanism can be beneficial during chronic infection as it prevents tissue damage [48], the hyporesponsiveness poses a challenge in the setting of cancer where it leads to reduced tumor control. The mechanisms leading to exhaustion are currently not fully understood and therefore cannot be avoided. Furthermore, promoting non-exhausted T cells could offer a way to circumvent exhaustion but it is unknown how to maintain this population in a chronic environment.

In order to gain new insights into the mechanism of exhaustion we set out to:

- determine early transcriptional drivers of T cell exhaustion using multi-OMICs data
- design a neural network that can identify therapeutic targets using cell-type specific public RNA-seq, ChIP-seq and ATAC-seq data
- identify ligands that can maintain non-exhausted T cells in a chronic environment

Immune celltypes are highly specialized cells in a complex system. As such, incorporating domain knowledge and tailoring the data analysis to the projects at hand is key to the success of the projects. In the end, gaining a deeper understanding of immune processes is the first step to curing diseases that result from immune dysfunction and/or dysregulation.

2. Methods

2.1. Clustering

Clustering approaches are commonly used in bioinformatics to find structures in high-dimensional data. They fall into the category of unsupervised learning, meaning that we are lacking class labels to drive the learning process. The goal of clustering is to group objects into classes in order to maximize intraclass similarity and minimize interclass similarity [136]. Out of the many existing methods, I am elaborating on two that I used throughout this thesis.

2.1.1. Hierarchical clustering

Hierarchical clustering is frequently used in genomic analyses to order samples and/or features based on their similarity when displaying data in heatmaps. The process of hierarchical clustering is commonly represented by dendrograms, which are tree-like structures visualizing the distance between objects and their hierarchical grouping. The dendrogram can be created in a divisive fashion from top down or agglomerative from bottom up. In divisive methods, all objects start in the same cluster and get sequentially divided into smaller clusters until each cluster only contains a single object. This top down approach is computationally more expensive, making the bottom up alternative more popular [136]. In these agglomerative methods, every object starts as a singleton cluster. The pairwise distance between all initial clusters is considered and the closest pair (A,B) is merged into a new cluster C = A ∪ B. A and B are then removed from the current set of clusters and C is added to the set. In case the set of current clusters at this point only contains C, the algorithm is done, otherwise we continue determining the next closest pair and merging.

Which pair is deemed closest depends on the distance measure used. The metric we turn to when applying hierarchical clustering in the analyses of this thesis is the Euclidean distance L_2 between two d-dimensional points $\mathbf{x} = (x_1, ..., x_d)$ and $\mathbf{y} = (y_1, ..., y_d)$:

$$L_2 = \sqrt{\sum_{i=1}^{d} (\mathbf{x_i} - \mathbf{y_i})^2} \tag{2.1}$$

When using Euclidean distance as metric in hierarchical clustering, for the downstream formulas we set:

$$dist(\mathbf{x}, \mathbf{y}) = L_2(\mathbf{x}, \mathbf{y}) \tag{2.2}$$

For clusters that contain more than one data point, we additionally have to decide how to define distance between clusters. Let X and Y be the sets of data points of two clusters. Three popular definitions of distance between the sets X and Y are:

Single link:

$$dist_{sl}(X,Y) = \min_{\forall \mathbf{x}\in X, \mathbf{y}\in Y} dist(\mathbf{x},\mathbf{y}) \tag{2.3}$$

Complete Link:

$$dist_{cl}(X,Y) = \max_{\forall \mathbf{x}\in X, \mathbf{y}\in Y} dist(\mathbf{x},\mathbf{y}) \tag{2.4}$$

Average Link:

$$dist_{al}(X,Y) = \frac{1}{|X| * |Y|} * \sum_{\forall \mathbf{x}\in X, \mathbf{y}\in Y} dist(\mathbf{x},\mathbf{y}) \tag{2.5}$$

2.1.2. Leiden clustering

Clustering is a crucial step in single-cell data analysis as it assigns cells into groups that can then be annotated into celltypes or states before proceeding with downstream processing. A common way to approach this is through community detection on graphs. Due to the sparse and noisy nature of single-cell measurements, this is commonly preceded by a feature selection and dimensionality reduction step (see also section 2.5 and subsection 1.2.4). Principal component reduced expression space is used to calculate pairwise Euclidean distances and generate a graph representation by applying a K-Nearest Neighbour (KNN) approach [117]. The constructed graph is a tuple $G = (V, E)$ with a set of nodes V, which represents the individual cells and a set of edges E which indicate the connection of each cell to its k-nearest neighbors. A KNN graph is scalable to large cell numbers since it required considerably less memory than a cell by cell similarity matrix

Leiden clustering is a graph-partitioning algorithm aiming to find communities of densely connected groups within the network. Two of the objective functions the Leiden algorithm uses for community detection are modularity and Constant Potts Model (CPM).

Modularity aims to maximise the difference between the actual number and the expected number of edges within a cluster based on the total graph structure and is given by [125]:

$$Q = \frac{1}{2|E|} \sum_{c} \left(e_c - \gamma \frac{K_c^2}{2|E|} \right) \tag{2.6}$$

where:

- K_c is the sum of the degrees of the nodes in cluster c
- $|E|$ is the total number of edges in the graph
- $\frac{K_c^2}{2|E|}$ the expected number of edges in the cluster c
- e_c is the actual number of edges in cluster
- $\gamma > 0$ is referred to as the resolution

CMP on the other hand is given by [125]:

$$E_{CMP} = \sum_{c} \left[e_c - \gamma \binom{n_c}{2} \right] \tag{2.7}$$

where:

- E_{CMP} is the energy associated with the clustering
- n_c is the number of nodes and $\binom{n_c}{2}$ the number of all possible edges within the cluster c

For both quality functions choosing a lower resolution γ leads to fewer clusters and higher resolution leads to more clusters.

The Leiden algorithm starts with each node being a singleton cluster. It then moves nodes into clusters until a local maximum of the quality function is reached. This is followed by a refinement step that allows the initially found clusters to be subdivided into smaller partitions. After that it creates an aggregated graph where all nodes of the refined clusters are represented as a single node. The initial cluster assignments for the nodes in the aggregated network are taken from the unrefined cluster assignments. Individual nodes in the aggregated graph get moved and the steps are repeated until there is no further improvement. Contrary to the previously popular Louvain algorithm, which can result in internally disconnected clusters, this guarantees that the resulting clusters are connected [125].

2.2. Supervised learning

In supervised learning we want to learn the association of input variables with one or more observed output variables. To this end we choose a model class that is appropriate for the data distribution and the assumed relationship between independent variables and target variables. We train the model to find the best set of parameters by minimizing a loss function capturing the quality of the fit.

2.2.1. Linear regression

The goal of linear models is to use P predictor variables to predict a continuous response variable y_i for each sample $i \in 1, ..., n$. The assumption is that the predictors have a linear additive effect on the response variable.

$$y_i = \beta_0 + \sum_{p \in 1,...,P} \beta_p * x_{ip} + \epsilon_i \tag{2.8}$$

where:

- x_{ip} are the predictor variables
- β_0,...,β_p are the coefficients for the predictor variables
- ϵ_i is the error term of the sample

This is equivalent to the matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.9}$$

The Ordinary Least Squares criterion (OLS) estimates the coefficients $\hat{\boldsymbol{\beta}}$ by minimizing the sum of squared residuals (SSE) between the observed values of $\mathbf{y}$ and those predicted by the linear model.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) \tag{2.10}$$

We assume that the errors of the linear regression are independent and identically distributed (i.i.d.), following a normal distribution with

$$\epsilon \sim N(0, \sigma^2) \tag{2.11}$$

In this case the Ordinary Least Squares criterion and Maximum Likelihood Estimation are equivalent. Maximum likelihood estimation (MLE) is a widely used method to find the best coefficient vector $\boldsymbol{\beta}$ for the linear function linking predictor to target. MLE determines the values for the coefficients such that the likelihood function gets maximized.

In general terms, for a model with parameters $\boldsymbol{\theta}$ and observed data $\mathbf{X}$, the likelihood function $L(\boldsymbol{\theta}|\mathbf{X})$ is expressed as joint probability density function of the data $\mathbf{X}$, given the parameters $\boldsymbol{\theta}$.

$$L(\boldsymbol{\theta}|\mathbf{X}) = f(\mathbf{X}|\boldsymbol{\theta}) \tag{2.12}$$

To find the Maximum Likelihood Estimation of $\boldsymbol{\theta}$, we look for the values of $\boldsymbol{\theta}$ that maximize the likelihood function.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{X}) \tag{2.13}$$

In detail, the likelihood function for a linear regression model is:

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} * \exp\left(-\frac{1}{2\sigma^2} * (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \tag{2.14}$$

where:

- $\mathbf{y}$ is the response variable
- $\mathbf{X}$ are the predictor variables
- $\boldsymbol{\beta}$ is the vector of coefficients
- σ^2 is the variance of the error term

To estimate $\boldsymbol{\beta}$ and σ^2 using MLE:

$$\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 = \arg\max_{\boldsymbol{\beta}, \sigma^2} L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}, \mathbf{X}) \tag{2.15}$$

For the given estimated $\hat{\boldsymbol{\beta}}$ we can evaluate whether the input variable has a significant effect on the prediction. The null hypothesis $H_0 : \beta_i = 0$ assumes no effect compared to the alternative hypothesis $H_1 : \beta_i \neq 0$. Since the coefficient $\hat{\beta}_i$ divided by its standard error $se(\hat{\beta}_i)$ follows a Student't t-distribution, we can calculate a t-test statistic with:

$$\frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \sim t_{n-p-1} \tag{2.16}$$

We reject the null hypothesis and consider the coefficient significant, if the p-value of the t-test statistic is smaller than the chosen significance threshold α.

2.2.2. Generalized linear models

For cases where the response variable is not continuous and normally distributed, generalized linear models (GLMs) offer an extension to the linear regression framework to handle a wider range of response variables [137]. In this case, rather than modelling the relationship of predictors with a response variable directly, we model the relationship with a function of the response variable (also referred to as link function). In addition to the linear predictor and the link function another component of the GLM is its distribution family. The model predicts the mean value and the distribution family specifies how the residual error is distributed. In general, a GLM can be written as:

$$g(E(\mathbf{y})) = \mathbf{X}\boldsymbol{\beta} \tag{2.17}$$

where:

- $\mathbf{y}$ is the response variable
- $\mathbf{X}$ is the matrix of predictor variables
- $\boldsymbol{\beta}$ is the vector of coefficients
- g() is the link function
- E(Y) is the expected value of y given the predictor variables

The coefficients of a GLM are usually estimated using MLE.

Logistic regression

To predict a binary response variable, we can use a variant of GLMs called logistic regression. The success probability of a Bernoulli random variable P(y=1) is mapped to the continuous outcome of the linear function using the logit function:

$$logit(P(y_i = 1)) = \log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \beta_0 + \sum_{p \in \{1,\ldots,P\}} \beta_p * x_{i,p} \tag{2.18}$$

Accordingly, we can transform it with the inverse link function, the logistic function, to predict the probability of an observation to be in class 1.

$$P(y_i = 1) = \frac{1}{1 + exp(-(\beta_0 + \sum_{p \in \{1,...,P\}} \beta_p * x_{i,p}))} \tag{2.19}$$

In the logistic regression model the effect sizes β_p are log odds ratios and we can interpret them such that a unit increase in the predictor changes the odds of $\frac{P(y_i=1)}{P(y_i=0)}$ by $exp(\beta_p)$.

Negative binomial regression

Another variant of generalized linear models is the negative binomial regression, which is deployed in the differential expression (DE) method *DESeq2* [138]. DE analysis aims at quantifying gene expression differences between two or more biological conditions, where the groups can represent e.g. treatments, genotypes or perturbations.

The read counts K_{ij} of gene i and sample j are assumed to follow a negative binomial distribution with mean μij and dispersion parameter ϕ_i: $K_{ij} \sim NB(\mu_{ij}, \phi_i)$. Since the mean μ_{ij} is influenced by the read depth of the sample, we correct for it by estimating a size factor s_{ij} used for scaling. From there, we can estimate $\mu_{ij} = s_{ij} * q_{ij}$ and fit the model in a gene-wise manner:

$$log_2(q_{ij}) = \beta_{i0} + \sum_{p \in \{1,...,P\}} \beta_{ip} * x_{jp} \tag{2.20}$$

The variables x_{jp} are defined by the design matrix, which encodes information about the predictor variables. In the simplest case it indicates the sample assignment to one of two groups but the use of linear models allows for more complex designs. The returned coefficient represents the logarithmic fold change between the two groups. The null hypothesis is that there is no linear relationship between the grouping and the gene expression and that the coefficient is zero $H_0 : \beta_i = 0$. In DESeq2 the hypothesis testing is performed with a Wald test. The method shares information between genes to get more reliable estimates of the dispersion parameter ϕ_i in cases of small sample size.

2.2.3. Regularization

With increasing number of parameters, models run the risk of overfitting on the training data. This leads to lack of generalization or, in other words, poor performance on unseen data. To counteract this, it is common practice to regularize the model coefficients. By adding a regularization term to the loss function, the model is penalized for having too many parameters or large weights.

The two most common types of regularization are L1 (Lasso) regularization and L2 (Ridge) regularization. L1 regularization adds a penalty term that is proportional to the sum of the absolute values of the coefficients which encourages the coefficients of less important features to be shrunk to zero and results in feature selection. The penalty of L2 regularization is proportional to the sum of the squared values of the coefficients, which encourages the coefficients to be small, but not necessarily zero.

In GLMs these two regularization types can be combined in the Elastic Net approach. It uses a weighted combination of L1 and L2 penalties, which is especially useful if there are many correlated predictors in the model [139].

$$\hat{\boldsymbol{\beta_0}}, \hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta_0},\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} l(y_i, \boldsymbol{\beta_0} + \boldsymbol{\beta}^T x_i) + \lambda[(1-\alpha)||\boldsymbol{\beta}||_2^2/2 + \alpha||\boldsymbol{\beta}||_1] \tag{2.21}$$

where:

- α is the elastic net mixing parameter (α=0 corresponds to pure ridge regression and α=1 corresponds to pure lasso regression)
- λ is the tuning parameter controlling the overall penalty
- $l(y_i, \boldsymbol{\beta_0} + \boldsymbol{\beta}^T x_i)$ is the negative log-likelihood for each observation i
- $||\boldsymbol{\beta}||_1$ is the L1 norm of the coefficient vector $\boldsymbol{\beta}$
- $||\boldsymbol{\beta}||_2^2$ is the squared L2 norm of the coefficient vector $\boldsymbol{\beta}$

2.2.4. Model performance

In the supervised learning context, we can compare the predicted labels of a binary classifier with the ground truth to generate a confusion matrix with four categories Table 2.1.

Table 2.1.: **Confusion matrix with two categories.** The performance of a binary classifier can be evaluated by quantifying correctness of its predictions.

		Predicted Label	
		True	False
Actual Label	True	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

This categorization can be used to derive a number of different performance metrics.

$$\textit{TPR (true positive rate) / Recall / Sensitivity} = \frac{TP}{TP + FN} \tag{2.22}$$

$$\textit{FPR (false positive rate)} = \frac{FP}{TN + FP} \tag{2.23}$$

$$\textit{Precision} = \frac{TP}{TP + FP} \tag{2.24}$$

$$\textit{TNR (true negative rate) / Specificity} = \frac{TN}{TN + FP} \tag{2.25}$$

$$\textit{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.26}$$

As most classifiers output a continuous prediction score rather than discrete class labels, using these metrics as performance indicators has the disadvantage that they require chosing a threshold. An alternative evaluation method that is not influenced by the choice of threshold is the calculation of the Area Under the Receiver Operating Characteristic (ROC) curve, often abbreviated as AUC [140]. The ROC curve visually represents how a classifier's true positive rate relates to its false positive rate across various thresholds.

This process systematically explores all possible thresholds, creating a curve that spans from the point (0,0) to (1,1). The closer the curve approaches the optimal point of (1,0), which represents 0% false positive rate and 100% true positive rate and hence a perfect class separation, the better the classifier's performance. To quantitatively assess this performance, the AUC is computed. AUC values range from 0 to 1, with 1 indicating excellent performance and an AUC value of 0.5 corresponding to a random classifier (represented by a diagonal line in the ROC curve).

2.2.5. Cross-validation

After training a model, we want to see how well it generalizes to unseen data, by assessing the generalization error. This way we make sure we are not overfitting on the training data, which is especially an issue in models with many coefficients fitted on small datasets ($p >> n$) as is still the case for many genomics datasets. Especially in situations when our model selection process includes hyperparameter tuning, we need to ensure that information from the test set does not leak into the training process. One way to approach this it by splitting the data into a training, validation and test set. In this case the validation set can be used for hyperparameter tuning and the test set stays untouched for evaluation of the generalization error. This approach was taken in subsection 4.2.1 and gives as a point estimate for the generalization error.

Another way to approach this is with (nested) cross-validation, which assesses the variability of the estimates and its dependence on the data split. In k-fold cross-validation the data is split into k equally sized subsets. Each subset is iteratively used for validation, while all the others are used for training. This way we compute multiple error estimates for each model which, especially for small datasets, results in more robust results. In order to avoid information leaking into the training process when performing hyperparameter tuning, we can apply cross-validation in a nested fashion or set aside a separate test set before performing the k-fold split. In the case of GLMs used in chapter 3 cross-validation allowed us to find the optimal regularization parameter λ.

2.3. Neural Networks

Artificial Neural Networks are a class of machine learning models inspired by biology that consist of neurons, organized into multiple layers. The first layer (representing the input data) and the final layer (representing the prediction) are linked through one or more sequential hidden layers. The number of layers is also referred to as network depth. Neurons are connected to the previous layer by weights and each neuron performs two basic functions to create its output value. First, they compute the weighted sum of inputs and then they apply an activation function to that sum. These activation functions are non-linear transformations that allow neural networks to capture non-linear relationships in the data.

In a fully connected layer the output a of an individual neuron is calculated by:

$$a = \tau(\sum_{i=1}^{n}(w_i * x_i) + b) \tag{2.27}$$

where:

- τ is an activation function
- n is the number of neurons in the input layer
- x_i is the i-th input neuron
- w_i is the weight of the i-th neuron
- b is the bias

2.3.1. Training

The weights and biases of the network, referred to jointly as $\boldsymbol{\theta}$, are learned during model training, with the goal to find the model parameters $\boldsymbol{\theta}$ that minimize a chosen loss function L.

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L(\mathbf{y}, f(\mathbf{X}, \boldsymbol{\theta})) \tag{2.28}$$

where:

- $\mathbf{y}$ is the vector of true targets
- $f(\mathbf{X}, \boldsymbol{\theta})$ is the predicted target of the network given input $\mathbf{X}$ and model parameters $\boldsymbol{\theta}$.

Model training is an iterative process consisting of a forward and a backward pass of information. During the forward pass information from the input is fed into the model to create a prediction. This prediction is used to compute the empirical loss:

$$R_{emp} = L(\mathbf{y}, f(\mathbf{X}, \boldsymbol{\theta})) \tag{2.29}$$

During the backward pass, we use the loss as input and apply backpropagation to learn the gradient of each weight with respect to the loss. This gradient is used to update the weights so that the loss is reduced.

At a given step t, the model parameters $\boldsymbol{\theta}$ are updated to move towards the steepest descent using the calculated gradient $\nabla R(\boldsymbol{\theta}_t)$ and learning rate α [141].

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \nabla R_{emp}(\boldsymbol{\theta}_t) \tag{2.30}$$

Since computing gradient updates based on the entire dataset is computationally very expensive, this is usually done on a subset of the data and hence referred to as stochastic gradient descent.

In practice, stochastic optimization algorithms often include adaptive learning rates, by keeping track of past weight updates, and regularization through weight decay. In subsection 4.2.1 we used an adaptation of the popular Adam optimizer [142], called AdamW, that decouples weight decay from the gradient updates [143].

2.3.2. Convolutional Neural Networks

The introduction of convolutional neural network (CNN) architectures represents a major breakthrough for the field of image processing [144]. Contrary to fully connected layers, convolutional layers do not consider every node of the input layer when computing the output of a neuron, but instead use a kernel with width and height dimensions smaller than the input.

The input is frequently a multi-dimensional array, also referred to as tensor. In the case of image processing the input could be a RGB image with dimensions L x H x C, denoting the width, height and number of colour channels, respectively. In subsection 4.2.1 the input dimensions of the genomic sequences are L x C, where L is the length, C are the 4 channels representing the one-hot encoding of the nucleotide sequence.

The convolution operation is performed by sliding the kernel over the input and computing the dot product between the filter and the current position of the input. The kernel depth matches the number of input channels so that convolution leads to a single value for every position of the input it is applied to.

For a 3-dimensional input, the convolutional operation can be represented as:

$$\mathbf{Y}_{i,j} = \sum_{c=1}^{C} \sum_{l=1}^{L} \sum_{h=1}^{H} \mathbf{W}_{c,l,h} \cdot \mathbf{X}_{i+l-1,j+h-1,c} \tag{2.31}$$

where:

- **X** is the input layer
- **W** is the filter
- **Y** is the output layer
- C, L, and M are the dimensions of the filter
- i and j are the spatial coordinates of the output layer

The stride or step size parameter controls by what amount the kernel is moved along the input matrix after each convolution. Together with the kernel size and optional padding the stride determines the resulting output size W_{out} of the layer after the convolution.

$$W_{out} = \frac{W_{in} - K + 2 * P}{S} + 1 \tag{2.32}$$

where:

- W_{in} is the input size (height or width) of the input feature map
- K is the size (height or width) of the convolutional filter

- P is the padding, i.e. the number of zero values added to the edges of the input feature map
- S is the stride of the convolution operation

Convolutional filters can detect patterns by taking into account the information of neighboring pixels and, since the same weights are applied to different parts of the input, they can detect patterns in a spatially invariant fashion. Furthermore, this weight sharing reduces the number of parameters that require training compared to other layer types. In image analysis, simple kernels can e.g. detect edges and sequentially applying multiple convolutional layers allows them to learn more complex structures or objects. Applied to genomic sequence analysis, simple patterns can resemble position weight matrices of TF binding motifs whereas complex patterns captured deeper in the network can represent regulatory syntax grammar.

Multiple different filters can be applied on the same input so that the number of output channels depends on the number of convolutional filters used in the step. By using several filters at each convolution step together with multiple hidden layers, the model can learn complex patterns and is well-suited for tasks such as image segmentation or genomic analyses.

2.4. Multiple testing correction

In genomic analyses, such as differential expression analysis in RNA-seq experiments, thousands of comparisons are tested for statistical significance. Using traditional cutoffs, this would lead to a very high number of false positives: for a single test the probability of obtaining one false positive result (also called Type I error or family-wise error rate (FWER)) is α, but for N tests it increases to $1-(1-\alpha)^N$. This increase in type I error means an increase in the probability of wrongfully rejecting the null hypothesis H_0, resulting in a false positive finding. There exist multiple approaches to correct for this multiple testing burden, either by adjusting the significance threshold α or by adjusting the p-values.

2.4.1. Bonferroni correction

If we have a family of N hypotheses and their corresponding p-values p_i with $i \in 1, ..., n$, the FWER is the probability to wrongfully reject the null hypothesis for at least one test within this family even though it is true. The Bonferroni correction ensures that the FWER is less than or equal α through adjusting α' by the number of tests n [145].

$$\alpha = \frac{\alpha'}{n} \tag{2.33}$$

Instead of adjusting α it is also possible to adjust the original p-value p_i of each test i in the family to

$$p'_i = max(p_i * n, 1) \tag{2.34}$$

Depending on the use case, the Bonferroni correction can be too conservative and leads to an increased type II error, i.e. false negatives.

2.4.2. Benjamini-Hochberg correction

An approach that is less conservative is the Benjamini-Hochberg correction. It is a method that controls the false discovery rate (FDR), which is the proportion of wrongly rejected H_0 among all rejected null hypotheses [146]. The procedure ranks the original p-values in ascending order $0 \leq p_1 \leq ... \leq p_N \leq 1$ and determines the largest index k where

$$p_k \leq \frac{k}{N} * \alpha \tag{2.35}$$

All p-values in the range $p_1, .., p_k$ are considered significant. Alternatively, the original p-values p_i can be adjusted to q_i using the ranked list of p-values starting from the largest p-value i=k in decreasing order.

$$q_i = min(p_i \frac{N}{k}, q_{i-1}) \tag{2.36}$$

2.5. Dimensionality reduction

In high dimensional data, such as genome wide sequencing data, dimensionality reduction is a crucial tool to facilitate visualization, denoising and improve computational efficiency. There are many different approaches, with methods such as Principal Component Analysis (PCA) [147] and MDS [148] focusing more on preserving the pairwise distance between all the samples, and others such as t-SNE [149] and UMAP [124] focusing more on preserving the local structure.

2.5.1. Principal Component Analysis

PCA is based on the idea that when you observe d variables for each sample and some of the variables are correlated, a k≤d number of uncorrelated principal components can be used to represent the data without loosing information on the relation between samples. The principal components are ordered, such that the first component captures the highest amount of variance in the data and the remaining components explain the highest amount of the residual variance in descending order.

PCA is especially popular to show the overall sample similarity in bulk sequencing experiments. In that case the samples are commonly visualized in PC space of the top components in 2D plots. In single-cell sequencing analyses it is frequently used to reduce the number of dimensions before computing a neighbourhood graph. Mathematically, the principal components can be determined by computing the eigenvectors of the covariance matrix of the data.

In detail, the steps include:

1. centering and scaling of the input data
2. computing the covariance matrix
3. finding eigenvalues and their corresponding eigenvectors

4. sorting eigenvectors based on decreasing eigenvalue
5. pick top k eigenvectors and multiply them with the input to project it into PC subspace

The eigenvalues measure how much variance of the data is captured by each component and can be used to decide how many components to retain based on elbow plots or the Kaiser criterion. A useful aspect of PCA is that it is straightforward to investigate which input features have a high contribution on what component.

2.5.2. Uniform Manifold Approximation and Projection

An alternative approach for dimension reduction is Uniform Manifold Approximation and Projection (UMAP). The memory and runtime scalability of the algorithm make it suitable for processing large datasets. In fact, UMAP is the methods of choice for visualizing single-cell data in lower dimensional space and is used by scanpy [113] and Seurat [150]. Additionally, it is arguably better than t-SNE at preserving the global structure and a benchmark study demonstrated that its results are also the most reproducible [151].

The algorithm is based on ideas from topological data analysis and manifold learning techniques. In general terms, it starts by constructing a k-neighbour graph and converting it to a topological structure. From there it finds a low dimensional representation of the data that is as close to that topological structure as possible [124].

3. GR-mediated gene expression in macrophages

Immunosuppressive drugs such as the synthetic glucocorticoid Dexamethasone (Dex), are commonly used to treat inflammatory conditions [11, 12, 13]. As described in subsection 1.1.1, Dexamethasone exerts its function by signalling through GR, however it is not yet understood how it can simultaneously activate some target genes and repress others.

In this work, we aim at gaining a deeper understanding by integrating genome-wide GR binding data with information on chromatin accessibility, H3K27ac marks, and 3D conformation data (HiC) during the feature engineering process. The goal is to utilize these feature in order to predict gene expression changes. Notably, we employ 4sU-seq to specifically examine nascent transcripts generated after Dex treatment, thus avoiding the confounding effects of preexisting transcripts. These genome-wide assays provide an unprecedented level of detail regarding various aspects of transcription, however, the best approach to integrate this wealth of information in order to derive mechanistic insights about Dex-induced gene expression changes, remains unclear.

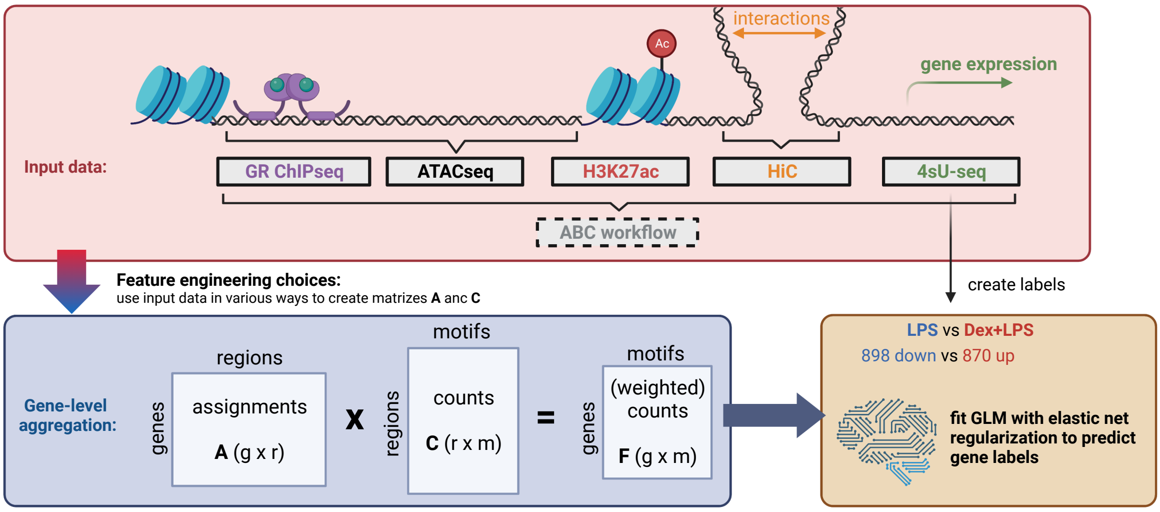

Figure 3.1.: Graphical abstract of the workflow deployed in the manuscript. Graphical abstract taken from [2].

We chose to tackle this challenge by approaching the biological question as a machine learning problem. Our aim is to predict the transcriptional outcome, defined as binary label for significantly up- and downregulated genes in response to Dex treatment, based on

DNA sequence patterns. To achieve this, we have devised a biologically-informed workflow that integrates genomic sequence and epigenetic assays into a unified model. We combine GR ChIP-seq, ATAC-seq and H3K27ac data in various ways to generate predictive features. Subsequently, we evaluated the performance of these different feature engineering strategies using an independent test set. Ultimately, we dive into interpreting the best performing models and conduct follow-up experiments to gain new insights into the regulatory grammar governing transcriptional changes (Figure 3.1).

Our findings reveal that models primarily based on information from GR binding locations outperform most others, demonstrating that the majority of information necessary for predicting Dex-induced transcriptional changes lies within these binding sites. Interestingly, the inclusion of 3D conformation data and region activity information, achieved by incorporating ABC score in the feature engineering process, does not improve the performance. Our analyses provide confirmation of the involvement of NF-κB binding sequences in gene repression and identify STATs as potential novel factors containing cryptic GR binding sites.

3.1. Results

3.1.1. Defining GR target genes

To understand GR-mediated transcriptional regulation, our first step is to define a set of genes whose expression is significantly altered in response to GR signaling, referred to as target genes. Leveraging previously published data [152] on 4-thiouridine labeled nascent transcripts in murine BMDMs, we conducted a comparative analysis of nascent transcript levels between samples treated with the pro-inflammatory agent LPS for 2 hours and those subjected to a combined treatment of the GR agonist Dexamethasone for 2.5 hours followed by a 2-hour LPS treatment (Dex+LPS). Binding of its ligand Dex allows GR to dimerize and translocate to the nucleus where it regulated gene expression. 4sU-seq allows us to track the levels of nascent transcripts and we can determine which genes are significantly up- or downregulated which serve as binary labels for subsequent computational analyses.

Labelling nascent transcripts for 1 hour before lysis, we uncovered notable changes in gene expression. Specifically, we observed a significant upregulation of 870 genes (adjusted p-value < 0.05, log2FC > 0.58) and a downregulation of 898 genes when comparing the Dex+LPS condition to the LPS-only condition (Supplemental Figure A.1). The number of DE genes comparing Dex+LPS and LPS treated samples are consistent with similar trends reported in microarray studies utilizing total RNA [27], albeit demonstrating more pronounced gene expression changes compared to shorter treatment times [153].

3.1.2. ChIP-seq summits provide high resolution binding locations

Characterizing GR-mediated transcriptional regulation requires an investigation of the genomic regions bound by GR to decipher the underlying mechanisms. It is reasonable to assume that these GR-bound regions harbor crucial information that determines the transcriptional response to Dex treatment, making them ideal candidates for predictive modeling.

To accomplish this, we leveraged previously published ChIP-seq data from macrophages treated with Dex and LPS for 3 hours [154]. The Dex+LPS treatment activates GR in an inflammatory context and we identified the genomic regions GR interacts with either through direct DNA binding or indirectly by interacting with other DNA-binding factors. We define the GR peak universe as reproducible peaks among replicates, resulting in a total of 13,431 peaks. This approach ensures our focus on robust peaks and aligns with the number of GR peaks observed in Dex+LPS-stimulated macrophages in earlier reports [27].

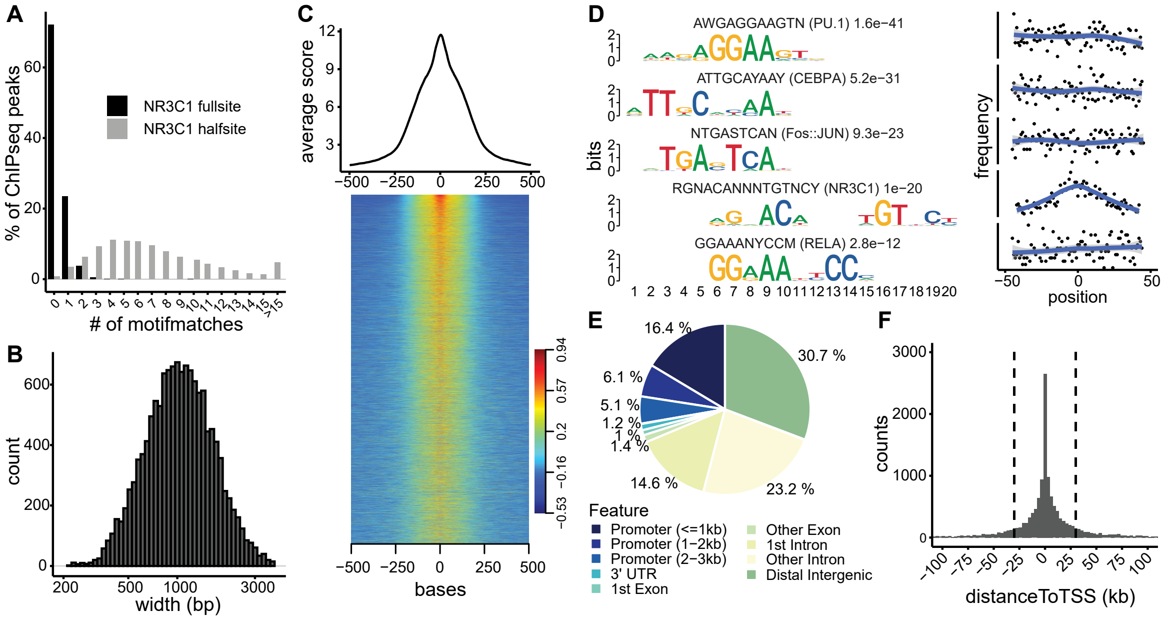

Figure 3.2.: **GR binding regions.** (A) Percentage of GR ChIP-seq peaks containing a certain amount of NR3C1 half site or full site matches. (B) Width distribution of GR ChIP-seq peaks called by MACS2. (C) Read distribution around GR ChIP-seq peak summits. Aggregated (top) and per location (bottom). Colors represent values scaled by row with (x - mean(x))/(max(x)-min(x)+1). (D) *De novo* motif enrichment in a 100 bp region centered on the GR ChIP-seq peak summit compared to shuffled control, with evalue and manual annotation to known motifs above the sequence (left). Positional distribution of found motif in relation to the GR ChIP-seq peak summit (right). (E) Genomic location of GR ChIP-seq peaks. (F) Distribution of distance from GR ChIPs-eq peaks to closest transcription start site (TSS). Figure and legend taken from [2].

Unexpectedly, only a fraction (27.82%, 3,737 peaks) of these ChIP-seq peaks contained at least one NR3C1 full site (Figure 3.2A). On the other hand, nearly all peaks (99.20%, 13,324 peaks) contained one or more NR3C1 half sites. However, given the relatively short length of the NR3C1 half site, its occurrence by chance without biological significance is statistically more likely. Our analysis revealed a total of 424,144 genome-wide matches for the full site and 14,427,667 for the half site.

Although the average peak width was of substantial size (Figure 3.2B) with a mean of 1,178.96 bp and a median of 1,034 bp, we observed a distinct pileup of reads around the peak summit (Figure 3.2C). Therefore, we selected a 100 bp region centered on the peak summits (GR summit regions) for *de novo* motif analysis using STREME of the MEME Suite [155]. We compared the discovered motifs with previously described ones and found enrichment (Figure 3.2D) of the macrophage lineage determining factor PU.1 (encoded by the gene *Spi1*) ($e\text{-}value = 1.6e^{-41}$), CEBPA ($e\text{-}value = 5.2e^{-31}$), FOS::JUN motifs ($e\text{-}value = 9.3e^{-23}$), and the classical GRE (deposited in motif databases as NR3C1 motif) ($e\text{-}value = 1e^{-20}$). The enrichment of the AP-1 complex (composed of FOS and JUN) and CEBPs is not surprising since they are known pioneering factors involved in GR recruitment [156, 157]. While the occurrence of motifs related to these pioneering factors appeared evenly distributed over the input window, the positional distribution of the NR3C1 motif within the GR summit regions exhibited central enrichment.

In terms of genomic location, 27.70% of the peaks were found in promoter regions within 3 kb of the transcription start site (TSS), 37.86% in introns, and 30.75% in distal intergenic regions (Figure 3.2E). These locations of GR peaks align with previous findings in mouse liver under activated GR conditions [158].

It is noteworthy that a majority of the identified peaks lack a GRE, suggesting potential scenarios where either certain GREs do not adhere to the consensus motif or alternative mechanisms facilitate the interaction between GR and DNA, either directly or indirectly. Nonetheless, the positional distribution of the NR3C1 motif within the GR summit regions confirms the informative value of summit information in accurately determining the binding location, making it invaluable for subsequent analyses. Consequently, the GR summit regions provide us with high-resolution binding information, serving as an instrumental starting point for predicting changes in gene expression. By assigning these summit regions to their putative target genes, we can explore differences in the GR summit regions that may account for the observed transcriptional outcome.

3.1.3. Activated and repressed GR target genes exhibit genetic and epigenetic differences in their proximal GR summit regions

While GR peaks located in close proximity to a TSS are likely involved in the regulation of the corresponding gene, accurately assigning peaks located far from a TSS is challenging. To mitigate the risk of false annotation during this proximity-based assignment, we applied a filter and considered only peaks within a 30 kb range of a TSS (Figure 3.2F). This filtering step resulted in 10,117 (75.33%) of all peaks being included and annotated them to a total of 5,412 unique genes. Notably, for genes downregulated in response to Dex treatment, the nearest ChIP-seq peak was found to be farther away compared to upregulated genes (Figure 3.3A). To validate the statistical significance of this difference, we conducted a permutation test, which confirmed that the observed gap of 9,909 bp exceeded what would be expected by chance ($p\text{-}value \sim 0$) (Figure 3.3B). These findings align with previous studies on GR binding in humans [18] and suggest the presence of distinct mechanisms underlying GR-mediated gene activation and repression, with the repression mechanism being more difficult to understand.

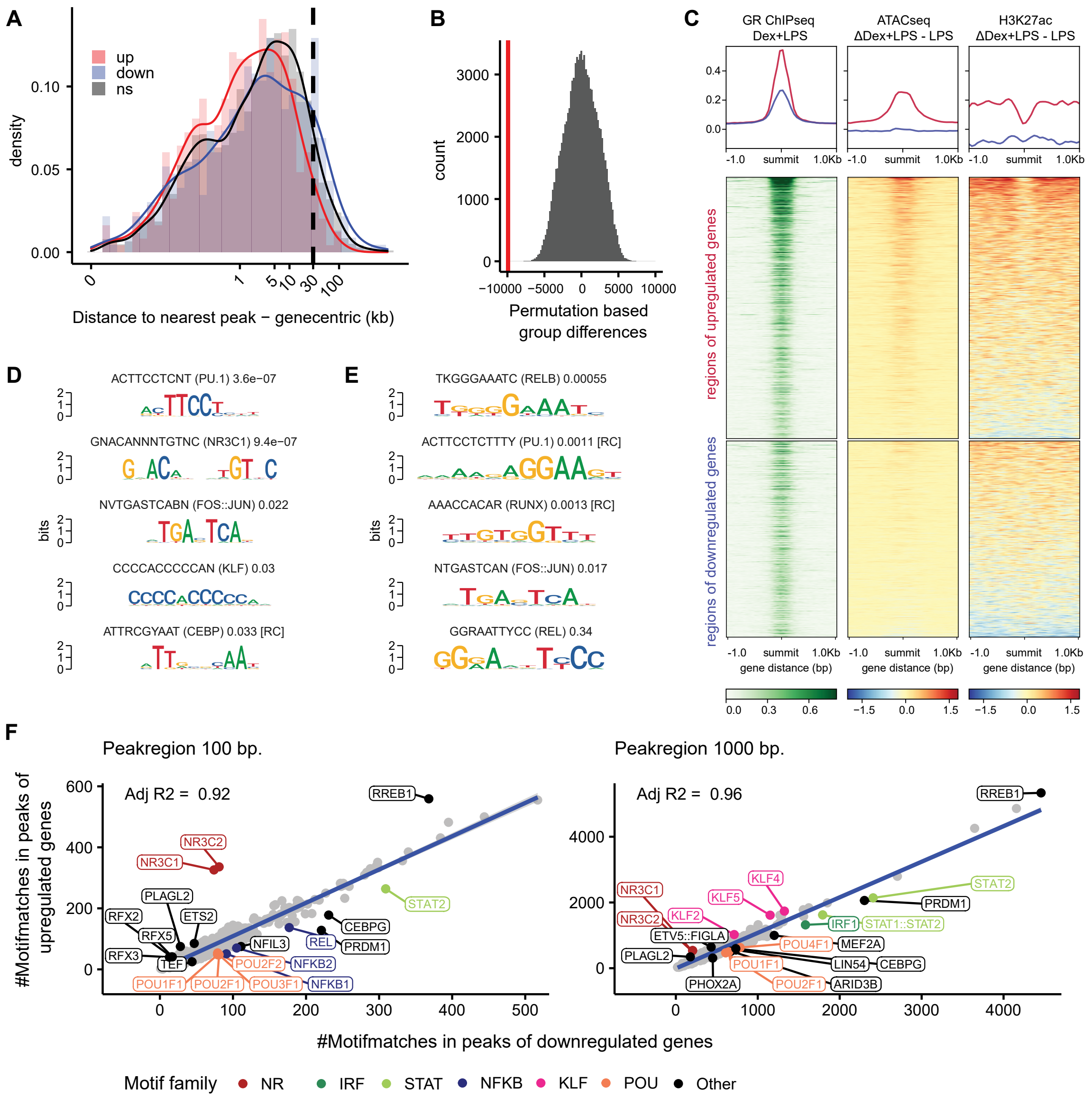

Figure 3.3.: **Differences in GR summitregions of activated and repressed GR targets.** (A) Distance from TSS of genes to nearest GR ChIP-seq peak split by direction of significant 4sU-seq expression change of the gene between Dex+LPS and LPS condition. The change is considered significant if it passes a threshold of adjusted p-value <0.05 and ∣log2FC∣ > 0.58. ns = not significant (B) Observed group difference of TSS to peak distance between up- and downregulated genes compared to differences expected with 100,000 permutations of randomized group labels. (C) GR ChIP-seq signal and Dex+LPS - LPS difference of normalized ATAC-seq and H3K27ac signal at GR ChIP-seq peaks annotated to up- and downregulated genes. (D+E) *De novo* motif analysis of 100 bp sequences centered around summits of peaks annotated to (D) upregulated genes and (E) downregulated genes (F) Chi-square results of motif occurrences in defined areas around the summit of peaks annotated to up- versus downregulated genes in an area of 100 bp (left) or 1000 bp (right) around the peak summits. Top 20 motifs with smallest adjusted p-value are labeled. Figure and legend taken from [2].

We examined whether epigenetic changes, such as variations in the accessibility or acetylation of lysine residue at position 27 of the H3 histone (H3K27ac), a marker for active enhancers [159], could account for the observed differences in gene expression. To investigate this, we utilized previously published ATAC-seq data from BMDMs treated with LPS or Dex+LPS for 3h, as well as H3K27ac histone data from BMDMs treated with LPS (3h) or Dex(16h)+LPS(3h). We calculated the difference in normalized scores between Dex+LPS and LPS-stimulated samples, where positive values indicate higher signals in the Dex+LPS condition. Our findings revealed that gene activation correlates with increased accessibility and a slight elevation in H3K27ac signal. However, the repression of genes cannot be attributed to a loss of accessibility (Figure 3.3C), which is is line with previous reports of sustained accessibility for repressed targets [154] and enhanced accessibility at Dex-induced genes [160].

We examined whether subtle sequence variations within the NR3C1 motif, present in the regulatory regions of genes that are upregulated or downregulated in response to Dex treatment, could potentially explain the observed changes in gene expression. To investigate this, we performed a *de novo* motif analysis separately for the GR summit regions associated with activating and repressing genes (hereafter referred to as activating and repressing GR summit regions, respectively). Surprisingly, we found a significant enrichment of the NR3C1 motif in the activating GR summit regions (Figure 3.3D), but not in the repressing regions (Figure 3.3E) ($p\text{-}value = 9.4e^{-7}$).

In order to systematically explore sequence differences while leveraging existing knowledge of motifs, we quantified the occurrences of motifs from the JASPAR database [91] within the activating and repressing GR summit regions. Subsequently, we performed a chi-square test to determine whether the distribution of motif matches was uneven between the two sets. Notably, within the original window size of 100 bp, the motif counts in the two peak sets showed an adjusted R2 of 0.92 (Figure 3.3F, left). When using an extended window size of 1,000 bp, the adjusted R2 increased to 0.96 (Figure 3.3F, right). This indicates that using larger input sequences leads to the dilution of differences, highlighting the importance of using peak summit information combined with a narrow window size to enhance the resolution of ChIP-seq data.

Analyzing the GR summit regions, we observed the most significant differences in the occurrence of the NR3C1 and NR3C2 motifs, with adjusted p-values of $6.44e^{-30}$ and $1.48e^{-29}$, respectively. In contrast to the *de novo* motif analysis, this approach allowed us to directly assess the occurrence of differential motifs and it successfully identified known cofactors that remained significant or marginally significant after correcting for multiple testing. Specifically, for the motifs associated with gene repression, we identified cofactors of the NF-κB (NFKB1: $p\text{-}adj = 1e^{-2}$), NFKB2: $p\text{-}adj = 5.9e^{-2}$), and REL ($p\text{-}adj = 9.71e^{-2}$), C/EBP (CEBPG: $p\text{-}adj = 3.93e^{-2}$), STAT (STAT2: $p\text{-}adj = 1.32e^{-1}$), and OCT (POU2F1: $p\text{-}adj = 4.03e^{-2}$, POU1F1: $p\text{-}adj = 5.9e^{-2}$, POU3F1: $p\text{-}adj = 6.96e^{-2}$, and POU2F2: $p\text{-}adj = 1.95e^{-1}$) families of transcription factors (Figure 3.3F, left).

By examining the total motif counts within the activating and repressing GR summit regions, we observed a predictive signal that can be further explored in subsequent analyses. This proximity-based assignment of regions to genes, along with the established labels, can also be employed in computational approaches that operate on a per-gene level. For our computational approach, we will utilize motif counts at GR summit regions with a window size of 100 bp, combined with proximity-based assignments, as the reference model.

3.1.4. ABC scores capture condition-specific differences in chromatin activity

The proximity-based assignment of regions to genes provides a simplified approximation that may overlook the intricate gene regulatory mechanisms by disregarding the three-dimensional architecture of chromatin. However, a promising alternative known as the activity-by-contact (ABC) model [161] offers a valuable approach to exploit chromatin structure data and epigenetic information in order to achieve more accurate region-gene assignments. The ABC workflow provides us with the opportunity to integrate publicly available HiC data [162] of BMDMs, along with ATAC-seq, H3K27ac [163], and 4sU-seq gene expression data [152] and returns multiple valuable outputs that we will use downstream. Firstly, it identifies potential regulatory regions (active regions) within the genome. Secondly, it generates ABC scores, which serve as indicators of regulatory potential. Additionally, it provides information on the genomic location of these regions relative to their associated target genes (e.g. whether the region is within the gene's promoter). Notably, since the data utilized is specific to macrophages and all samples, except for the HiC data, pertain to specific conditions (LPS and Dex+LPS), the resulting ABC scores are condition-specific for these contexts as well.

The ABC workflow determined an average of 2.37 enhancers per gene in the Dex+LPS condition and 2.36 in the LPS condition (Figure 3.4A, top) surpassing the predefined threshold of 0.02, which aligns with the expected range specified by the ABC authors. Considering that a specific genomic region can regulate multiple genes, the workflow allows ABC regions to be assigned scores for multiple target genes. Specifically, in the Dex+LPS condition, enhancers hold scores for an average of 1.45 genes, and in the LPS condition, they have scores for an average of 1.44 genes (Figure 3.4A, bottom).

The ABC scores are determined in a condition-specific manner. As a result there are cases where the workflow identifies a regulatory region for one treatment, but not the other. This can happen if the region does not even meet the cutoff for accessibility or if the activity (as determined by a combination of accessibility, H3K27ac marks and HiC contacts) does not pass the threshold. In order to compare scores between conditions, we displayed scores that are not present in the second condition as 0. We find that overall, the ABC scores between the two conditions are highly correlated with $r = 0.92$ (Figure 3.4B).

Interestingly, when we selectively visualize the ABC scores associated with genes that are either upregulated or downregulated in response to Dex treatment, we observe differences in the marginal distributions (Figure 3.4C). Specifically, genes activated by GR are more likely to have ABC scores for the Dex+LPS condition, while having no scores exceeding the threshold (indicated as 0) for the LPS-specific scores. By calculating the difference in ABC scores between the two conditions and plotting it against the log2 fold change in gene expression for genes exhibiting changes in response to Dex treatment, we find a moderate correlation ($r = 0.25$) (Figure 3.4D). This indicates that variations in the ABC score can account for a portion of the stimulus-induced gene expression change, particularly for extreme cases near

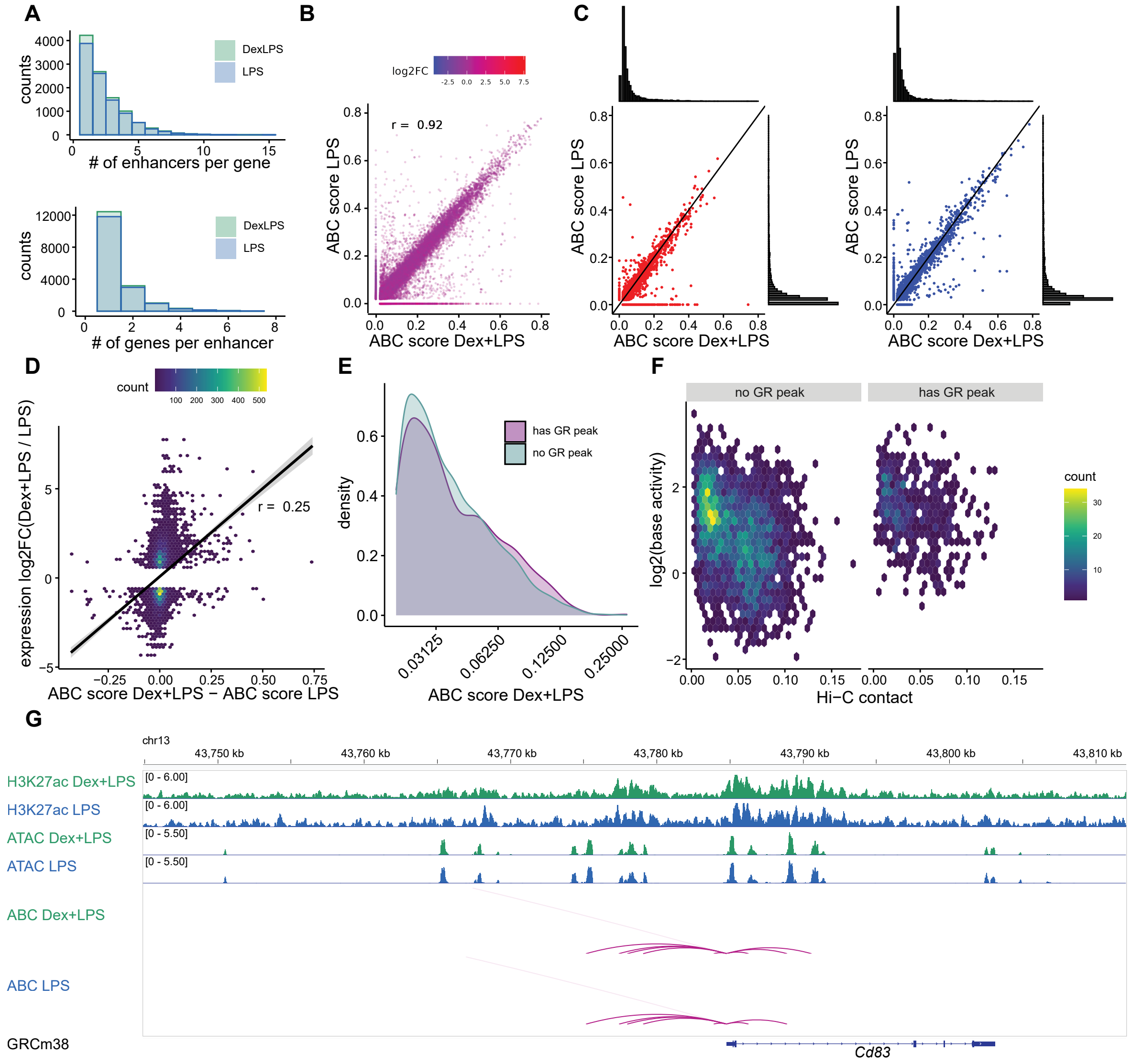

Figure 3.4.: **Condition specific region-gene assignments using the activity-by-contact workflow.** (A) Number of enhancers per gene and genes per enhancer resulting from the activity-by-contact (ABC) workflow. (B) Correlation of ABC scores between Dex+LPS and LPS conditions. (C) ABC scores for regions annotated to (left) upregulated genes and (right) downregulated genes. Scores that only passed the threshold of 0.02 in one condition, are displayed as 0 in the second condition. (D) Difference in ABC scores between Dex+LPS and LPS condition versus log2fold expression change from 4sU-seq (E) ABC scores of the Dex+LPS condition for differentially expressed genes, split by whether the ABC enhancer region overlaps with a GR peak or not. (F) Base activity versus powerlaw scaled and adjusted HiC contacts split by whether the ABC enhancer region overlaps with a GR ChIP-seq peak or not. (G) IGV snapshot of normalized H3K27ac and ATAC-seq signals as well as the condition specific ABC scores for the CD83 locus. Figure and legend taken from [2].

the axes (Figure 3.4C).

Active regions are identified through the ABC workflow, and it is worth noting that not all of these regions contain GR peaks. Interestingly, the active regions containing GR peaks tend to have higher ABC scores ($mean = 0.043$) compared to those without GR peaks ($mean = 0.041$) (Figure 3.4E). By considering that the score combines base activity with the number of HiC contacts, we can infer that regions with GR peaks exhibit higher scores due to elevated base activity (Figure 3.4F).

In the case of differentially expressed genes like *Cd83*, the condition-specific ABC scores (Figure 3.4G) identify potential regulatory regions responsible for gene expression changes. The ABC workflow integrates multiple epigenetic assays to make condition-specific predictions of regulatory regions, not only for individual genes but on a genome-wide scale. At present it is the gold-standard for enhancer-promoter assignments that other methods use as benchmark [134]. However, the current challenge lies in determining how to combine ABC results with additional genomic assays to predict gene expression responses. In our approach, we will incorporate this information into our machine learning framework to uncover patterns associated with Dex-induced gene repression for all differentially expressed genes.

3.1.5. Feature engineering

We want to predict gene-expression changes in a genome-wide fashion with a model using tabular data as input, which makes the integration of all available information into a unified feature matrix indispensable. This process involves making critical decisions at multiple stages of feature engineering, including the identification of regions of interest, region-gene assignments, and data aggregation for gene-level predictions. The ABC model provides additional options at each of these levels.

The first decision is how to choose the input regions we consider when quantifying motif occurrences. While it is reasonable to assume that much of the information governing the transcriptional response to Dex treatment is found within the GR summit regions, focusing solely on GR-bound regions might overlook other significant factors present in accessible but non-GR-bound regions. Alternatively, active regions identified by the ABC workflow in either the Dex+LPS or the LPS condition can be utilized as input for quantifying motif occurrences. Within the active regions, we can further discriminate between promoters, located within 500 bp of a TSS, and enhancers, situated in genic or intergenic regions. Analyzing the features extracted from these regulatory regions separately allows for flexibility in excluding some or all promoter regions, on the other hand aggregating features from promoters and enhancers into a single value increases interpretability. By feeding the motif counts from promoters and enhancers to the model as individual features, we make the assumption that those genomic regions capture different aspects of the sequences driving transcriptional outcome. Conversely, considering that the majority of GR binding locations occur outside of promoter regions, it can be argued that enhancer regions are predominantly responsible for GR-mediated gene regulation, and thus, promoter regions could be excluded from the model. In a hybrid approach, we exclude motif counts from promoter regions that are not the promoter of the target gene in question but instead promoters of surrounding genes (nonself promoters).

On top of choosing the input regions we also need to decide on how to assign these regions to their corresponding target genes. The simplest way is to assign each region to its nearest gene by linear proximity, resulting in a one-to-one relationship. However, this approach oversimplifies the biological reality because it ignores how genetic regions can interact in three-dimensional space, even if they are far apart in sequence. Additionally, one region can influence the activity of multiple genes. Using the ABC model allows us to have a more flexible one-to-many mapping, considering situations where a single region is involved in the regulation of several target genes.

Another factor to consider is that multiple regions can work together to regulate one specific gene. At the aggregation step, we must decide whether each region's contribution to the gene-level prediction should be equal, or if we should weigh them based on their activity. This opens the door to incorporating further epigenetic information about the regions by using ABC scores as weights when we add up the motif occurrences. Lastly, instead of using a feature matrix based on one specific treatment condition, we can explore the difference between conditions. We do this after the aggregation step by subtracting the gene-by-motif matrices of the Dex+LPS and LPS conditions (details see subsubsection 3.2.7).

3.1.6. GLMs identify motifs predicting GR mediated gene repression

Rather than making arbitrary choices in our feature engineering, we adopt a systematic approach by exploring all possible combinations to identify the best one before delving into the biological interpretation. It is a challenge that the number of features in some of our matrices greatly exceed the number of data points, increasing the risk of overfitting complex machine learning models. To mitigate this risk, we opt for simple logistic regression models and apply regularization techniques to further reduce overfitting. Furthermore, linear models offer a more straightforward interpretation of model parameters compared to more complex models like tree-based ones. We recognize that motif counts of TFs from the same motif family are correlated (Supplemental Figure A.2). To address this, we employ elastic net regularization during feature selection in our GLMs. Elastic net regularization, unlike lasso regularization, considers correlated features together [139], which is better suited for our specific use-case. Moreover, the elastic net penalty helps exclude non-informative features from the final model by assigning zero-coefficients, thereby improving biological interpretability.

As a reference model, we counted the occurrences of motifs in GR summit regions, assigned regions to genes based on proximity, and then combined motif counts by simply summing those from all summit regions mapped to the same gene without any weighting. Alternatively, we quantified motif counts in active regions identified by the ABC workflow, assigned these regions to genes using ABC-based assignments, and iterated the other modeling options such as weighting and inclusion of promoters. In a hybrid approach, we combined motif counts from GR summit regions but incorporated ABC information from the overlapping ABC regions. This means we focused on summit regions located within regulatory regions identified by the ABC workflow (Supplemental Figure A.3).

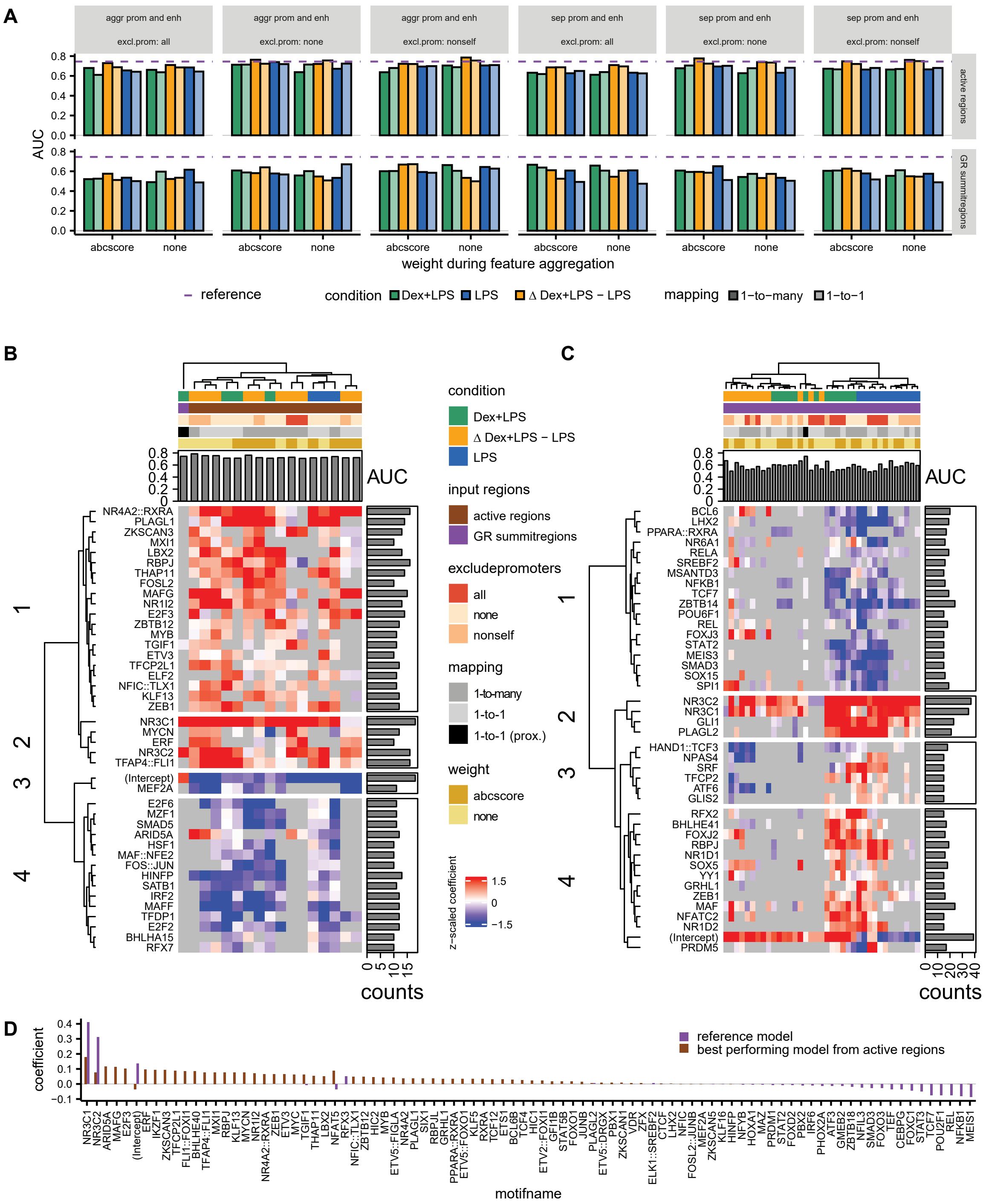

Figure 3.5.: See legend on next page.

Figure 3.5 *(previous page)*: **Systematic model comparison.** (A) Model performance on the test set for models generated iterating through combinations of feature engineering choices. (B+C) Model coefficients displayed as heatmaps. Coefficients were scaled within each model for plotting purposes. Euclidean distance was used as input for row and column clustering with the ward.D2 method. Row barplot indicates number of models in which the feature received a non-zero coefficient. 1-to-1 (prox) = proximity-based 1-to-1 mapping used in the reference model. (B) Top 17 best performing models in which features in promoter and enhancer regions are aggregated and their top 40 most frequently selected coefficients. (C) Coefficients of models based on GR summitregions, filtered for factors with non-zero coefficients in at least 15 models. (D) All coefficients of the reference model and the best performing one based on all active regions. aggr prom and enh = motif counts from promoter and enhancer regions get aggregated and fed into the model jointly. sep prom and enh = genomic location (promoter, genic, intergenic) of the regions of interest is considered in the feature engineering process (leading to up to 3 features per motif). excl prom = whether all, none or only promoters that are not assigned to the respective target gene (nonself promoters) should be excluded in the feature engineering process. Figure and legend taken from [2].

By comparing models based on motif counts in ABC regions to those based on motif counts in GR summit regions, we can disentangle different mechanisms at play: one group that results from changes in accessibility and region activity (well performing models based on the difference between Dex+LPS and LPS indicated in yellow), and another group that arises from direct binding with GR. The overall best performing model achieved an $AUC = 0.79$ on the test set, with $AUC = 0.83$ on the training set, suggesting no substantial overfitting (Supplemental Figure A.4A). This model utilized the difference in motif counts between Dex+LPS and LPS conditions at ABC regions (Figure 3.5A). It excluded promoter regions unless they were specifically the promoter of the target gene and omitted ABC scores in the aggregation step. To assess whether this performance significantly outperformed other models, we compared ROC curves using the Delong method (Supplemental Figure A.4B) and estimated the proportion π_0 of true null hypotheses using Storey's q-value method from the resulting p-value distribution. The analysis revealed that only a fraction of approximately $\pi_0 = 0.0054$ of all 144 tests (~ 0.8) showed ROC curves that did not differ significantly from the best model and thus performed equally well.

Several motifs were shared between the coefficients selected by the model based on GR summitregions and those selected by the best performing models based on active regions (Figure 3.5B). Both approaches highlighted the significance of the classic GRE motif, as evident from the strong positive coefficients assigned to NR3C1 and the highly similar motif NR3C2 in most models. Models relying on active regions suggested a connection between GR-mediated gene regulation and SMADs. Notably, the models based on active regions showed both positive (FOSL2) and negative (FOS::JUN) coefficients for AP-1 family members.

The model also detected factors that can impact the transcriptional outcome by influencing the epigenetic landscape. For instance, ARID5A is part of the SWI/SNF chromatin remodeling complex, and HINFP plays a role in a histone deacetylase (HDAC) complex. As our main objective was to comprehend transcriptional regulation through direct interactions with GR, rather than general activation patterns, we focused on models derived from features of the GR summit regions (Figure 3.5C). Notably, among the factors most frequently included in the final model were several members of the NF-κB family (REL, RELA, NFKB1) and STAT2. This finding corroborates our earlier results on total motif counts from the peak-based analysis, as presented in Figure 3.3F.

Our reference model, which used proximity-based assignments, achieved an $AUC = 0.74$ on the test set (Figure 3.5A, dashed purple line) and an $AUC = 0.77$ on the training set, indicating no substantial overfitting (Supplemental Figure A.4A). To compare this performance with all other models, we conducted pairwise comparisons and used Storey's q-value method on the resulting p-value distribution. The analysis returned a $\pi_0 = 0.026$, meaning that only a small fraction (approximately 3.7) out of all 144 tests were estimated to perform as well or better than the reference model. This demonstrates that our reference model outperformed the majority of other models (Supplemental Figure A.4C), providing evidence that most of the necessary information to predict Dex-induced transcriptional changes is indeed contained within GR-binding locations.

The reference model showed better performance compared to the hybrid models (all $q < 0.05$), where motif counts from GR summit regions were combined with ABC scores derived from epigenetic data. As a result, we decided to investigate the selected coefficients for both the reference model and the best model based on active regions. In the multivariate model based on GR summit regions, the features with the most negative coefficients were MEIS1 (-0.087), NFKB1 (-0.080), REL (-0.076), POU2F1 (-0.076), TCF7 (-0.074), and STAT3 (-0.049). Conversely, the features with the largest positive coefficients were NR3C1 (0.410) and NR3C2 (0.313). However, it is important to note that in a multivariate model, the direction and magnitude of a predictor depend on the values of all other variables in the model. So, a positive coefficient does not necessarily mean a positive correlation with the target variable, and the coefficient of a predictor can change depending on the other predictors included in the model. Therefore, to confirm the direction, significance, and magnitude of the coefficients for the identified putative repressive motifs, we also conducted a bivariate analysis. In order of magnitude, the analysis returned NFKB1 (-0.260, $p = 0.0007$), POU2F1 (-0.258, $p = 0.0014$), MEIS1 (-0.227, $p = 0.0043$), REL (-0.222, $p = 0.0028$), TCF7 (-0.203, $p = 0.0074$), and STAT3 (-0.178, $p = 0.0166$) (Supplemental Figure A.4D). All motifs of interest maintained a negative coefficient (all $p < 0.05$). In short, the presence of these motifs within 100 bp around GR peak summits predicts gene repression.

It is essential to remember that these models represent a computational approach that merely identifies candidate sequences. Validating whether the identified factors are indeed expressed in the cells and determining whether the motifs appear in GR ChIP-seq data due to GR being tethered to their cognate factor or GR directly binding to those sequences can be explored with further analyses.

3.1.7. Protein interaction, expression and localization analysis of repressive factors

After identifying potential candidates for gene repression through our computational approach, we compared the results with experimental data to identify TFs likely to have a regulatory role *in vivo*. In our models, we noticed predictors from several members of the signal transducer and activator of transcription (STAT) family. For example, STAT3 showed up as a strong negative coefficient in the reference model, while STAT2 was selected by a large number of models based on GR summit regions (Figure 3.5C). Since all STATs recognize similar DNA sequence motifs (Supplemental Figure A.5), we decided to explore all family members further. To independently validate our computational results, we examined the expression levels of the TFs in macrophages. A TF can only play a regulatory role if it is expressed in these cells. To compare the expression levels of different transcripts, we considered the fragments per kilobase million (FPKM) to adjust for any biases caused by transcript length. Our analysis revealed that the expression of *Meis1* and *Tcf7* was low (<1 FPKM) across all conditions (Figure 3.6A), suggesting that these TFs are unlikely to be involved in GR-mediated gene regulation. Despite their low expression, the model assigned non-zero coefficients to their binding sequences, which might have been caused by sequence similarity between the motifs of MEIS1 and TCF7 and those of the true regulatory TF.

In response to LPS treatment, the expression increases for *Nfkb1* ($padj = 6.19e - 33$) and *Rel* ($padj = 1.75e - 20$), and decreases when adding Dex ($padj = 0.004$ and 0.012, for *Nfkb1* and *Rel* respectively). Among the *Stat* genes, all are expressed except for *Stat4*, but their expression levels do not show significant changes between the Dex+LPS and LPS conditions. *Nr3c1* itself is significantly downregulated ($padj = 0.0145$) in that comparison (Figure 3.6A), which likely represents a negative feedback loop following GR activation. In this context it is important to remember that the activity of many proteins depends on mechanisms beyond transcription, such as translocation and phosphorylation, which are not captured by expression levels.

GR is well-known for its direct protein-protein interactions with the inflammatory transcription factors AP-1 and NF-κB. Additionally, it has also been observed to have cryptic binding sites within their binding motifs. To explore whether GR has protein-protein interactions with other factors, we examined ChIP-MS data of BMDMs treated with Dex for 16 hours and LPS for 3 hours. Through this analysis, we confirmed interactions with NF-κB family members (NFKB1, REL, RELA) and AP-1 (JUNB). While we did not find significant interactions between GR and STAT1, STAT3, or STAT6, we did observe a significant interaction between GR and STAT5A:STAT5B (Figure 3.6B). The ChIP-MS assay could not identify POU2F1 and STAT2.

We explored whether there was additional evidence for tethering that could shed light on our observations. Phosphorylation is essential for STATs to dimerize and translocate to the nucleus, where they bind to DNA and trigger inflammatory gene expression [164, 165]. Examining the phosphorylation levels of STAT5 at different time points after treatment revealed a significant reduction in STAT5 activity in the Dex+LPS condition compared to the LPS condition (Figure 3.6C, left), while the total amount of STAT5 remained unchanged (Supplemental Figure A.6A). To further investigate this, we conducted STAT5 western blots on samples separated into nuclear and cytoplasmic fractions, which confirmed that the majority of STAT5 in the Dex+LPS condition is located in the cytoplasm and not active in

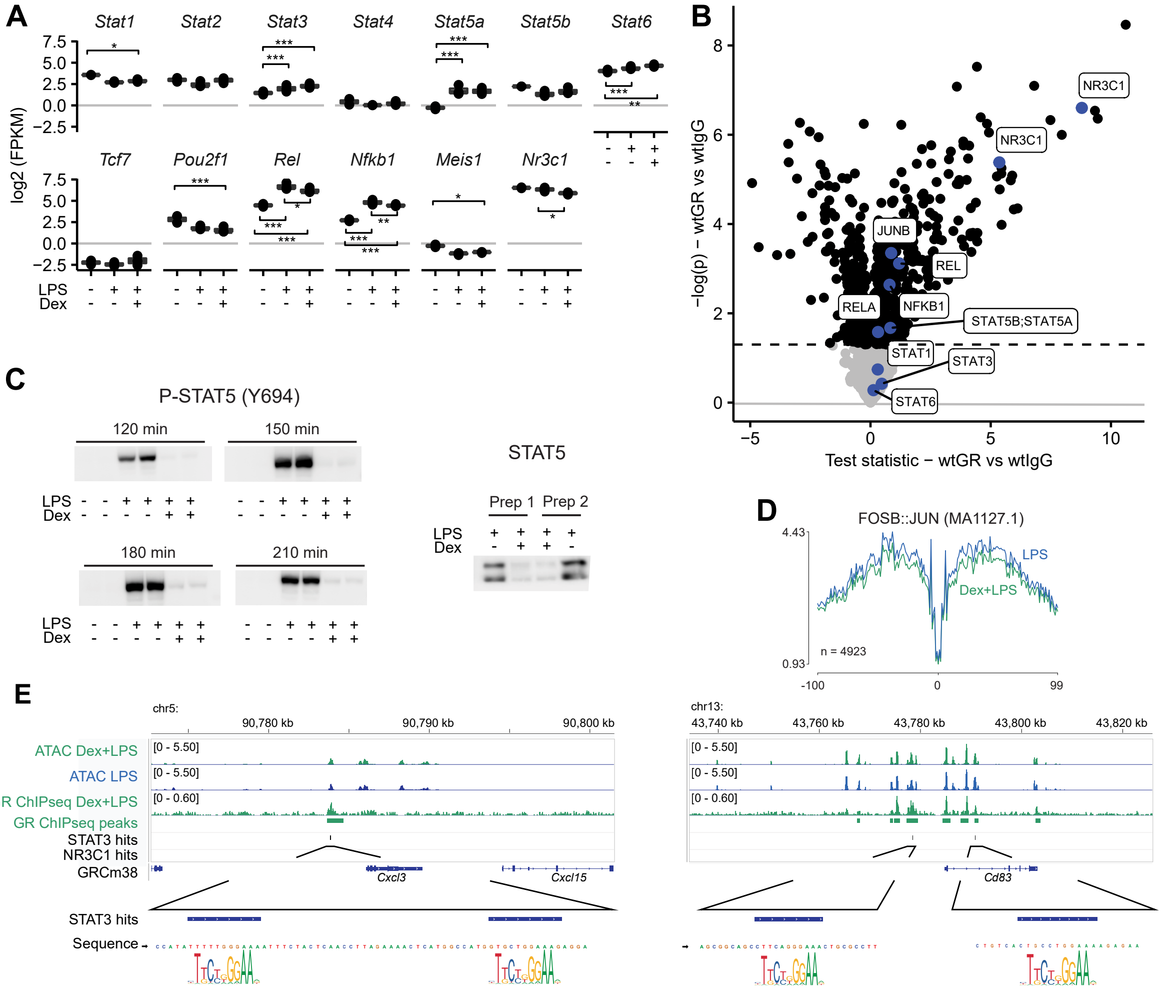

Figure 3.6.: **STAT expression, protein interactions and footprints.** (A) Expression levels of candidate factor for gene repression. (B) GR ChIP-MS data of samples treated with Dex+LPS. Factors of interest are labeled. (C) Western blot of STAT5. n=2. (Left) P-STAT5 of whole cell lysates at various timepoints. (Right) Western blot of STAT5 on nuclear extracts after 3h of Dex+LPS or LPS stimulation. (D) Lineplots for region with differential footprint around FOSB::JUN motif comparing a total of n=4923 regions. (F) Examples for downregulated target genes with STAT3 hits (but no NR3C1 hits) in peak regions. (left) Cxcl3 locus (right) Cd83 locus. STAT3 (MA0144.2) motif displayed underneath genomic sequence.
* = adj.p-value < 0.05, ** = adj.p-value < 0.01, *** = adj.p-value < 0.001 . Figure and legend taken from [2].

the nucleus (Figure 3.6C, right and Supplemental Figure A.6B). Similar reductions in activity were observed for STAT1 and STAT3 (Supplemental Figure A.6B). Based on this finding, it appears that the increased presence of STAT motifs within GR-bound regions is unlikely to be a result of tethering. Instead, the data is suggestive of an alternative mechanism, such as direct binding of GR to STAT motifs.

To further investigate potential TF binding at the motifs identified by the GLMs, we conducted footprinting analysis on the ATAC-seq data. This analysis aimed to investigate whether we could observe footprints formed by DNA-bound TFs that protect the DNA from transposase activity. Among others, the factors from the AP-1 and IRF families that we identified with the GLMs built on all regulatory regions, exhibited trends for differential footprints between the Dex+LPS and LPS conditions. Particularly noteworthy were the differential footprints associated with AP-1 family members (Supplemental Figure A.7), where regions around the footprints showed more accessibility in the LPS condition compared to the Dex+LPS condition (Figure 3.6D). However, it is worth mentioning that these effects did not retain statistical significance after correcting the p-values for multiple testing. This suggests that the predictive nature of these features might be attributed to changes in their accessibility between conditions, rather than through a direct GR-mediated mechanism.

For STATs, the absence of differential footprints suggests that changes in accessibility at those loci alone cannot explain their regulatory role (Supplemental Figure A.7). This becomes evident when examining regions with STAT3 motifs located in GR summit regions, where no visible changes in accessibility or region activity are observed (Supplemental Figure A.8). Two examples for genes downregulated in response to Dex+LPS treatment are *Cxcl3* and *Cd83*.They both harbor GR binding regions containing STAT motifs, but no NR3C1 motifs within 30 kb of their TSS (Figure 3.6E).

3.1.8. Conclusion

We set out to find sequence motifs that determine the direction of GR-mediated expression changes. Both, conventional motif enrichment tools and a tailored machine learning workflow that integrates various epigenetic assays reveal a connection between the NR3C1 motif and GR-mediated gene activation. Our approach suggests that gene suppression entails a variety of underlying mechanisms, including members of the AP-1, NF-κB and STAT families. The observation that STAT motifs are present in GR-bound regions despite a low abundance of nuclear STAT in the Dex+LPS condition suggests that this is not the result of GR tethering to STAT proteins but instead a direct interaction of GR with genomic STAT binding motifs.

3.2. Materials and additional methods

3.2.1. ATACseq

We merged fastq files from samples sequenced on multiple lanes and removed adapter sequences with trim-galore (v0.6.7). The reference `ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M24/GRCm38.primary_assembly.genome.fa.gz` was downloaded and filtered for standard chromosomes before creating a bowtie index from it. We mapped the reads with bowtie (v1.2.3) and its options *"-k 1 -m 1 -v 2"*. Unmapped mates were removed with Samtools (v1.9) and its flag *"-F 4"*, before filtering out duplicates with picard (v2.27.1). Finally, peaks were called with the macs2 (v2.2.7.1) command *"macs2 callpeak -f BAMPE –nomodel –keep-dup 1 -g mm -t bam "*

We retrieved a file with blacklisted regions from `http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/mm10.blacklist.bed.gz` and filtered out peaks overlapping those regions with bedtools. Fastqc (v0.11.9) and deepTools (v3.5.1) were applied for quality control, samtools was used to generate alignment stats and ultimately all quality control information was gathered into a report using MultiQC (v1.12).

We merged replicates from the same treatment groups into one bam file and one peaks file, which we then used as input for the footprinting analysis with HINT - Regulatory Analysis Toolbox (RGT) (v0.13.2).

3.2.2. ChIP-seq

Adapter sequences for the H3K27ac and GR ChIP-seq samples were removed with cutadapt (v4.0) before mapping the reads to the reference GRCm38 with bowtie (v1.2.3). GR ChIP-seq peaks were called compared against input controls using MACS2 (v2.2.7.1) and a relaxed threshold of $p = 0.1$, followed by determining reproducible peaks using idr (v2.0.4.2) and a threshold of $IDR = 0.05$.

The adjusted library size of H3K27ac and ATAC-seq samples was computed by merging peaks from both conditions for each assay and quantifying how many reads overlapped the respective peak universe. This adjusted library size was then used to scale the tracks of the samples. DeepTools (v3.5.1) was deployed to first compute the difference in read count tracks between the two conditions with *bigwigCompare* and then create visualizations with *computeMatrix* followed by the *plotHeatmap* function.

3.2.3. ChIP-MS

For details on the processing and the statistical analysis of the ChIP-MS data, please refer to the original publication [163].

3.2.4. Gene expression analysis (4sU-seq)

We trimmed the sequencing reads with trimmomatic (v0.39) and aligned them to a GRCm38 reference of rRNA with bowtie2 (v2.3.4.3), to remove ribosomal reads. From there we used the splice-aware aligner STAR (v2.7.0d) to align all remaining reads and removed duplicates with picard (v2.18.27) and then computed gene level counts with *featureCounts* from the subread tool (v1.6.3).

We removed genes with $CPM < 0.2$ in all samples and then performed differential expression analysis with with DESeq2 (v1.32.0). We mapped Ensembl gene IDs to MGI gene names using biomart. Genes were considered differentially expressed for downstream analyses if they had an adjusted p-value < 0.05 and absolute log2 fold change > 0.58 in the comparison of LPS to Dex+LPS treated samples.

3.2.5. ABC workflow

In order to integrate multiple assays into information that could be used during the feature engineering process, we turned to the ABC workflow. This workflow not only determines active regions which can serve as regions of interest, but also computes an ABC score linking regions to genes. Please refer to the github repository[1] for details on the ABC workflow.

TSS

We established macrophage specific transcription start sites through macrophage CAGE data retrieved from the FANTOM5 project. We determined tag clusters with CAGEr (v2.0.1) and lifted over coordinates of the dominant CTSS to mm10 assembly coordinated. From there we used the `TxDb.Mmusculus.UCSC.mm10.knownGene` (v3.10.0) database to annotate the TSS locations to genes using ChIPseeker (v1.32.0). In order to retrieve exactly one transcription start site per gene, we fetched the TSS location with maximum CAGE score within 30 kb of the gene's reference TSS.

HiC

We downloaded JuicerTools[2] and used the ABC provided scripts *Juicebox_dump.py* and *compute_powerlaw_fit_from_hic.py* to prepare our macrophage specific HiC data for the ABC workflow.

ABC candidate regions

We determined candidate regions by using reads from both replicates with the setting *"--nStrongestPeaks 150000"*. We used the *"--regions_includelist" argument* to add macrophage specific TSS, promoter regions determined from a genomic reference as well as areas +- 250 bp around GR summits and excluded blacklisted regions.

[1] https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction

[2] https://hicfiles.tc4ga.com/public/juicer/juicer_tools.1.9.9_jcuda.0.8.jar

ABC scores

The workflow generates ABC scores for the assignments of regions E to their putative target genes G. The activity of a region gets weighted by the amount of 3D contacts with the target gene and divided by the total effect of all regions on that gene.

$$ABCscore_{E,G} = \frac{A_E * C_{E,G}}{\sum_e A_e * C_{e,G}} \tag{3.1}$$

where:

- A_E is the enhancer activity
- $C_{E,G}$ is the 3D contact of the enhancer E with the promoter of gene G
- e are all elements within 5Mb of G

In detail, the contacts are quantified as powerlaw scaled HiC contacts between the putative regulatory regions with promoters of the genes. For our project, we computed the region activity based on H3K27ac data from 2 replicates. To ensure that region-gene connections would only be assessed for genes that are expressed within the samples, we also provided the workflow with condition specific average TPM values from the 4sU-seq experiment. We provided condition specific (LPS and Dex+LPS treated) samples for the ATACseq data, 4sU-seq data and H3K27ac data, resulting in condition specific ABC scores. We retained regions with $ABCscore \geq 0.02$ for downstream analyses.

3.2.6. Motif analysis

Genome wide motif scans

We used homer (v4.11) and its script *scanMotifGenomeWide.pl* to perform genome wide scans with a simplified NR3C1 fullsite or halfsite motif. The motif was simplified by assigning non-dominant bases a weight of 0.001 and the dominant base a weight so that the total of the position would sum up to 1. For the full length motif we applied a threshold of 5 when searching with the pattern [AG]GNACANNNTGTNC[CT] and for the halfsite motif we applied a threshold of 6 and searched with the pattern [AG]G[ACGT]ACA.

De novo motif analysis

We used *STREME* from the tool MEME Suite (v.5.4.1) to find enriched motifs compared to a shuffled input control.

Motif matches

We used *FIMO* from the tool MEME Suite (v.5.4.1) to find matches of known motifs within our input sequences. Motifs were retrieved from the 2022 release of the JASPARdb database and filtered for binding motifs whose transcription factors were expressed in our 4sU-seq data. In the case of composite motifs at least one of the transcription factor partners had to be present in our expression dataset in order for us to consider them expressed. Visualization of the motifs was done with the R packages memes (v1.0.0) and universalmotif (v1.10.1).

3.2.7. Generalized linear models

Feature engineering

Our objective is to perform binary predictions regarding whether genes get up- or downregulated in response to glucocorticoid treatment. We base this prediction on the occurrence of known TF binding motifs within the selected input regions, which is represented by a region x motif countmatrix **C**. It is important to note that the labels are on the gene level, while the motif counts are derived from regions. To bridge this gap, we assign regions to their likely target genes and represent this assignment by a gene x region matrix **A**. This allows us to compute a weighted sum of motif counts from all regions mapping to the same gene and results in the final genes x motifs feature matrix **F**.

$$\mathbf{F} = \mathbf{A} \cdot \mathbf{C} \tag{3.2}$$

$$f_{gm} = \sum_{i=1}^{r} a_{gi} \cdot c_{im} \tag{3.3}$$

The final feature matrix **F** is determined by several choices we can make when constructing the matrices **C** and **A**. We refer to these choices as "feature engineering choices" throughout the project.

When constructing **C**, we need to decide what regions to use as input when quantifying motif occurrences. We decided to investigate 100 bp regions centered on summits of GR peaks (GR summitregions) and active regions identified by the ABC workflow (active regions).

When constructing **A**, there are considerably more choices to make. The simplest approach is used as our reference model. In this case we derive the region-gene assignment through linear proximity using ChIPseeker's (v1.28.3) *annotatePeak* function in combination with TxDb.Mmusculus.UCSC.mm10.knownGene (v.3.10.0) as genomic reference and a maximum distance of 30 kb. It follows that the assignments are binary in that each region is assigned to only one gene. In more elaborate approaches, we derive the assignments from the ABC workflow and either treat them as binary variable (whether or not the score passes a predefined threshold) or as a continuous one by using the ABC score itself. Binary values result in an unweighted aggregation during the matrix multiplication step, whereas continues values make it a weighted aggregation. Of note, when testing the combination of GR summitregions with ABC-based assignments, we assigned a zero weight to those summitregions that did not overlap with putative regulatory regions identified in the ABC workflow.

In contrast to the annotation method based on proximity, ABC scores do not return a one-to-one assignment between regions and genes. This has provided an opportunity to systematically assess whether it would yield better results to utilize the one-to-many mappings generated by the ABC workflow or to simplify them into one-to-one mappings, associating regions with specific gene targets. Attaining a one-to-one mapping based on the ABC outcomes can be achieved by considering only the highest ABC score for each region and assigning zero weights to all other associations for that region within the matrix **A**.

Furthermore, the ABC workflow provides details regarding whether a region overlaps with a genic, intergenic, or promoter region, as well as whether it represents the promoter of the target gene or a different gene (nonself). We tested models that treat all regions uniformly, regardless of their genomic positioning relative to the gene, and models that incorporate this information into the feature engineering procedure. If we incorporate the genomic location of features concerning genes as part of the feature engineering process, each pairing of a gene and region provides additional information about whether the region resides within the promoter, genic, or intergenic region of that specific gene. A weighted average is separately calculated for each category of genomic location, and the resulting individual feature matrices are concatenated to construct the final feature matrix used for subsequent model fitting. Consequently, in this scenario, the dimensions of the final feature matrix become **F** (g x 3m). In the project, we collectively refer to genic and intergenic regions as enhancers.

Additionally, this approach allows us the flexibility to either omit all promoters or solely nonself promoters from the fitting process, leading to different number of features in the downstream workflow. Depending on the decisions made during feature engineering, the number of input features ranged from 398 to 1,224.

ABC scores and the active regions identified using the ABC workflow are specific to particular conditions, which in turn renders the feature matrix **F** condition-specific. Instead of selecting features based on information from a single condition, an alternative approach involves subtracting the matrices and utilizing the difference between the Dex+LPS and LPS conditions after aggregation as input for the modeling process.

$$\mathbf{F}_{Dex+LPS-LPS} = \mathbf{F}_{Dex+LPS} - \mathbf{F}_{LPS} \tag{3.4}$$

Lastly, depending on the decisions made during feature matrix generation, the number of genes integrated into the model also fluctuated. In instances where a target gene lacks associated input features, it will be dropped. Within the training dataset, the count of negative and positive labels ranged from 126 to 709 and 139 to 673, respectively. Within the test dataset, the count of negative and positive labels ranged from 35 to 168 and from 31 to 154, respectively. It is worth noting that the labels exhibit a balanced distribution, with an average positive-to-negative label ratio of 0.4955 (median=0.4952) in the training set and an average of 0.4749 (median=0.4870) in the test set.

Model fitting

Before fitting the generalized linear model (GLM), we removed features without any matches across all genes. Subsequently, all remaining features underwent scaling and centering using *scale* from the R-base package (v4.1.3) to prevent frequent motifs from dominating the model outcomes. For the evaluation process, genes located on chromosomes 1, 8, and 9 were set aside to form the test set. The remaining chromosomes constituted the training set. Within the training set, we conducted a 6-fold cross-validation, employing a GLM with elastic net regularization via the R package glmnet (v4.1.2). In detail, we configured the GLM with the family parameter set to "binomial" and the mixing parameter alpha set to 0.5. The primary goal during cross-validation was to identify the optimal regularization parameter λ, utilizing the Area Under the Curve (AUC) as the performance metric. The optimal λ, determined through the training set, was then employed when making predictions and assessing the model's performance for the test set using the R package ROCR (v1.0.11).

We compared ROC curves and checked the model performance differences for significance with the R package pROC(v.1.18). To test whether a model of interest (either the best performing model or our reference model) performed significantly better than the other models (making pairwise comparisons), we used a directional alternative hypothesis in *roc.test* with DeLong's method. We utilized Storey's q-value method, which is implemented in the R package qvalue (v2.24.0), to process the resulting p-values. In this analysis, the lambda search space was confined to the range where p-values were observed, which allowed us to estimate the percentage π_0 of true null hypotheses.

Regarding heatmap visualizations, it is important to note that coefficients from models containing separate features based on the genomic location of a motif cannot be effectively displayed alongside coefficients from models where these features are aggregated. This is because they involve different sets of features, with one set comprising multiple features derived from a single motif, while the other set consists of a single feature. To address this, we assessed which approach yielded better results. Among the top 25 best-performing models based on AUC on the test data, 17 models aggregated information from both promoter and enhancer regions. In the heatmaps displaying coefficients of these top models Figure 3.5B, we thus displayed those 17 models and determined the number of models in which each motif had a non-zero coefficient. This information was visualized as a barplot in the row annotations.

3.2.8. HINT footprinting

We used the python library Regulatory Genomics Toolbox (RGT) (v0.13.2) to run *"rgt-hint footprinting"* and *"rgt-motifanalysis matching"* for each condition. Finally, we check for changes in binding by comparing cleavage profiles of the matched footprints from Dex+LPS and LPS with *"rgt-hint differential"*.

3.2.9. Bone-marrow derived macrophage cell culture

We surgically removed the tibia, femur, and humerus from male C57BL6/N mice aged between 8 to 12 weeks. These bones were cleaned and surface-disinfected using ethanol before bone marrow extraction in RPMI-1640. We then lysed erythrocytes using AKC lysis buffer (1M NH4Cl, 1M KHCO3, 0.5M EDTA). Subsequently, the cells underwent density centrifugation through a Ficoll-Paque gradient and were cultured in differentiation medium (DMEM containing 30% L929 supernatant, 20% FBS, and 1% penicillin/streptomycin) for 6 days on bacterial plates at 37°C with 5% CO2. Afterward, cells were detached using Versene, counted, and seeded in macrophage serum-free medium. Following an overnight incubation, the cells were treated with either Vehicle (PBS), LPS (100ng/mL, Sigma Aldrich, LPS25), or Dex+LPS (100ng/mL LPS, 1µM Dexamethasone, Sigma Aldrich, D2915) for 3 hours or as indicated.

3.2.10. Nuclear extracts

For nuclear extracts, 2×10^7 cells on 15cm cell culture dishes were washed once with ice-cold PBS. They were then transferred to a 1.5mL microcentrifugation tube and lysed on ice in V1 lysis buffer (10mM HEPES-KOH pH 7.9, 1.5mM MgCl2, 10mM KCl, and freshly added 1µM Dexamethasone, 0.5mM DTT, 0.15% NP40, protease inhibitors, and phosphatase inhibitors) using a micro-pistil. After centrifugation (2700 xg, 20min), nuclei were collected and lysed in V2 buffer (420mM NaCl, 20mM HEPES-KOH pH 7.9, 20% glycerol, 2mM MgCl2, 0.2mM EDTA, and freshly added 1µM Dexamethasone, 0.5mM DTT, 0.1% NP40, protease inhibitors, and phosphatase inhibitors) by rolling for 1 hour at 4°C. Subsequently, nuclear lysates were subjected to centrifugation (21000 xg, 45 minutes, 4°C), and the resulting supernatants were used for western blot analysis.

3.2.11. Western blots

For western blot analysis, cells lysed in RIPA buffer (containing 150mM NaCl, 50mM Tris pH 7.4, 1% NP40, 0.5% DOC, 0.1% SDS) or nuclear extracts underwent sonication for three cycles of 10 seconds each. Following sonication, these samples were boiled in Laemmli buffer (62.5mM Tris pH 6.8, 1% SDS, 0.8% glycerol, 1.5% 2-mercaptoethanol, 0.005% bromophenol blue) at 95°C for 10 minutes. Standard western blot procedures were then carried out using appropriate antibodies. For a complete list of antibodies please refer to the original publication [2].

4. Exhaustion in CD8+ T cells

T cell exhaustion, a state of reduced effector function and reduced proliferation, poses a major challenge for cancer immunotherapy. Looking at viral diseases, exhausted effector T cells and their long-lived progenitor population have been described in chronic infections, which are associated with prolonged antigen exposure, since the infection cannot be fully cleared by the immune system. However, the mechanisms leading to exhaustion remain poorly understood. For the project in section 4.1 our goal was to understand the early developmental steps driving the generation of these TCF1+ TOX+ hypofunctional progenitors of exhausted T cells (Tpex).

Knocking out the transcription factor TOX increases the effector function of cytotoxic T cells in the context of chronic infection. Regardless, this knowledge can currently not be therapeutically harnessed since these reactivated T cells fail to be maintained long-term. For the project in section 4.2 we set out to find a way to sustain these cells and what is more, to preferentially expand non-exhausted cells in a context of chronic antigen exposure.

4.1. Development of exhausted progenitors in acute infection

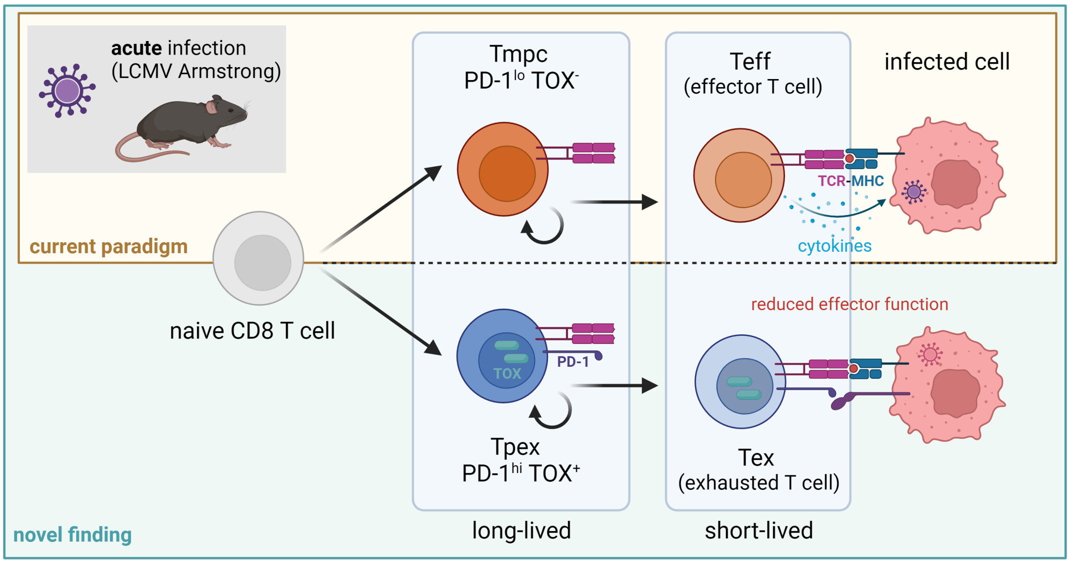

Figure 4.1.: **Preliminary findings of Tpex in acute infection challenge the current paradigm.** Prior to me joining this project our collaborators discovered TCF1+ PD-1hi TOX+ cytotoxic T cells following infection with LCMV Armstrong. Schematic created with BioRender.

Looking into early timepoints after infection, my collaborators found that Tpex are not only present during chronic infection but, in a smaller number, TCF1+ PD-1hi TOX+ cytotoxic T cells are also present during acute infection (Figure 4.1). This observation challenged the current paradigm, which assumes that exhausted T cells are exclusively generated in chronic infection or tumor settings.

4.1.1. Bulk RNA-seq identifies Tpex in acute infection

To investigate whether the TCF1+ TOX+ cells found in acute infection are bona fide progenitors of exhausted T cells (Tpex), we decided to do deeper phenotyping with RNA-seq and investigate whether their transcriptional profile matched that previously reported for exhausted T cells. Our collaborators used a reporter mouse, which allows for selection of early TCF1 expressing P14 progenitors, which were adaptively transferred into host mice. 7 days after infecting the host mice with LCMV, TCF1 expressing P14 progenitors were isolated and sorted into two groups based on PD-1 expression levels, namely PD-1lo and PD-1hi progenitors. This gating scheme allows to enrich for TOX+ and TOX- cells based on surface receptors, since direct staining of TOX is incompatible with sequencing library preparation (Figure B.1). Finally, the sorted progenitors were submitted for RNA-sequencing.

To visually inspect the similarity between samples in lower dimensional space, we performed principal component analysis (PCA). The PCA revealed that the component responsible for the majority of variation in the data corresponds to the PD-1 status of the samples, indicating that are two groups are transcriptionally different (Supplemental Figure B.2A,B). Comparing the PD-1hi progenitors to the PD-1lo progenitors with differential gene expression analysis, we identified a total of 2550 differentially expressed genes (DEGs) (Figure 4.2A). Among these DEGs, 1236 were found to be more highly expressed in the PD-1hi population, while 1314 were more highly expressed in the PD-1lo population (adjusted p-values < 0.05 and log2FC > log2(1.5) or < -log2(1.5)) (Figure B.2C).

To bring biological meaning to our set of DEGs, we utilized publicly available data of acute and chronic genesets and found that our PD-1hi population matches previously reported chronic gene signatures while our PD-1lo population concurs with acute signatures (Figure 4.2B). We independently validated these findings with public data of CD8 T cells at day 21 post infection with chronic (Docile) or acute (Armstrong) LCMV [166]. In detail, we compared the transcriptional differences between our PD-1hi and PD-1lo progenitors with the differences observed between CD8 T cells in chronic and acute infections and find that they correlate with r=0.517 (Figure 4.2C). These results indicate that PD-1hi sorted progenitors exhibit a strong exhaustion signature similar to Tpex cells described in chronic infections [167, 166].

Among the DEGs of the two progenitor types, we observed lower levels of *Id2* and effector cytokines (*Ifng* and *Tnf*) in PD-1hi progenitors. Given that our samples were isolated at day 7, this suggests that these cells are programmed early on, even during acute infections, to express fewer inflammatory cytokines and maintain limited effector function (Figure 4.2D). Furthermore, we found increased expression of inhibitory receptors associated with exhaustion such as *Pdcd1*, *Havcr2*, and *Tnfrsf9*, as well as other genes significantly upregulated in

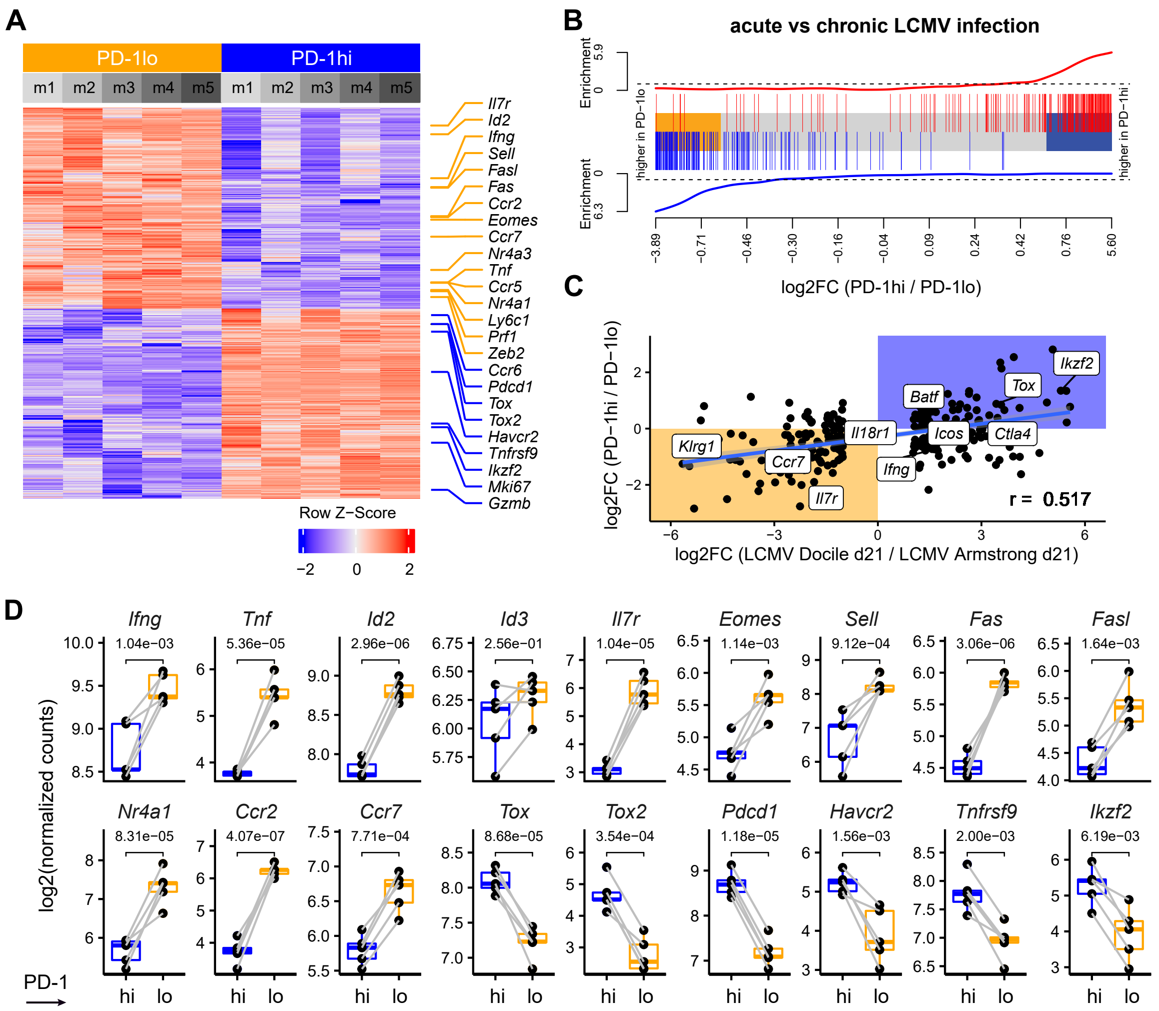

Figure 4.2.: **Transcriptional differences between PD-1hi and PD-1lo progenitors confirm exhausted phenotype.** (A) Heatmap of differentially expressed genes (adjusted p-value < 0.05 and |log2FC| > 0.58) between PD-1hi and PD-1lo P14 samples. (B) Gene signatures of acute vs chronic LCMV infection plotted onto log2FC between PD-1hi and PD-1lo samples. Red dashes indicate genes higher in chronic infections, blue dashes indicate genes that are higher in acute infection. Blue and yellow boxes mark |log2FC|>0.58. (C) Correlation of log2FC between PD-1hi and PD-1lo P14 with the log2FC between LCMV Armstrong and LCMV Docile infected ID3+ cells at day 21 post infection. (D) Log2 normalized counts of bulk RNA-seq with paired PD-1hi and PD-1lo P14 samples from the same donor mouse. Significance values represent adjusted p-values determined with a linear model. n=5 for both groups.

exhaustion, including *Tox*, *Tox2*, *Nr4a1*, and *Ikzf2*, in PD-1hi progenitors (Figure 4.2D). In contrast, PD-1lo progenitors exhibited characteristics resembling traditional memory cells, with high expression of *Il7r*, *Eomes*, *Sell*, and *Id3* (Figure 4.2D).

Our observations show that during the early phase of acute infection, the formed progenitors range from PD-1lo progenitors displaying the transcriptional signature previously reported in acute infections, to PD-1hi progenitors resembling Tpex. This is remarkable, since Tpex have so far been thought to be exclusively associated with chronic infections. In summary, the data demonstrate the generation of a diverse set of progenitor T cells in the early stages of infection, irrespective of whether the infection eventually becomes chronic or resolves.

4.1.2. Tpex are epigenetically different from Tmpc

After we identified Tpex with transcriptional profiling, we wondered whether the Tpex found in early infection also resemble the Tpex found in chronic infection on an epigenetic level. We turned to a public dataset on total splenic CD8 T cells that were isolated 7 days after acute (LCMV Armstrong) or chronic (LCMV clone13) infection and analyzed by scATAC-seq [168]. We started our investigation by performing a joint clustering of the cells from both infections (Figure 4.3A (left)). In order to annotate the resulting clusters, similar to the original study [168], we used signatures obtained from bulk ATAC-seq samples (for details see subsection 4.3.2). Two of the clusters contained cells with high scores for the exhaustion signature; one that matched the signature of terminally exhausted cells (Tex) and one that matched the signature of exhausted progenitors (Tpex) (Figure 4.3B). Additionally a group of cells displayed high scores for the memory precursor effector cell (MPEC) signature but low scores for exhaustion, so we termed it memory precursor T cells (Tmpc).

To validate the cluster identity, we compared the Tpex and Tmpc clusters with differential accessibility analysis. We annotated differentially accessible (DA) regions to their closest gene and aggregated them based on their gene annotation. Our investigation revealed that the Tpex cluster is significantly more accessible at multiple regions proximal to the exhaustion markers *Tox*, *Tox2*, and *Pdcd1* in comparison to the Tmpc cluster (Figure 4.3C). Of note, the Tpex cluster contained cells from both acute and chronic infections. We wondered whether these cells from the two infection types would show substantial heterogeneity despite being in the same cluster and directly compared accessibility of Tpex from Armstrong and clone13 infection. None of the 4686 tested regions showed significant differences (all adjusted p-values > 0.1), indicating that Tpex from the two infections indeed share similar epigenetic profiles. Accessibility for regions proximal to *Pdcd1* are displayed in Supplemental Figure B.3A.

Moving forward, we focused on cells infected with LCMV Armstrong to investigate the previously unrecognized heterogeneity within the CD8 T cells in acute infection. Since we previously used PD-1 to enrich for exhausted cells (see subsection 4.1.1), we decided to look into the accessibility of its genomic locus. We visualized coverage at the locus of *Pdcd1* (the gene encoding for PD-1), split by cluster (Figure 4.3D) and highlighted the regions identified as DA in the comparison of Tpex and Tmpc by shading those areas in grey. Accessibility patterns are similar between Tex and Tpex while the Tpex and Tmpc populations show substantial differences. Taken together this shows that there are differences between Tpex and Tmpc at the *Pdcd1* locus even when we exclusively look into cells originating from acute infection.

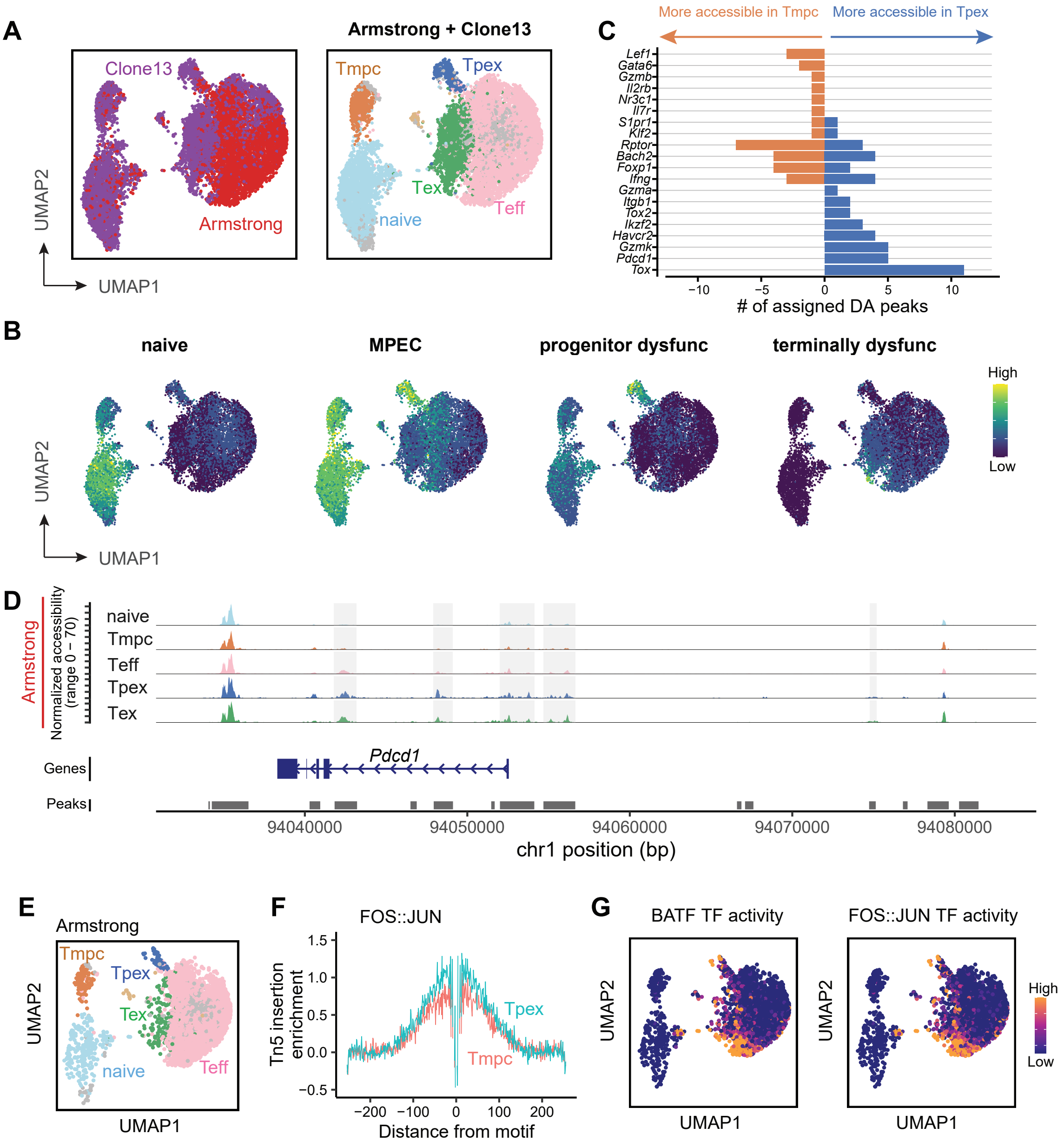

Figure 4.3.: See legend on next page.

Figure 4.3 *(previous page)*: **Tpex from acute infection epigenetically resemble their counterpart from chronic infections.** Public scATAC-seq data of total splenic CD8 T cells analyzed 7 days after mice were infected with either LCMV Armstrong or LCMV clone13. (A) UMAP embedding of cell similarity with cells colored by infection type (left) or results of Leiden clustering (right). (B) Signatures for naive, exhausted, progenitor exhausted and terminally exhausted cells derived from bulk ATAC-seq data are used to compute signature scores for each cell. Colors represent the mean expression level of the signature within the cell shown on the UMAP embedding. (C) Number of differential accessible regions between Tpex and Tmpc cluster annotated to genes of interest. (D) *Pdcd1* locus accessibility signal of cells from acute infection. Marker regions of Tpex cluster are highlighted in gray. (E) UMAP embedding of cell similarity subset on cells from the LCMV Armstrong infection. Colors represent cell assignments to Leiden clusters. (F) Accessibility signal at FOS::JUN motif locations. (G) TF activity for BATF and FOS::JUN motif. MPEC = memory precursor effector cell

We were wondering whether we could find putative regulators that drive the differences between Tpex and Tmpc within acute infection. We fetched known transcription factor binding motifs from the JASPAR database and assessed motif activity by computing the deviation from the expected accessibility [169]. Comparing motif activity between Tpex and Tmpc, we discovered a total of 154 motifs that are significantly (adjusted p-value $<$ 0.05) more active in Tpex. Remarkably, a substantial proportion of the most significant hits was associated with members of the AP-1 transcription factor family (Supplemental Figure B.3B). Our analysis revealed increased accessibility around FOS::JUN sites in Tpex compared to Tmpc (Figure 4.3F), consistent with the heightened activity of AP-1 family motifs (Figure 4.3G).

Taken together, our observations underscore that the open chromatin landscape typically associated with Tpex in chronic infection is also observable in the Tpex population found in the early phase of acute infection.

4.1.3. Exhaustion is linked to TCR sequence

TCR signalling strength and as such the T cell activation, has been reported to drive exhaustion in tumor models [170]. We sought to investigate whether activation strength had the same effect on exhausted cells in acute infection. Wet-lab experiments encouraged us to investigate the endogenous response which, contrary to the P14 system, has a diverse TCR repertoire. We used np396 and gp33-loaded tetramers, which have previously been described to lead to high and medium activation, respectively [171]. Through this tetramer staining we can biochemically pull down T cells recognizing the gp33 and np396 epitopes. We followed the Tpex into the memory phase and isolate TCF1 expressing memory gp33+ or np396+ T cells from reporter mice 4 weeks after infection with the acute LCMV strain. By performing scRNA-seq combined with scTCR-seq analysis we set out to investigate the potential link between exhaustion and TCR sequence (see Figure B.4 for an experimental overview).

We decided to first investigate the effects of TCR sequence in a visual manner. The tool mvTCR [172] allowed us to create a joint embedding of TCR-seq and GEX so that both data modalities contribute to the resulting UMAP representation (Figure B.5A). We used the amino acid sequence from the CDR3 region of the TCR α and β chain to define clonotypes. Additionally, we identified Leiden clusters based on the cells gene expression and passed both group assignments to mvTCR to optimize for those class labels when creating the embedding. The clonotype modules are successfully captured within the UMAP representation, as seen by cells with the same clonotype grouping together in distinct areas of the plot (Supplemental Figure B.5B). Normalized Tox levels were higher in cells recognizing the high activation (np396) compared to medium activation (gp33) epitope and seemed to be associated with certain clonotype modules (Supplemental Figure B.5D).

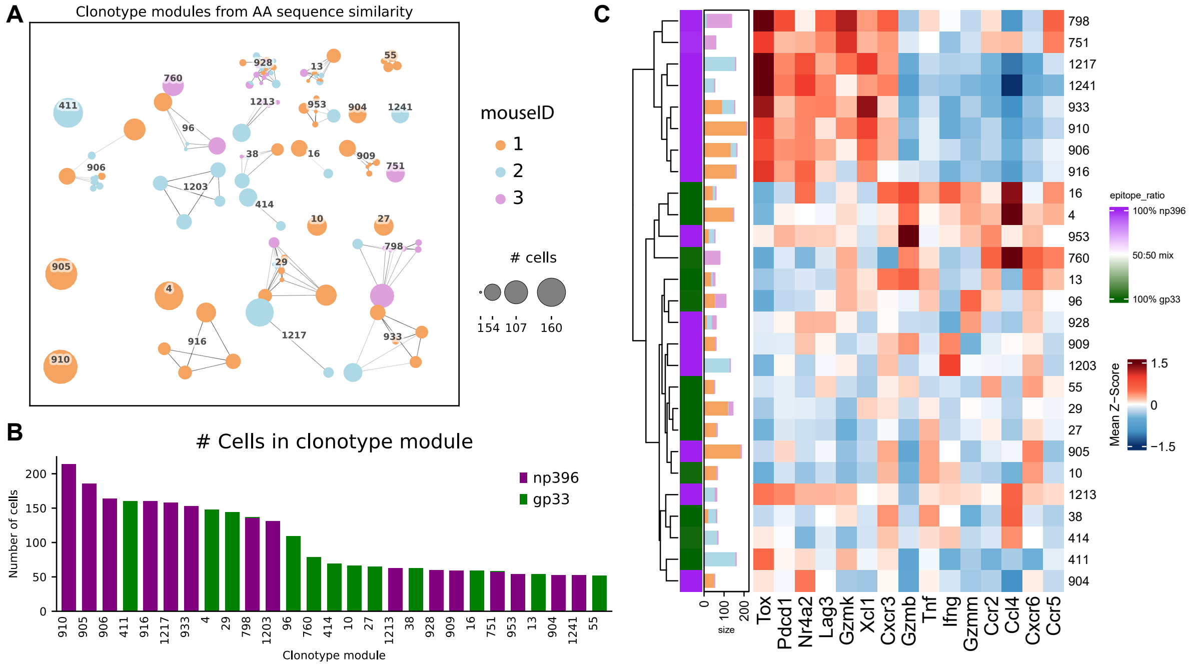

Figure 4.4.: **scRNA-seq analysis shows that exhaustion is dependent on TCR signalling.** gp33 and np396-specific TCF1+ progenitors were purified from Tcf1 reporter mice infected with LCMV 4 weeks post infection. Results shown for all mice (n=3). (A) Clonotype network showing all modules containing at least 50 cells. Clonotype pairs were aligned using their amino acid sequence and similar clonotypes (i.e. distance below 10 when utilizing a BLOSUM62 matrix) are connected by an edge. Node size depicts clonotype size, color represents the mouseID. (B) Number of cells per clonotype module, colored by the tetramer they were pulled down with. (C) Mean z-score of signature genes shown per clonotype module. Row annotations show the epitope ratio of np396 and gp33 within the module. Dendrogram was generated using euclidean distance and complete linkage. Barplots in the row annotation show the number of cells within the clusters colored by mouseID.

Independently, we used the above mentioned clonotype definition to group them together into modules based on sequence similarity (Figure 4.4A). Due to the high likeness of the

CDR3 sequences within a module, they likely posses similar biochemical properties which is supported by a module containing cells pulled down with gp33 or np396, but rarely a mix of both (Figure 4.4B). Looking at mean gene expression levels per module, we find that the exhaustion phenotype comprised of high levels of *Tox*, *Pdcd1*, *Nr4a2* and *Lag3* is specific to the np396+ clonotype modules, although not all np396+ modules are exhausted (Figure 4.4C). This matches observations we made using flow cytometry (data not shown) and confirms that recognition of the high activation epitope makes cells more likely to become exhausted.

Taking into consideration that np396 is deemed a higher affinity epitope compared to gp33, we conclude that the formation of a long-lived Tpex population requires strong TCR stimulation. Weaker stimulation could either lead to the formation of short-lived Tex cells or the Tpex cell might simply become outnumbered by the more frequent Tmpc over time. Regardless, our data show that strongly activated Tpex in acute infection have the capacity to survive long-term, even after the antigen is cleared, as was previously shown for Tpex derived from chronic infection [173].

4.1.4. Conclusion

In this project we show that the TOX+ cells found in acute infection are transcriptionally and epigenetically bona fide Tpex and that they are maintained past the clearance of antigen. Furthermore, wet-lab follow-ups and scRNA-seq and scTCR-seq identified that activation strength is the underlying mechanism driving exhaustion (see Figure 4.5).

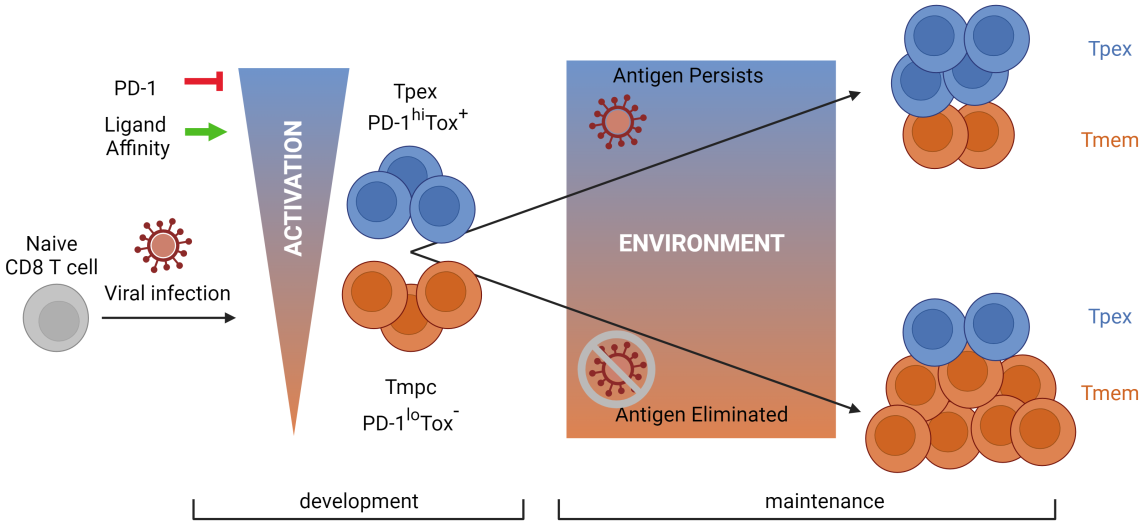

Figure 4.5.: **Graphical abstract of Tpex generation.** Stronger T cell activation leads to generation of higher numbers of Tpex cells compared to Tmpc. Environmental stimuli preferentially maintain and expand one cell population over the other. Schematic created with BioRender.

4.2. Maintenance of non-exhausted progenitors in chronic infection and cancer

In section 4.1 we have shown that there is a small proportion of exhausted cells present in acute infection. While this is conceptually very interesting, from a therapeutic standpoint the more pressing question is how to revert that exhaustion phenotype or expand non-exhausted cells in a targeted fashion. Furthermore, it is of interest to find genetic manipulations that render T cells insusceptible to exhaustion, as this could therapeutically be deployed through engineered CAR T cells [174]. Since TOX has been identified as central transcription factor of exhausted cells, researchers hoped that targeting TOX might offer a strategy to prevent cells from becoming exhausted in scenarios of chronic antigen exposure. While rendering TOX inactive through genetic engineering, i.e. TOX knockout (TOX KO), initially leads to increased effector phenotype, the number of TOX KO progenitors starts to massively decline about two weeks into a chronic infection [48]. In other words, TOX KO successfully leads to a non-exhausted phenotype in CD8 T cells, but in an environment of chronic antigen exposure they fail to be maintained. Our goal is to understand how we can maintain TOX KO progenitors and thereby open this up as potential therapeutic avenue.

4.2.1. Neural networks show link between KLF and TOX KO phenotype

Our first approach to understanding the underlying gene regulation was to investigate whether we could find DNA sequences linked to the transcriptional differences between WT and TOX KO CD8 T cells. The idea behind this is that if the downstream effects of TOX KO were mediated through a specific transcription factor, the corresponding binding motif would predict the transcriptional response of its target gene.

Existing sequence models such as Enformer are good at predicting gene expression of samples that the model has been trained with but predicting transcriptional changes in response to stimuli is still an open challenge [134]. Our goal was to devise an interpretable neural network approach that would identify sequences in the regulatory region of a gene which determine whether it is up- or downregulated in response to a perturbation or stimulus. We decided to do this in a knowledge-guided fashion by looking into regulatory regions predefined by celltype specific assays and opted for a computationally inexpensive architecture using local attention instead of the self-attention with quadratically growing complexity.

Since the number of differentially expressed genes between TOX KO and WT are too small to train a classification task from scratch, we decided to turn to transfer learning. We pretrained the model on read count tracks from epigenetic data modalities generated using similar experimental designs as the the RNA-seq data (same infection, same timepoint and comparable genotypes), in order to learn a related task with more available training data. Namely, we retrieved publicly available ATAC-seq data [48] and ChIP-seq data [175] from CD8 T cells of WT and TOX KO mice at day 8 after LCMV clone13 infection. We performed preprocessing for the samples (see section 4.3 for details) and used peaks from the ChIP-seq and ATAC-seq samples to define a peak universe of 55,313 peaks. These peaks served to define regions of interest during the pretraining. Intuitively, the model learns the link between genetic sequence and TOX binding as well as the TOX KO driven changes in accessibility in this step.

Our input data contains information on regulatory sequences, but the labels we have for training the expression change are on gene level. We approach this by deploying multiple instance learning (MIL), which can utilize this weakly labelled data. In this setting, regulatory regions annotated to the same target gene are placed in the same bag and the direction of the gene expression change is the bag label. During the expression change prediction, we apply transfer learning by using our pretrained model weights to initialize an architecturally similar network. In addition to the body, which is identical to the architecture used for pretraining, it contains a head that uses local attention to accommodate a multiple-instance learning problem (Figure 4.6A). The attention mechanism learns the attention of each region and aggregates region level features on gene level through a weighted sum. This forms latent gene level features that are used to classify on the gene level.

Performance of the pretraining was good with correlation of the observed and the predicted counts ranging between 0.48 and 0.69 (Figure 4.6B). Of note, the sample that the model performed worst on, was ChIP-seq data from TOX KO cells. Contrary to all other TOX KO models used within this project, these samples were generated from a KO model that was missing exon 1 rather than exon 5 which contains the DNA-binding region. We had the suspicion that this model might have residual activity of TOX, which is why we included the TOX KO samples in the model training.

We performed hyperparameter tuning (see Table B.1 for details) to determine the best architectural choices for the expression change prediction. The final model achieved a rather poor AUC of 0.57 in the test set (Figure 4.6C), which suggests that the model only manages to capture a small part of the underlying regulatory mechanism. Despite its humble performance, we decided to investigate what patterns the model had learned and found that a sequence pattern reminiscent of KLF binding motifs is associated with genes which are more highly expressed in the TOX KO cells (see Figure 4.6D). We did not find a pattern associated with genes that are more highly expressed in WT.

A possible explanation for the mediocre performance of the model is that TOX binding is sequence-independent [176], which would make pretraining on TOX ChIP-seq data ill-fitted for finding sequence determinants of the transcriptional changes. Another explanation is that the underlying changes happen specifically in the TCF1+ progenitor population and using data of bulk sequencing from a mixed population, the signal gets washed out by the majority of effector cells.

4.2.2. TOX KO progenitors intrinsically differ from their WT counterpart

In order to directly look at transcriptional differences in the respective progenitor populations, our collaborators used FACS to enrich for progenitors before performing RNA-seq at day 7, day 10 and day 14 after infection with LCMV clone13. Having snapshots at multiple timepoints allows us to tease apart early phenotypic differences that drive subsequent downstream effects. We knew that TOX KO progenitors are still present at numbers comparable to WT at day 8 after infection but decline thereafter [48].

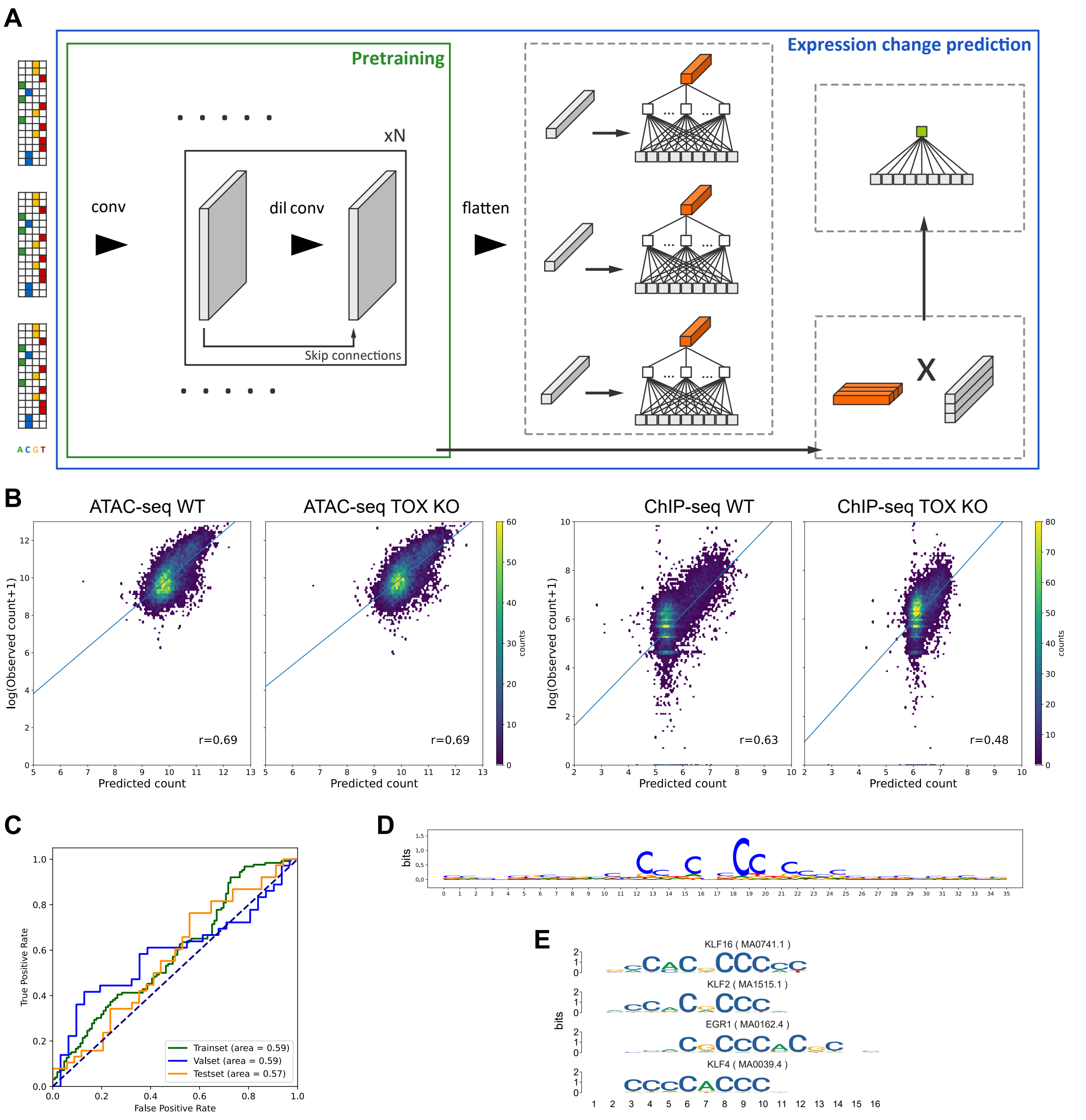

Figure 4.6.: **Neural network model identifies link between KLF binding sequence and TOX KO phenotype.** ATAC-seq and ChIP-seq data from TOX KO and WT CD8 T cells were used for pretraining and individual instances were combined for prediction of expression change on the gene-level. (A) Architecture of multiple instance learning neural network approach. (B) Correlation between predicted and observed counts of the ATAcseq (left) and ChIP-seq (right) samples used for pretraining. Colour indicates the number of points within the hexagonal bin. Counts exceeding the colour range are depicted with the max value. (C) Model performance for predicting labels of expression change between WT and TOX KO samples. (D) MoDISco sequence motif associated with elevated gene expression in TOX KO cells. (E) Sequence enriched in promoter regions of genes with significant interaction of genotype and time compared to promoters of non significant genes as background.

Based on this we decided to investigate a geneset consisting of genes with significant coefficient for the interaction term in the linear model $expression \sim genotype$ x $time$. We were curious whether we could reproduce the results found with our neural network approach, so we tested for sequence enrichment in the promoter regions of this geneset. Specifically, we compared the promoter sequences of the resulting 318 genes to promoter sequences of non-differentially expressed genes as background. Indeed, we found motifs resembling KLF binding motifs enriched in genes that are changed in response to the perturbation, confirming the motif found through our neural network approach (Figure 4.6E).

Sequence analyses indicated that KLF might be involved in the differences between WT and TOX KO cells. From there, we set out to identify a transcriptional link that could be therapeutically targeted with an external factor by doing an in-depth analysis of the transcriptional data. Differential expression analysis comparing WT and TOX KO progenitors at day 7 shows that even at this early timepoint a total of 727 genes (343 higher in TOX KO and 384 higher in WT) are differentially expressed (Figure 4.7A). This means that progenitors of the two genotypes show strong transcriptional changes before their numbers start declining [48]. Looking at the expression of known exhaustion markers *Ikzf2* (Helios) and *Pdcd1* (Figure 4.7C), we can confirm that the TOX KO progenitors are indeed less exhausted. A total of 211 genes were DE across all 3 timepoints (Figure 4.7B). One of those genes is the IL-2 receptor alpha chain *Il2ra* which is significantly higher expressed in TOX KO progenitors at all three timepoints (Figure 4.7C).

Pathway enrichment shows that at day 14 we find an enrichment for the apoptosis pathway (mmu04210, adjusted p-value$= 3.36e^{-04}$). This is in line with previous reports that the ratio of TOX KO compared to WT P14 cells is already reduced 2 weeks after infection [48]. Since we are interested in the early drivers of this effect, rather than the transcriptional changes that come hand in hand with cell death, we decided to focus on the changes at day 7 for downstream analyses.

4.2.3. IL-2 signalling is linked to changes in TOX KO progenitors

Given the sharp decline of TOX KO progenitors we wondered if we could leverage our RNA-seq data to find a target by which to treat them in order to maintain their numbers and function long-term. Cytokines and other ligands are crucial molecules that regulate immune processes by signalling through their cognate receptors, so we decided to investigate our data in the light of ligand analysis. To do this, we turned to the tool NicheNet [177] which uses prior knowledge on gene regulatory networks to predict links between ligands and gene targets.

We define the TOX KO progenitors from day 7 as receiver population and want to find out what ligands can best predict the DEGs between the KO and WT conditions. We filtered for ligands whose receptors were expressed in the receiver population and determined the ligands with highest ligand activity. Among the best upstream ligands we noticed the cytokine IL-2 which regulates a high number of target genes in our target set (Figure 4.8A). Its performance to predict DE genes had a corrected AUPR = 0.04, representing its performance increase over a random prediction. Looking at the receptors IL-2 is known to interact with, it comes as no

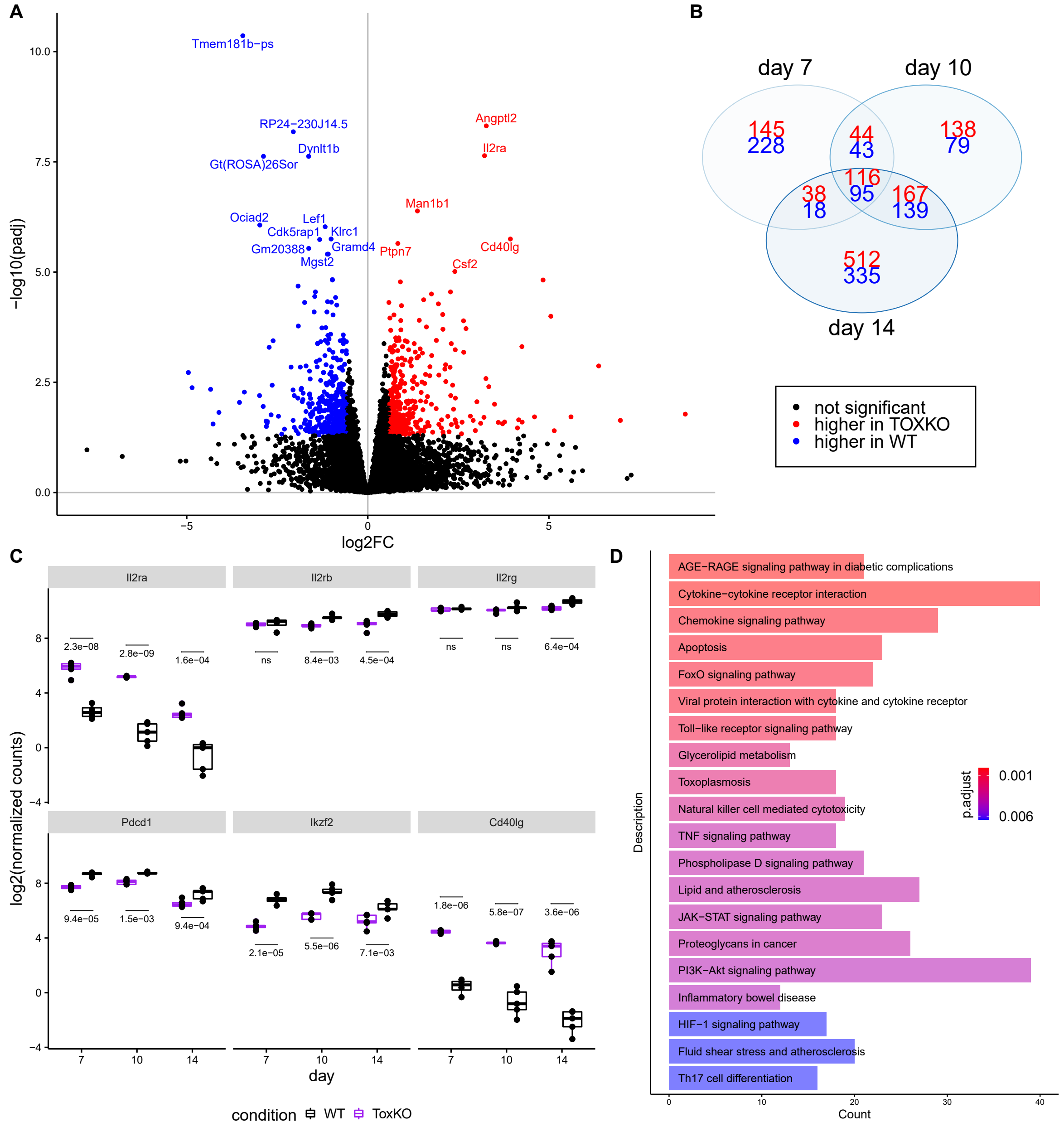

Figure 4.7.: ***Il2ra* expression is higher in TOX KO progenitors at all measured timepoints.** WT and TOX KO progenitors were sequenced at day 7, 10 and 14 after LCMV clone13 infection. (A) Differential expression analysis of TOX KO vs WT progenitors at day 7. Genes passing with adjusted p-value cutoff < 0.5 and log2FC greater/smaller 0.58 and highlighted in red/blue. (B) Venn diagram showing overlap in differentially expressed genes between 3 timepoints. (C) Boxplot of expression levels shown for select genes. Displayed values represent adjusted p-values as determined by limma. (D) Pathway enrichment of genes differentially expressed at day 14.

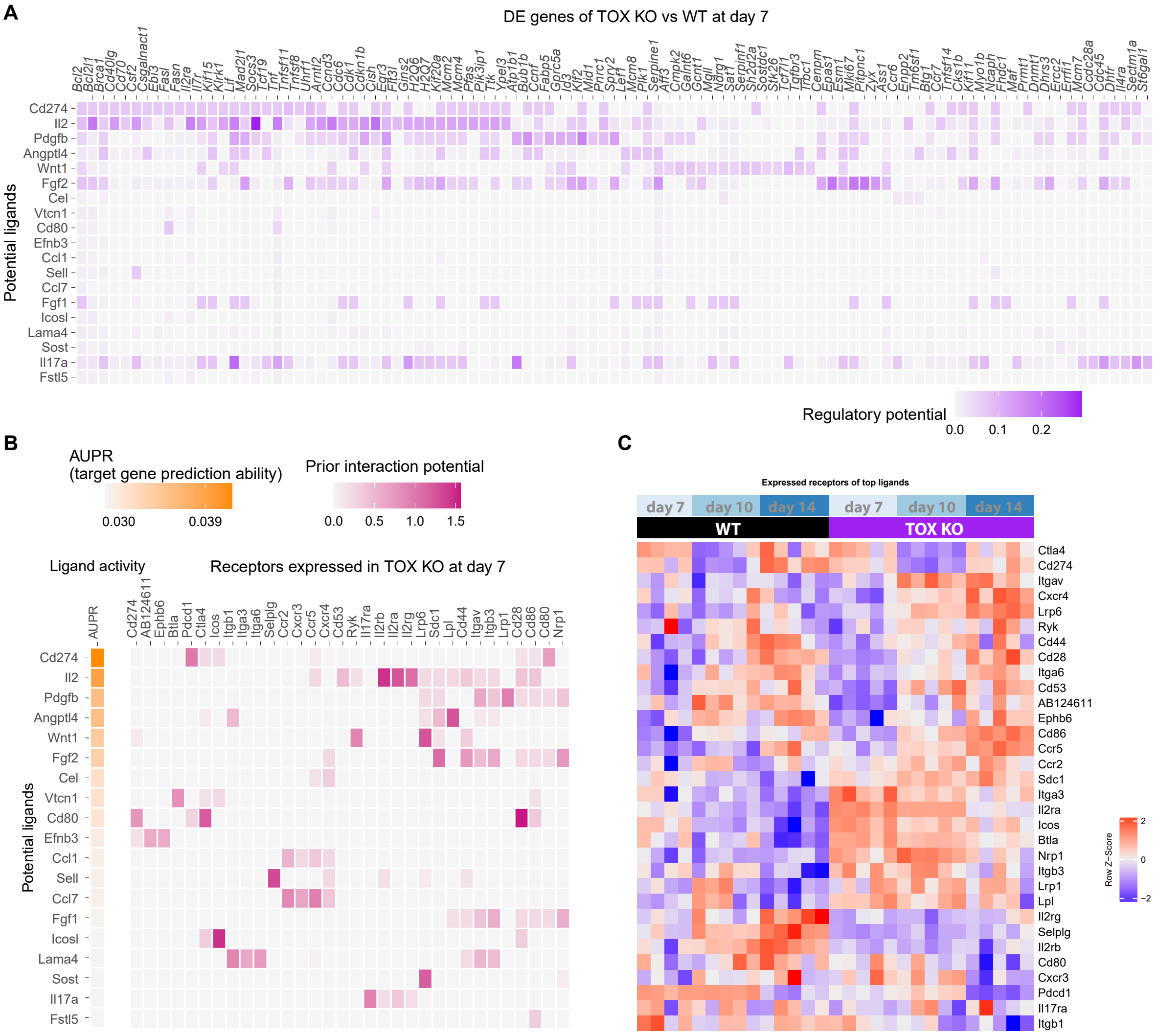

Figure 4.8.: **IL2 ligand activity explains differences between WT and TOX KO progenitors.** (A) Regulatory potential scores between best upstream ligands and targets of the target gene set. Scores below the quantile cutoff of 0.25 are displayed as white. (B) Prior interaction potential of best upstream ligands and their expressed receptors. (C) Expression pattern of receptors to best upstream ligands across all samples.

surprise to see that its prior interaction potential is highest with IL-2Rα, IL-2Rβ and IL-2Rγ, the components forming the high affinity trimeric receptor (Figure 4.8B). Visualizing the receptor expression of the best upstream ligands for both conditions across all timepoints, we noticed that only a couple of them displayed clear differences at day 7 and a consistent pattern across all timepoints (Figure 4.8C). Among them was *Il2ra* as mentioned before, whereas the expression of the receptor chains IL-2Rβ and IL-2Rγ was comparable or even lower in the TOX KO (Figure 4.7C).

From our project on early exhaustion in section 4.1 we know that in response to infection the immune system preemptively forms a wide range of heterogeneous CD8 T cell progenitor populations. Artificially pushing cells into a non-exhausted phenotype, as is done by the TOX KO system, leads to changes within the population that alter its IL-2 pathway activity. Knowing that IL-2 is an important survival factor for some T cell subsets [57], this could be the explanation for the declining numbers of TOX KO progenitors in a chronic environment. We hypothesize that the elevated *Il2ra* expression and expression changes in its downstream targets are caused by an increased dependence on IL-2 as survival signal.

4.2.4. IL-2 expands non-exhausted progenitors in chronic infection

We hypothesized that non-exhausted progenitors depend on IL-2 and that providing them with this cytokine would maintain their numbers. In order to test this hypothesis *in vivo*, our collaborators transferred a 1:1 ratio of WT and TOX KO CD8 P14 cells, into recipient mice, treated with IL-2 (daily for six days starting 12 days post infection) and assessed the number of exhausted and non-exhausted cells at day 18 post infection with LCMV clone13 compared to untreated controls.

At the endpoint, TOX KO cells were identified by their expression of the fluorescent transgene YFP, whereas WT cells were YFP negative (Figure 4.9A). We found that IL-2 treatment increased the total frequency of CD8 T cells and this appears to be due to increased expansion of both WT and TOX KO CD8 T cells (data not shown). Interestingly, the fraction of TOX KO cells among the total number of recovered CD8 P14 T cells increased significantly (p-value = 0.0286) from a median of 16.5% in the untreated control to 36.5% in the IL-2 treated samples (Figure 4.9A).

We continued by investigating the effect of IL-2 on exhausted vs non-exhausted CD8, separately for progenitor and effector cells. Interestingly, looking at flow cytometric data, we see that the progenitor population, identified as SLAMF6+ TIM3-, is highly increased in the TOX KO P14 cells treated with IL-2, whereas WT P14 appear to have an expansion of the SLAMF6- effector cells (Figure 4.9C). In other words, the treatment lead to an increase within the effector population of WT cells, whereas the TOX KO cells showed a successful expansion within the progenitor compartment (Figure 4.9D). This massive expansion in the non-exhausted, TOX KO progenitor population upon IL-2 treatment (Figure 4.9D) mirrors the increased ratio of TOX KO cells compared to WT cells (Figure 4.9A,B). To determine whether this IL-2 effect is specific to the P14 model system or whether it generalizes to the endogenous CD8 response, we turned to a public data set that explores the creation of a better effector CD8 population upon treatment with a modified IL-2 compound [61].

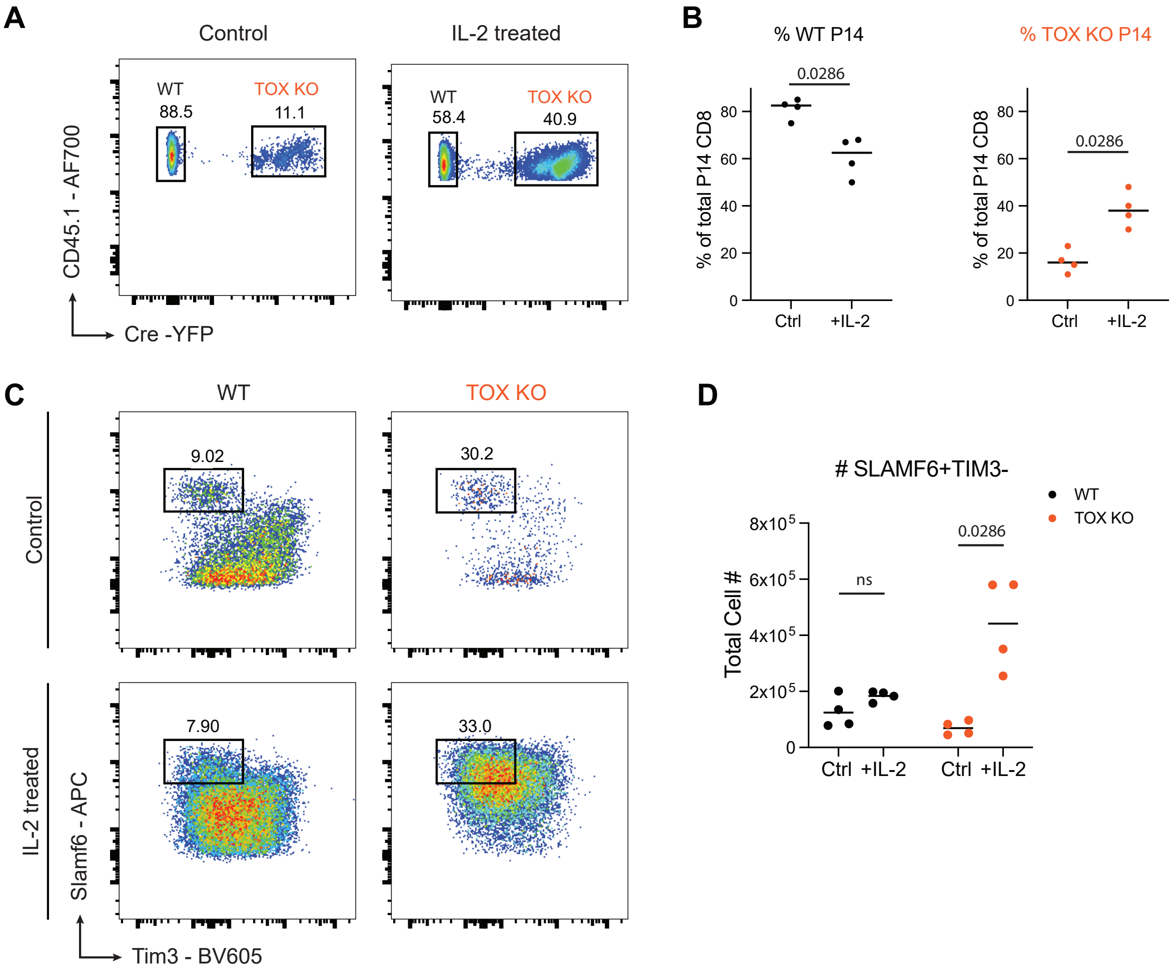

Figure 4.9.: ***In vivo* IL-2 treatment expands TOX KO progenitors** WT and TOX KO P14 CD8 T cells were transferred at 1:1 ratio into host mice and infected with LCMV clone13. 12 days post infection mice were treated with 45,000 IU of IL-2 once daily. Mice were harvested at 18 days post infection. (A-B) Ratio and frequency of WT vs TOX KO P14 in the presence or absence of IL-2 treatment. (C) SLAMF6+ TIM3- progenitors were gated on singlets/ FSC-SSC / live / CD8 / P14 / TOX KO (YFP +) or WT (YFP-). (D) SLAMF6+ TIM3- total progenitor numbers and frequency of total P14 population. Significance determined by Mann-Whitney Test. Bar represents median. Plots provided by Dr. Talyn Chu.

4.2.5. PD1-IL2v treatment expands non-exhausted progenitors in tumor environment

IL-2 treatment is actively used in therapies in a wide range of variations. Given our findings that progenitors are more heterogeneous than previously acknowledged and that IL-2 might preferentially expand progenitors with a non-exhausted phenotype, we wondered what this meant for IL-2 treatment in the context of cancer.

We turned to a recent publication that described a $\beta\gamma$-biased IL-2 variant coupled to PD-1 (PD1-IL2v), which delivers IL-2 to PD-1 expressing cells and leads to the generation of better effectors from stem-like CD8+ T cells. In this study, a subcutaneous tumor model was used to investigate the effect of treatment with FAP-IL2v, PD1, FAP-IL2v and PD1, and PD1-IL2v as conjugated compound compared to a vehicle control. FAP-IL2v is a fusion of IL2v to an antibody against fibroblast activation protein (FAP), which targets the compound to cancer-associated fibroblasts [60] and PD1 is a compound blocking the PD-1/PD-L1 interaction. (Also see the introduction part 1.1.2 for further introduction on cancer immunotherapy).

We retrieved the scRNA-seq and scTCR-seq data as well as their published clustering results (Figure 4.10A) to identify populations of progenitor and terminally differentiated CD8 T cells. Cluster 5 has uniquely high levels for *Sell*, which encodes for CD62L also known as L-selectin and is a marker for naive T cells (Figure 4.10B). Cluster 6 showed high levels of *Slamf6* and *Tcf7* and while the authors of the original paper refer to as stem-like, we will use the comparable term Tpex for the sake of consistency. We noticed that besides the Tpex population there were additional cells expressing these progenitor markers present in other clusters. Computing a gene-wise mean z-score of the expression per cluster and performing hierarchical clustering showed that besides cluster 6 (Tpex), cluster 12 and cluster 14 represent two additional albeit smaller populations with a similar phenotype (Figure 4.10C). For downstream analyses we will refer to these additional progenitor populations as alternate1 (cluster 12) and alternate2 (cluster 14). On the side of differentiated cells we considered cluster 1, 8 and 10 exhausted (following the original annotation of [61]) and labeled all remaining clusters effector T cells.

Since our previous findings revealed that IL-2 signalling affects the various progenitor populations in different ways, we leveraged the TCR sequencing data to investigate the clonal relationship between the 3 progenitor clusters and differentiated cells. For every sample we looked at the clonotypes present in each progenitor population and summed up the number of cells in the differentiated populations that shared those clonotypes. As expected, clonotype sharing is low in the vehicle control and only one clone was highly expanded (Supplemental Figure B.7A). The treatments including uncoupled PD-1 (PD1 and PD1 + FAP-IL2v) show a trend of expanded clones from the Tpex population (Figure 4.10E and Supplemental Figure B.7B).

Strikingly, treatment with PD-IL2v leads to an increase of shared clones with the effector population, confirming the conclusion of the original paper that the compound leads to better effector function [61]. What is interesting is, that the number of effectors sharing a clonotype with alternate1 and alternate2 seems higher than the sharing with cluster 6, which indicates that those clusters are populations targeted by the treatment (Figure 4.10D). However, this trend does not reach statistical significant at n=3.

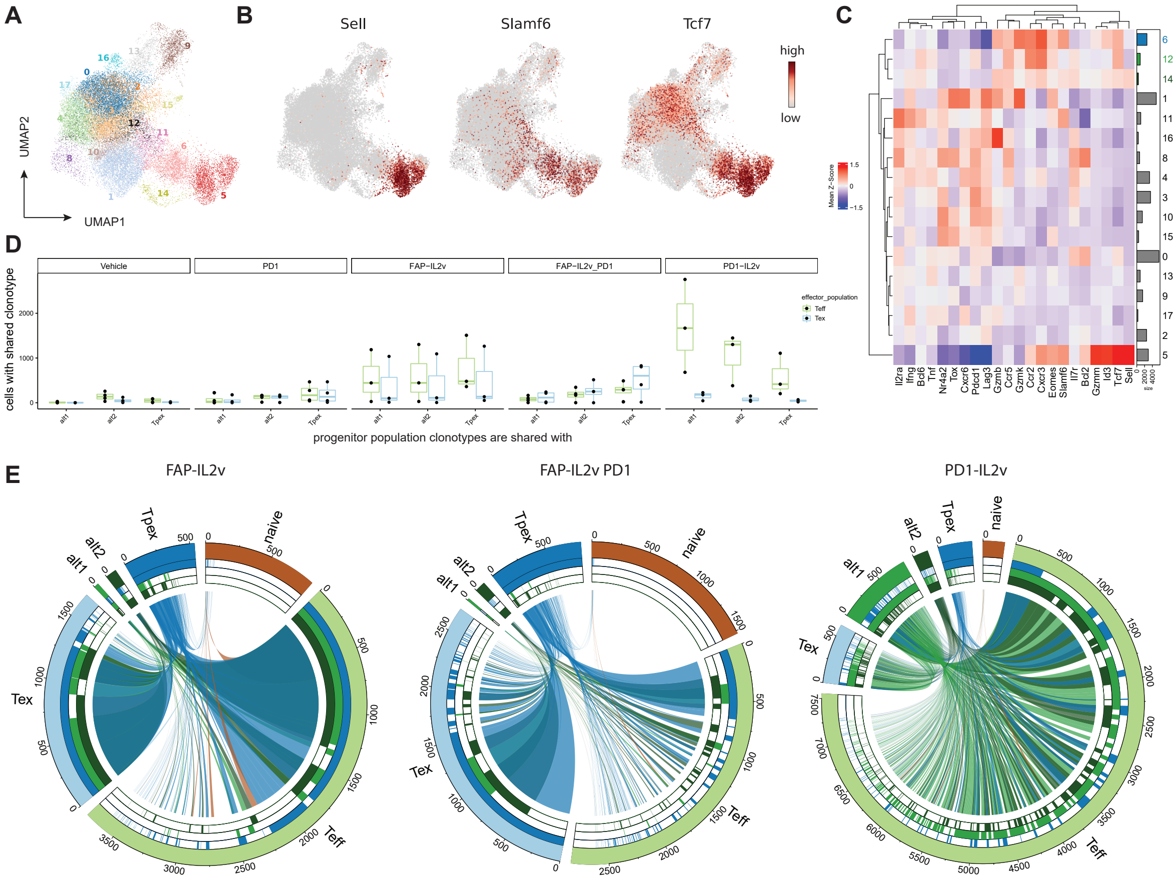

Figure 4.10.: **Clonotype sharing between progenitor populations and differentiated cells.** (A) UMAP embedding and Leiden clusters from the original publication [61]. (B) Expression of the naive marker *Sell* and progenitor markers *Slamf6* and *Tcf7*. (C) Mean group z-score of gene expression. Barplot indicates cluster size. (D) Total number of cells in the differentiated populations sharing clonotype with the indicated progenitor population. Dots represent individual mice. (E) Each cell is shown as one fragment on the circos plot, with the cluster assignment of the cell shown in the outermost ring. Cells sharing the same clonotype are connected with a link that is coloured based on the progenitor population. Additionally, whether a cells shares clonotype with a progenitor population is indicated by the three inner rings (one for each progenitor population). Cells from the same treatment condition were combined into one plot. Links between Teff and Tex are omitted to avoid overplotting. alt1 = alternate1, alt2 = alternate2, Teff = effector cells

An effector cell that has a clonotype which is shared with more than one progenitor population will be counted multiple times in the above approach. In order to avoid any potential resulting misinterpretations, we additionally visualized the clonal relationship between the populations as circos plot. In this representation each cell is a single segment on the ring and cells that share the same clonotype are connected by links (Figure 4.10E). Looking at the circos plots it becomes evident that treatment with FAP-IL2v leads to expansion of differentiated cells that are found across all progenitor clusters, whereas PD1-IL2v increases the number of alternate1 progenitors. Putting this into context with our *in vivo* studies of IL-2, it seems that when coupled to PD-1, IL2v acts similar to IL-2 treatment in that it preferentially expands alternate1 progenitor population that gives rise to effector cells which appear completely non-exhausted. This suggests the possibility that instead of converting Tpex into non-exhausted effectors, IL-2 treatment preferentially expands alternate progenitors that are non-exhausted to give rise to non-exhausted effector populations. This is significant because, while most successful immune checkpoint therapies are designed to directly target Tpex, this is not the progenitor population that the majority of effectors is actually differentiating from.

4.2.6. Conclusion

In this section, we built on our findings from section 4.1 showing that both, exhausted and non-exhausted progenitors, are generated at an early timepoint irrespective of the outcome of an infection. We investigated the maintenance of non-exhausted CD8 T cells in chronic settings and, using a combination of computational approaches, we identified IL-2 as potential ligand. Indeed, *in vivo* validations revealed that IL-2 preferentially expands non-exhausted progenitors in the P14 mouse system. Furthermore, our data supports that this finding extends to the endogenous T cell response in a tumor environment when conjugating PD-1 to IL-2v. Taken together, we found that IL-2 is a crucial environmental factor that preferentially expands and maintains non-exhausted T cells in chronic infection and tumor, which represents an exciting therapeutic avenue to potentially circumvent exhaustion.

4.3. Materials and additional methods

4.3.1. Bulk RNA-seq of PD-1hi vs PD-1lo progenitors

A P14 Tcf7yfp(bright)mCherry reporter mouse was used to enable selection of TCF1 expressing progenitor cells through fluorescence. In order to prevent rejection of the transferred mCherry expressing cells during LCMV infections, P14 TCF1 reporter cells were adaptively transferred into Vβ5 TCRβ chain only transgenic hosts. These host mice express the same TCRβ chain as OT-1 TCR transgenic mice and the presence of the transgene prevents rejection of the Tcf7yfp(bright)mCherry cells. Host mice were infected with LCMV Armstrong and TCF1 expressing P14 precursors were isolated 7 days after infection, before sorting them into PD-1hi and PD-1lo populations (Figure B.1) and performing RNA-seq.

Preprocessing of RNA-seq data was performed with a customized nf-core [178] pipeline in Nextflow (v22.04) [179]. TrimGalore (v0.6.7), which is a wrapper tools around Cutadapt (v3.4) and FastQC (v0.11.9), was utilized to remove adapter sequences and low quality base calls before aligning the reads to GRCm38 with STAR (v2.6.1d) [102]. We employed SAMtools (v1.14) [180] to sort and index BAM files and quantified reads with Salmon [104]. Further QC results were generated with Picard (v2.26.10), RSeQC (v3.0.1) [181] and QualiMap (v2.2.2)[182] and gathered in MultiQC(v1.11) [183] report.

Reads were normalized for sequencing depth with edgeR's (v3.36.0)[111] cpm function and log transformed before using them to run a PCA. Limma's (v3.50.3) voom function was used to estimate the mean-variance trend based on the design matrix $expression \sim population$. The log2 expression values from the EList object were used to visualize gene wise expression in boxplots. We fit a linear model for each gene and used empirical Bayes moderation to decrease the individual variances and squeeze them towards a common value. The resulting gene-wise coefficients for the population factor were considered significant if they passed a threshold of adjusted p-value < 0.05 and |log2FC| > 0.58.

For comparisons with gene expression differences of precursor T cells after LCMV Armstrong vs LCMV Docile infections [166] we retrieved log2FPKM values from GEO dataset GSE142687. We averaged log2FPKM values per condition, assessed the ratio of unlogged values and logged the results to get one log2FC value per gene. We selected the genes that comprise the "core exhaustion signature" described in [166] and correlated them with the log2FC of the PD-1hi vs PD-1lo progenitor comparison. The tool clusterProfiler (v4.2.2) tested enrichment of previously reported MSigDB genesets `https://www.gsea-msigdb.org/gsea/msigdb/` within our list of differentially expressed genes. We visualized the enrichment of the MSigDB geneset for acute versus chronic LCMV infection (published as part of GSE30962) on the ranked list of log2FC between PD-1 populations using limma's barcodeplot function

4.3.2. scATAC-seq

Mice were infected with either LCMV Armstrong or LCMV clone13. 7 days post infection, lymphocytes were extracted from their spleens and prepared for sequencing. Further details on how the dataset was generated can be found in [168].

Barcodes, peaks and matrix file were available via GEO accession number GSE164978. In order to produce the fragment files needed for coverage plots, we downloaded the raw fastq files using sratoolkit (v3.0.0) and ran cellranger-atac (v2.0.0) to map reads to the reference refdata-cellranger-arc-mm10- 2020-A-2.0.0.tar.gz provided by 10x. We performed downstream analyses and visualizations using R (v4.1.0), Seurat (v4.0.3) [114] and Signac (v1.3.0) [115].

We read the data into a ChromatinAssay object, selected features present in at least 10 cells and filtered the cells to those with at least 200 features. We determined the top features of the assay with a lenient threshold of at least 10 total counts to consider the feature. We ran term frequency inverse document frequency normalization, followed by singular value decomposition for dimensionality reduction. The first principal component correlated with sequencing depth, so we removed it and selected component 2 - 30 as input to create the neighborhood graph. We performed clustering using Leiden algorithm with resolution 0.25.

Previously reported ATAC-seq signatures [168] were used to help annotate the clusters. We visualized the signature scores with Seurat's AddModuleScore function which computes the average expression of a signature subtracted by aggregated control features from the same expression bins. We deployed logistic regression to find markers that are differentially accessible between Tpex and other clusters, while limiting the test to features detected in at least 10% of either population. We deemed a region differentially accessible if it passed the thresholds adjusted p-value < 0.05 and average log2FC > 0.3 and used linear proximity on the EnsDb.Mmusculus.v79 (v2.99.0) reference to annotate region to their likely target gene.

Read coverage of individual genomic regions was visualized with Signac's CoveragePlot, highlighting differentially accessible regions between Tpex and other clusters in grey. We also performed differential accessibility analysis comparing Tpex and Tmpc populations specifically in the same fashion as above with the only difference that the cutoff for significance used here was adjusted p-value < 0.05 and average log2FC > 0.15. After annotating the regions, the number of differentially accessible regions was aggregated per gene.

TFBSTools (v1.32.0) [184] and the package JASPAR2020 (v0.99.10) provided info on known transcription factor binding sites, which together with chromVAR (v1.16.0) [169] can find transcription factors associated with differential accessibility between Tpex and Tmpc. Tn5 insertion frequency around motifs of interest was visualized with Signac's Footprint function.

4.3.3. scRNA-seq and scTCR-seq of progenitors at late timepoint after acute infection

Tcf7yfp(bright)mCherry reporter mice were directly infected with LCMV Armstrong. After 4 weeks, splenocytes were collected and then stained for np396 and gp33 with tetramers. CD8 T cells that were positive for both Tetramer and the Tcf1 reporter were isolated using FACS. Following the manufacturer's protocol (CG000331 Rev E), we prepared gene expression and T-cell receptor V(D)J libraries using the Chromium Next GEM Single Cell 5' Reagent Kit v2 (PN-1000265, 10X Genomics), Chromium Single Cell Mouse TCR Amplification Kit (PN-1000254, 10X Genomics), and Chromium Next GEM Chip K Single Cell Kit (PN-1000287, 10X Genomics). For multiplexing (i7 and i5 index read, 10bp), the Dual Index Kit TT Set A (PN-1000215, 10X Genomics) was used. Sequencing was performed in a paired-end run (read

1: 26bp, read 2: 90bp) on a NovaSeq6000 platform utilizing S1 v1.5 (100 cycles) sequencing kits (20028319, Illumina). Demultiplexing and generation of .fastq files were carried out using Bcl2fastq software (v2.20.0.422).

We obtained the reference files refdata-cellranger-arc-mm10-2020-A-2.0.0.tar.gz and refdata-cellranger-vdj-GRCm38-alts-ensembl-7.0.0.tar.gz from the 10x Genomics website. These references were used to align the gene expression and TCR assay reads using cellranger multi (v7.1.0). Downstream analyses were conducted in Jupyter notebooks (v6.4.3), utilizing a combination of Python (3.9.6) and R (4.1.1) code. To facilitate sharing of data between programming languages, we made use of the python modules rpy2 (v3.4.5), anndata2ri (v1.1) and the Bioconductor package SingleCellExperiment (v1.16.0). Correction for ambient RNA within the droplets was accomplished using the R package soupX (v1.6.1) [116]. This correction process involved preliminary Leiden clusters computed through the standard scanpy (v1.8.1) [113] workflow. The corrected counts were then concatenated for the six samples, comprising two epitope stainings for each of the three mice. Subsequently, each sample underwent doublet detection using the bioconductor package scDblFinder (v1.8.0) [118], which excluded 875 cells out of the initial 16,880 cells.

We investigated the distributions of quality metrics and applied MAD thresholds to filter out low quality cells. In detail, we filtered out 363 cell if the exceeded the cutoff of 5 MAD for log1p transformed number genes detected within the cell, the percentage of counts in the top 20 genes and log1p transformed total counts. Based on the percentage of mitochondrial counts, we removed 597 cells since they had a percentage > 8% or exceeded 3 MAD. Based on the percentage of ribosomal counts, we removed 491 cells since they had a percentage < 8% or exceeded 3 MAD. Lastly, we performed feature selection by removing genes that were not found in at least 20 cells. The final matrix contained 14,952 cells and 13,956 genes after filtering.

We created a Seurat (v4.1.1) [114] object from the filtered data and performed variance stabilization by computing Pearson residuals with respect to sequencing depth and percentage of mitochondrial counts using SCTransform (v0.3.3) [121]. We carried out PCA to determine the top 10 principal components, followed by computation of the neighborhood graph. This graph representation was then embedded into a UMAP visualization and used as input for Leiden clustering with resolution set to 0.15. We identified a distinct cluster of 135 cell that showed high expression of the APC marker *Cd74* and excluded them before rerunning SCTransform to ensure this APC population would not bias the identification of most variable genes.

We converted the processed transcriptional data to an AnnData (0.7.6) object and used the module scirpy (v0.12.2) to join the TCR data to it. Cells lacking at least one complete pair of receptor sequences were excluded. The similarity between CDR3 sequences was assessed via amino acid alignment and the BLOSUM62 matrix. If the scores of both CDR3 regions (from the α and the β chain) were smaller than 10, cells were grouped together into a clonotype module. All modules containing a minimum of 50 cells were displayed in the clonotype network and barplot.

For the heatmap, data were scaled gene-wise and clonotype modules' mean z-scores were computed. Heatmap row annotations illustrate the ratio of gp33+ to np396+ cells within each clonotype module, along with a barplot showing module cell counts colored by mouse.

4.3.4. mvTCR

mvTCR was cloned from github [1] on February 3rd 2023 and executed in python(v3.8.8). The architecture of this neural network model follows the structure of a Variational Autoencoder and generates a shared embedding which lets us capture both, the TCR-seq and the scRNA-seq, data modalities in a joint UMAP representation [172]. Utilizing The top 10 principal components served as input to generate a neighborhood graph, which was in turn used to compute Leiden clusters (resolution=0.25). These Leiden clusters and the aforementioned clonotype modules were passed to mvTCR, where they were weighted 5-to-1 to create a pseudometric for optimization. The latent embedding of the trained model was then utilized to create a neighborhood graph and UMAP representation, effectively capturing both data modalities.

4.3.5. Neural network training data

Public TOX ChIP-seq data

ChIP-seq for the transcription factor TOX was performed on WT and TOX KO P14 T cells at day 8 after LCMV clone13 infection [175]. The TOX KO mice are missing exon 1 and 1.7kb upstream of the Tox gene [185].

Fastq files were downloaded from ENA with the accession numbers SRR5195618 - SRR5195620. Sequence trimming was performed with TrimGalore(v0.6.7) using cutadapt (v4.0). We aligned the reads to the GRCm38 reference using bowtie (v1.2.3). We proceeded to sort BAM files with samtools (v1.9) and remove overlap with blacklist regions [2]. Duplicates were removed with picard (v2.27.1) and peak calling was performed with MACS2(v2.2.7.1) using the narrowpeak setting and q=0.05 utilizing an input control. We generated bigwig files for the forward and reverse strand with a binsize of 1 using the deepTools(v3.5.2) function bamCoverage.

Public bulk RNA-seq of WT and TOX KO

The TOX KO model used for this study was a conditional deletion of exon 5, which has the effect of removing about two thirds of the DNA-binding domain as well as the nuclear translocation sequence [48]. Samples were generated from P14 cells isolated after LCMV clone13 infection. We determined differentially expressed genes at day 8 as done for the original publication [48] using the provided scripts [3] with slightly more relaxed cutoff of log2FC=0.58 (compared to the original log2FC=1).

[1] git@github.com:SchubertLab/mvTCR.git

[2] `mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/mm10.blacklist.bed.gz`

[3] `https://github.com/zehnlab/microarray_ngs_scripts/blob/master/code/DEG.R`

Public bulk ATAC-seq of WT and TOX KO

Analogous to the bulk RNA-seq described in the paragraph above, the TOX KO model used in this study is missing exon 5. Samples were generated from P14 cells isolated at day 8 after LCMV clone13 infection. ATAC-seq data was downloaded from ENA with the accession numbers SRR9108760-SRR9108761. TrimGalore (v0.6.5) with cutadapt(v2.6) was used to perform sequence trimming. The reads were mapped with bowtie2 (v2.4.5) to a precreated index for bowtie2 [4]. We removed reads mapping to chrM and MT as well as unlocalised and unplaced scaffolds. PCR duplicates were removed with picard (2.21.2). We generated bigwig files for the forward and reverse strand with a binsize of 1 using the deepTools(v3.5.1) function bamCoverage. We defined fragments from open chromatin as having insert size between 20 and 150 base pairs and used these reads to perform peak calling with macs2 (v2.1.1).

4.3.6. Neural network sequence models

Pretraining

We combined peaks from TOX ChIP-seq (6,978 peaks from WT and 2,320 peaks from the KO) samples and ATAC-seq samples (54,820 peaks) and merged overlapping regions to create a peak universe of a total of 55,313 peaks. Following the procedure described in BPNet [129], we created training, validation and test sets based on chromosomes. This ensures that all regions that regulate a gene in cis, are in the same partition. Namely, regions on the chromosomes 5, 6, 7, 10, 11, 12, 13, 14, 15, 16, 17, 18 and 19 were used for training, regions on the chromosomes 1, 8 and 9 were used for validation and regions on the chromosomes 2, 3 and 4 were used for testing. This resulted in 34,213 peaks in the training, 9,321 peaks in the validation and 10,409 peaks in the test set.

We performed pretraining with a PyTorch (v1.8.1) implementation of the BPNet model [129] and customized code from the kipoi (v0.8.0) codebase for datasets [186] to structure them in a way that would fit our MIL architecture later on. In brief, the model is a convolutional neural network that inputs one-hot-encoded DNA sequences to predict read count profiles at base-resolution. The first convolutional layer uses 64 filters, followed by dilated convolutional layers with 64 filters and a kernel size of 3, where the dilation rate doubles for every layer. Contrary to the original publication we decided to give the architecture more flexibility and tuned the filter size of the first layer as well as the number of dilated convolutional layers as part of the hyperparameter optimization. To ensure good gradient flow, we added skip connections and batch normalization after every dilation layer.

The output of these convolutions is a bottleneck layer that serves as input to two heads; a profile head that predicts the base-resolution profile of the track and a count head that predicts the total number of read counts aligned to the input region. Loss for the count head was defined as the mean squared error and for the profile head we used the negative log-likelihood of the observed profiles given the predicted counts and the total number of

[4] https://registry.opendata.aws/jhu-indexesonjuly-06-2022

read counts in the region. Both loss terms are combined with a weighting factor λ which is determined during hyperparameter tuning performed with optuna (v2.10.0) (Table 4.1).

We trained a total of 50 trials for maximum 5 epochs with AdamW as optimizer and pruned them based on the median correlation of the count head for the validation set.

Table 4.1.: Hyperparameter tuning for pretraining

parameter	from	to	type	stepsize	further info
kernel size of first convolution	7	27	int	4	
number of dilated conv layers	4	10	int	1	
lambda	4	20	int	2	
input sequence width	100	800	int	100	
initial learning rate	1e-5	1e-3	float		log=True

For the best performing trial, training was resumed from the last checkpoint and continued for a total of 10 epochs.

Expression head

In order to predict gene expression change, we need to bridge the gap between genomic regions and their target genes. We used the putative regulatory regions defined by ChIP-seq and ATAC-seq peaks described earlier and annotate them to their closest gene. From there, we decided to apply strategies from attention-based multiple instance learning [187]. Each region is part of a set of instances $X = \{x1, ..., xK\}$, referred to as bag containing a varying number of instances k for every gene. The bag label $y \in \{0, 1\}$, represents the gene expression change between WT and TOX KO genotype. Specifically, a bag label 1 corresponds to the gene being significantly higher in the WT samples and 0 corresponds to the gene being significantly higher in the TOX KO genotype.

To incorporate this concept into a neural network architecture, we implemented a MIL approach similar to what has successfully been applied to medical image analysis [188]. For the expression head the architecture was kept identical to the one described for pretraining up to the bottleneck layer. From there we further aggregated the bottleneck activation map (with the pooling done based on average pooling, max pooling or convolution, decided based on hyperparameter tuning). before passing them into a local attention subnetwork. This subnetwork consisted of 2 fully connected layers, separated by dropout layer and tanh activation and returned a scalar per instance. We used the softmax function to rescale the attention layers so that they would sum up to a total of 1 per bag and applied this scaled attention to compute a weighted average of bottleneck features. Whether to follow the weighted aggregation step by a batchnorm layer and what type of nonlinearity to use were part of the hyperparameters. The weighted average of the featuremap as well as the featuremaps for each instance were passed to a fully connected network to make a class prediction.

We included a hyperparameter in the tuning procedure to decide whether to simply use the bag loss for the training or whether to include the instance loss as well. In this more elaborate approach, the total loss is composed of a mix of bag loss and instance loss and gradually focuses more on the bag loss with every epoch.

$$sil\_ratio = 1 * (1 - 0.2)^{currentepoch} \tag{4.1}$$

$$loss = (1 - sil\_ratio) * bag\_loss + sil\_ratio * instance\_loss \tag{4.2}$$

Since we had fewer data points available for this task, training time was less of a constraint and we trained the models for 100 epochs. Kernel size, number of dilated convolutional layers and input sequence width were kept as in the pretraining to allow for a weight transfer. We used AdamW as optimizer to run 40 trials with various hyperparameter combinations (Supplemental Table B.1) and used the model that performed best on the validation set for model interpretation.

Feature interpretation

To determine what input sequences had a high contribution to the prediction, we used captum (v0.4.1) to compute attributions with *InputXGradient* providing bag IDs as additional forward argument

Additionaly, we used captum's *saliency* to get hypothetical contributions and then used both, together with the one hot encoded input data as input for MoDISco (v0.5.16.0) [189].

MoDISco finds high importance seqlets and clusters similar seqlets into motifs. The algorithm was executed with a sliding window size = 15 and flank size = 5.

4.3.7. Bulk RNA-seq of WT vs TOX KO progenitors

Contrary to the public bulk RNA-seq of WT vs TOX KO described for the neural network training, this dataset was sorted on the progenitor population before sequencing. In detail, antigen specific P14 T cells from WT and TOX KO (missing exon 5) were isolated and transferred into congenically marked recipient mice. The recipient mice were infected with LCMV clone13 and progenitors (CXCR5+ Tim3-) CD8 T cells were sorted at d7, d10 or d14 after infection and submitted for sequencing. Each condition and timepoint was sequenced with n = 5 with the exception of the WT phenotype at day 7 where we had an n of 4.

Preprocessing of the RNA-seq samples was done as described for bulk RNA-seq of PD-1hi vs PD-1lo progenitors.

We created a DGEList object from the raw counts and retrained genes with CPM greater 1 in at least 4 samples. The three different timepoint were treated as levels in a categorical variable and the model matrix was set up as $expression \sim genotype$ x *time*. We estimated the mean-variance trend with limma's (v3.50.3) voom function and extracted log2 expression values from the EList object for plotting. Coefficient estimation was done by fitting a linear model followed by fitting contrasts to retrieve the coefficients of interest (namely, the effect of TOX KO at each of the 3 timepoints). After running empirical Bayes moderation, we deemed genes with adjusted p-value < 0.05 and absolute log2FC > 0.58 as significant. The venn diagram shown in Figure 4.7 uses identical cutoffs.

In order to perform enrichment analysis we first created a local version of the KEGG.db using the package createKEGGdb(v.0.0.3).We then used clusterProfiler (v4.2.2) [190] to check for enrichment of the genes that are DE at day 7 within KEGG pathways. Pathview (v1.34.0) was used to visualize the enriched pathways and highlight the DE genes within it.

We utilized the NicheNet (v1.1.1) [177] workflow to investigate whether certain ligands could explain the transcriptional changes between WT and TOX KO progenitors. Prior knowledge was obtained as networks and ligand-target matrices from `https://zenodo.org/record/7074291/`. TOX KO samples of day7 were defined as the receiver population, so we assessed what receptors they express and deemed all ligands to those receptors as potential ligands for the analysis. We defined a receptor to be expressed if its log2 expression value was > 1 in all 5 samples of the condition. Genes that are DE at day 7 serve as gene set of interest and all other analyzed genes as background gene set. We predicted ligand activities with these two genesets, the provided ligand target matrix and the potential ligands defined above. Following the recommendations of the package developers, we utilize the area under the precision-recall curve (AUPR) between the observed transcriptional response and the ligand's target prediction to determine the best upstream ligands.

For the heatmap visualizing receptor expression in all our samples, we transformed the gene expression to Z-scores in a row wise fashion and reordered the genes based on a dendrogram generated using euclidean distance and complete linkage.

Promoter analysis was performed on a set of genes that had a significant coefficient for the interaction between time and genotype. For this we treated the timepoint as a continuous variable before fitting the linear model as described above. We ran *AME* from the MEME Suite (v5.4.1) in discriminative mode using the promoter sequences of the 318 significant genes with adjusted p-value < 0.05 as input and using promoter sequences as non-DE genes as control.

4.3.8. scRNA-seq and scTCR-seq data of tumors from PD1-IL2v treated mice

A Panc02-Fluc pancreatic subcutaneous tumor model was used to assess the efficacy of various intravenously delivered treatments. Treatment groups were 1.5 mg/kg muFAP-IL2v, 10 mg/kg muPD1, 0.5 mg/kg muPD-1-IL2v, 10 mg/kg muPD1 + 1.5 mg/kg muFAP-IL2v and a Vehicle control. Treatment was started when the tumor reached $200m^3$ in size and then administered twice daily. Mice were sacrificed 3 days after the second antibody treatment. Tumor tissue was isolated, transferred into a single cell suspension and enriched for viable single CD45+CD8+CD11c-CD4- cells using FACS. The scRNA-seq experiment included feature barcoding and scTCR-seq. For details on sample preparation please refer to the original publication [61].

In addition to the files downloaded through the ArrayExpress accession E-MTAB-11773, an .h5ad file containing the Anndata object including the complete AIRR data from the cellranger output was provided by the authors upon request.

Mirroring the filtering done in the original publication, we excluded the clusters 7, 18, 19, 20, 21 and 22 since they captured a variety of non-CD8 T cell populations (see [61] for details). After removing these contaminants, we had matched scRNA-seq and scTCR-seq information on 33,735 cells. Thereof 3899 were from Vehicle, 5245 from PD-1, 6929 from FAP-IL2v, 7956 from FAP-IL2v + PD1 and 9706 from PD1-IL2v treated samples. For the expression heatmap we computed the mean gene expression z-score per cluster and clustered rows and columns based on euclidean distance with complete linkage. Circos plots were generated with the circlize package v(0.4.13) [191].

5. Discussion and Outlook

5.1. GR-mediated gene expression in macrophages

Our goal was to find sequence determinants that decide the polarity of GR-mediated gene expression changes. The mechanism behind GR-mediated gene repression has long puzzled researchers and while individual aspects of it have been investigated before, we are the first to formulate the questions as a machine learning problem. On the level of individual genetic and epigenetic assays, we could confirm previously reported observations and by combining the assays we were able to identify STAT motifs as predictor for gene repression.

We looked into the genomic sequence of GR binding sites and found data suggesting that GR might interact with non-GRE sequences on the DNA, either directly or indirectly. Looking into these alternative mechanisms, we assessed simple motif occurences as a baseline model and more sophisticated regularized logistic regression models to identify sequences associated with transcriptional outcome. The factors we identified included members of the SMAD, NF-κB, C/EBP, OCT and STAT family of transcription factors. Some binding sites of other factors, like POU2F1 and STAT3, have been found in the promoters of GR-regulated genes, but their role was previously considered as coactivators [192]. While several of these factors have been reported to act in conjunction with GR for transactivation [192], our results suggest that the presence of their binding motifs is associated with gene repression. In fact, prior research has demonstrated that GR inhibits TGF-β signaling through SMAD3 [193]. Surprisingly, we found both positive and negative coefficients for AP-1 family members. This observation could be due to the method applied, where the polarity of the predictor changes based on other coefficients present in the model. Alternatively, it might indicate the complex biological relationship between GR and AP-1. On one hand, GR inhibits AP-1 target genes [194], while on the other hand, AP-1 acts as a pioneering factor for GR and enhances its binding [157].

Although the identified motifs might also predict repression due to their enrichment near genes that are activated by LPS and return to lower levels in the Dex+LPS condition, this reasoning does not explain why these motifs occur at GR binding sites. Instead, the close proximity to a GR peak summit suggests that GR is either directly binding to the motif or binding to another protein that, in turn, binds to the DNA. Recent research has highlighted the importance of DNA binding for GR-mediated suppression [24], with GR being found to bind sequences within NF-κB [195] and AP-1 [196] binding sites.

The STAT family is of particular interest due to its importance for the regulation of immune responses. When cytokines like interferons bind to their cognate receptors, STAT complexes become activated, initiating a signaling cascade that triggers the expression of inflammatory genes [197]. The association between GR and STAT3 signaling has been known for more than 25 years, originally described as a protein-protein interaction [198]. Recent research has added to this link with the discovery of a composite GR-STAT3 binding motif that shows a strong responsiveness to GR [199]. By employing ChIPexo, a higher resolution alternative to ChIP-seq, researchers have found STAT motifs at GR binding locations [26]. While some researchers argued that the absence of a GRE suggested the interaction was based on tethering [26], the low levels of active STATs observed in Dex+LPS conditions lead us to believe that GR might directly bind to STAT motifs on the DNA. After Dex treatment the available STAT is low whereas GR levels are high which suggests a possible scenario where the two factors compete for DNA binding, which leads to the suppression of inflammatory STAT target genes. However, it is important to note that further experimental studies are required to confirm this hypothesis and what is more, not all downregulated genes contain STAT motifs, indicating that this represents only one aspect of a complex repressive machinery.

Our novel machine learning method provides an unbiased approach to distinguish mechanisms involving indirect effects through chromatin changes from those involving direct GR binding. However, by using GLMs, we may overlook non-linear relationships between features and labels, such as cooperative TF-TF interactions. While more experimental studies are needed to validate our conclusion using independent assays, these results mark a promising step in comprehending how GR controls gene repression. In the future, this knowledge might benefit drug development and lead to the discovery of safer immunosuppressive treatments.

5.2. Development of exhausted progenitors in acute infection

While looking into the early drivers of exhaustion, we unexpectedly noticed the presence of TOX+ PD1hi progenitors in acute infection. Comprehensive data analysis on a combination of public and novel sequencing datasets confirmed that this population is transcriptionally and epigenetically exhausted, thereby fitting the criteria of *bona fide* progenitors of exhausted T cells (Tpex). So far, exhaustion has been described as a progressive loss of T cell effector function observed in chronic infection and cancer [46, 54, 200, 201] and we are the first to report Tpex in acute infection.

Our multi-OMICs data integration of single cell transcriptome and immune repertoire sequencing data uncovered a link between Tpex formation and TCR binding affinity. The resulting theory that, mechanistically, Tpex formation is driven by T cell activation strength, has since been confirmed by our collaborators using two additional experimental models (data not shown in this thesis). In the context of cancer, the observation that strong TCR activation leads to an exhaustion phenotype has been made before [170], but has been overlooked in the context of acute infection [202].

We leveraged the resolution of single-cell sequencing methods to show that Tpex form during the early stage of infection, before it is evident whether the infection resolves or becomes chronic. Experimental validations showed that while these progenitors present a small population at later timepoints, they are initially found at roughly equal numbers compared to typical memory precursor T cells. This proves that the organism initially forms a heterogeneous progenitor repertoire that can subsequently by shaped and steered based on the environment. Intuitively, such an adaptability brings advantages for the infected host organism, as it prepares the specialized Tpex population in the eventuality that the infection will turn chronic. This theory is in accordance with the finding that, in the context of chronic antigen exposure, T cell exhaustion is a beneficial mechanism preventing immunopathologies and is hence a functional adaptation rather than a defect [48, 203]. Further research is needed to investigate why their numbers decline between early and later timepoints.

Taken together, we report a previously unappreciated heterogeneity in the progenitor populations following infection. Exhausted T cells are formed independent of the outcome of the infection and chronic infections merely maintain and expand the preemptively formed Tpex. This observation challenges the current paradigm that exhausted T cells are exclusive to chronic infection and cancer [46, 54, 174, 200, 201, 204] and prompts us to reassess previous reports in this new light. Expanding on the finding that a heterogeneous progenitor pool is formed independently of the infection outcome brings up the idea that there might be non-exhausted progenitors in a chronic antigen environment, that could be therapeutically targeted in cancer.

5.3. Maintenance of non-exhausted progenitors in chronic infection and cancer

The finding from section 4.1 that a heterogeneous progenitor pool is created early on, irrespective of the outcome of the infection, made us investigate the environmental factors driving their maintenance. We designed a multiple-instance learning neural network architecture to link DNA sequence with gene expression outcome and found that the transcriptional changes in non-exhausted progenitors are associated with KLF binding sequences in regulatory regions of the target genes. Differential expression and ligand analyses revealed that non-exhausted progenitors exhibit changes in the IL-2 pathway activity and express higher levels of CD25, i.e. the high affinity α chain of the IL-2 receptor. This aligns with our results from the neural network approach as a link between KLF and IL-2 signalling has been suggested previously [205]. *In-vivo* validations confirmed that IL-2 treatment shifts the ratio in favor of non-exhausted progenitors, highlighting their dependence on IL-2 to expand. While the importance of IL-2 for expansion and maintenance of various T cell populations has been well established [206, 207, 208], its potential to expand non-exhausted cells in chronic infection is novel.

We applied these new insights to shed light on the results of PD-1 and IL-2 treatment strategies in cancer. We found that treatments involving unconjugated PD-1 act on Tex and Tpex populations which is consistent with reports that PD-1 can reinvigorate exhausted T cells but not permanently reprogram them [63]. Looking into the effect of PD1-IL2v, we observed that the compound increased functional effectors by expanding an alternative progenitor population rather than Tpex. Previous research on combination treatment of PD-1 and IL-2 in infection models acknowledged the importance of CD25 expression on the targeted progenitors and showed that blocking CD25 signalling lead to a loss of the synergistic effect [50], but failed to appreciate the heterogeneity of progenitors and instead assumed conversion from Tpex to non-exhausted effector cells. Deducting developmental relationships from a single timepoint is challenging. We can tell which cells developed from the same clone, but lack information on the phenotype of that clone at the time of the developmental branching. Sampling before and after treatment in combination with scRNA-seq and scTCR-seq could provide further evidence to test our hypothesis that PD1-IL2v expands non-exhausted progenitors rather than converting exhausted progenitors. A limitation of our studies is that our conclusions are based on mouse models and further validations are required to determine whether the same observations hold true in the human system.

It may seem surprising that the PD1-IL2v compound acts on CD25 expressing cells despite the IL-2 mutations at the interface with CD25 (see section 1.1.2 for further details on IL-2). The IL-2 variant abrogates the advantage that the CD25 brings when competing for low concentrations of IL-2 but does not prevent binding to the trimeric receptor [59], especially in clinical settings when the administered levels highly exceed the endogenous ones [57]. The conjugate compound PD1-IL2v is more efficient at expanding functional effector cells than the unconjugated combination therapy of FAP-IL2v and PD1 or the FAP-IL2v monotherapy. We speculate that the PD-1 expressed by activated non-exhausted progenitors can retain the IL-2v in close proximity to the receptor and hence increase its signalling capacity. Furthermore, it has been shown that PD1-IL2v is internalized slower than FAP-IL2v [61] which results in longer half-life before receptor mediated clearance and could explain its increased effectiveness.

The presence of non-exhausted progenitors in a chronic environment and the finding that IL-2 preferentially expands them has major implications for T cell based cancer immunotherapies. It opens the door to avoiding exhaustion in CAR T cells by introducing a TOX KO through gene editing and then maintaining their numbers *in vivo* with IL-2 administration. In the same vein, exhaustion within the host in settings such as cancer can potentially be circumvented by shaping the immune response with environmental factors such as IL-2 that preferentially expand acute progenitors. Lastly, so far bi-specific drugs have been designed with the objective to target exhausted progenitors, hence the population of non-exhausted progenitors represents an auspicious target for novel therapeutics.

5.4. Conclusion

In this thesis we successfully designed customized data analysis strategies that were tailored to the scientific questions at hand in order to gain new insights on the immune system. Looking back at the scope of the thesis discussed in section 1.4, we can assuredly say that we reached the set goals.

Within the innate immune system we investigated macrophages and the glucocorticoids' mode of action. We devised a machine learning workflow that incorporates assay from different parts of the transcriptional regulation and identifies STAT motifs as novel predictor for GR-mediated gene repression.

Within the adaptive immune system, we used single cell sequencing assays, to investigate the development of early progenitors. We discovered previously unappreciated heterogeneity within early CD8 T cell progenitors and found signals promoting the formation of exhausted T cells. We built on this discovery and performed multi-OMICs data integration and successfully uncovered factors that preferentially expand non-exhausted T cells in a chronic environment.

A. Supplements for Chapter 3

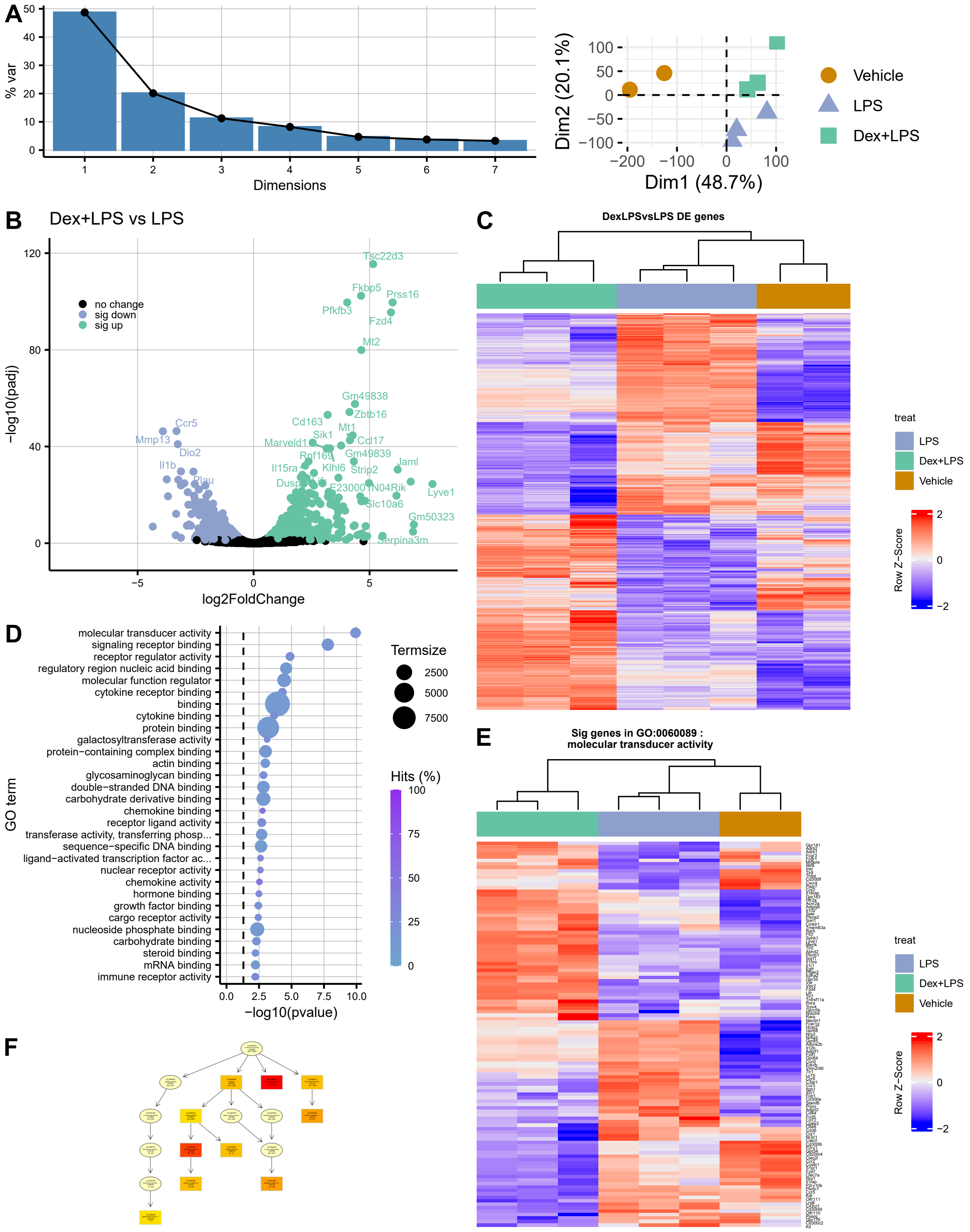

Figure A.1.: See legend on next page.

Figure A.1 *(previous page)*: **Differential gene expression analysis.** (A) Principal component analysis (left) Elbow plot showing % of total variance explained by top principal components (right) samples projected onto space by PC1 and PC2. (B) Volcano plot with genes significantly higher (adjusted p-value <0.05 log2FC > 0.58) expressed in the Dex+LPS and LPS condition highlighted in green and blue, respectively. (C) Heatmap with genes differentially expressed between the Dex+LPS and the LPS condition. (D) GO term enrichment analysis for the differentially expressed genes. (E) Heatmap with the differentially expressed genes from the category "Molecular transducer activity" (F) GO hierarchical tree of the most enriched categories. Figure and legend taken from [2].

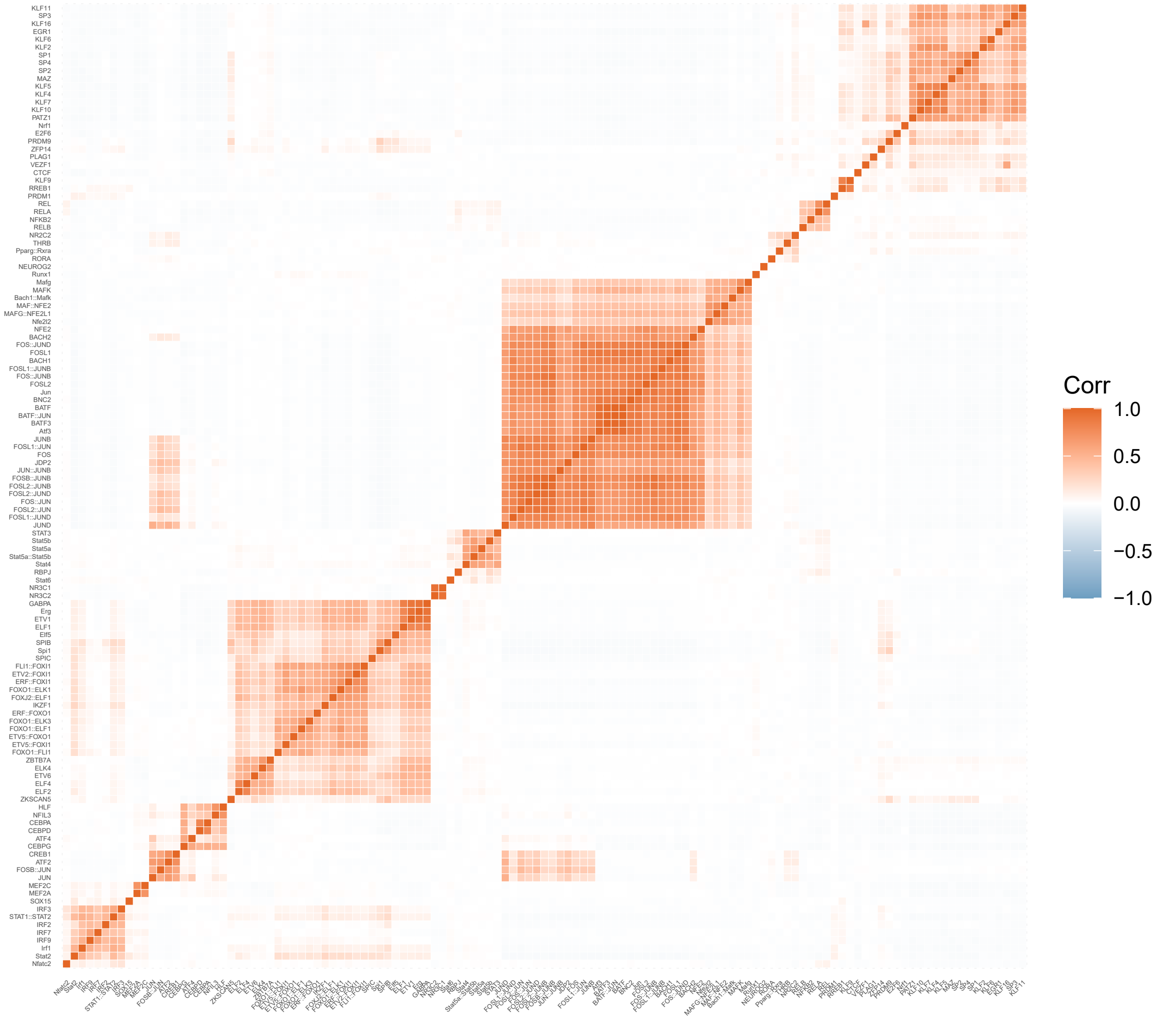

Figure A.2.: **Motifcorrelations.** Pearson correlation of counts of 30% most variable motifs at 100bp around GR ChIP-seq peak summits. Figure and legend taken from [2].

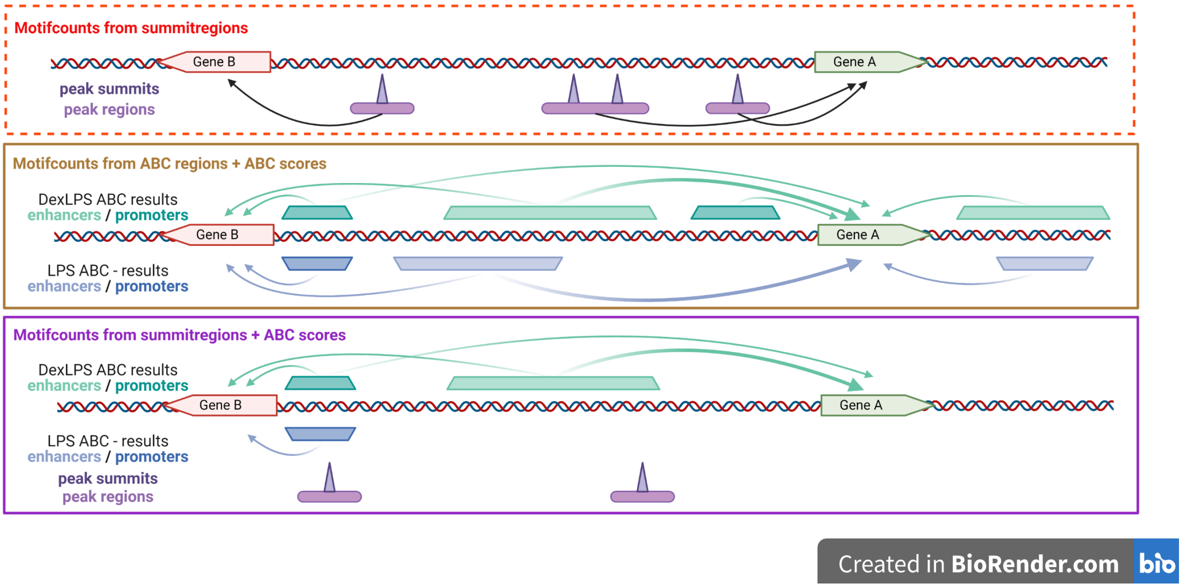

Figure A.3.: **Visualization of different approaches of GLM feature generation.** Motif matches within the respective regions are used as input for the feature engineering process. Top row shows assignment of GR ChIP-seq peak regions to the closest gene in a 1-to-1 mapping used by the reference model. Middle row shows active regions identified with the ABC workflow and corresponding region-gene assignments. The same region can have assignments to more than one gene. ABC scores (represented as line thickness in the assignment) can be used as weight during gene-level aggregation. Enhancer regions can be genic or intergenic. The bottom row represents the hybrid approach where motif counts within GR summitregions are combined with assignments from the ABC workflow. Only the subset of GR ChIP-seq regions with peaks within an active region are used for feature engineering. Figure and legend taken from [2].

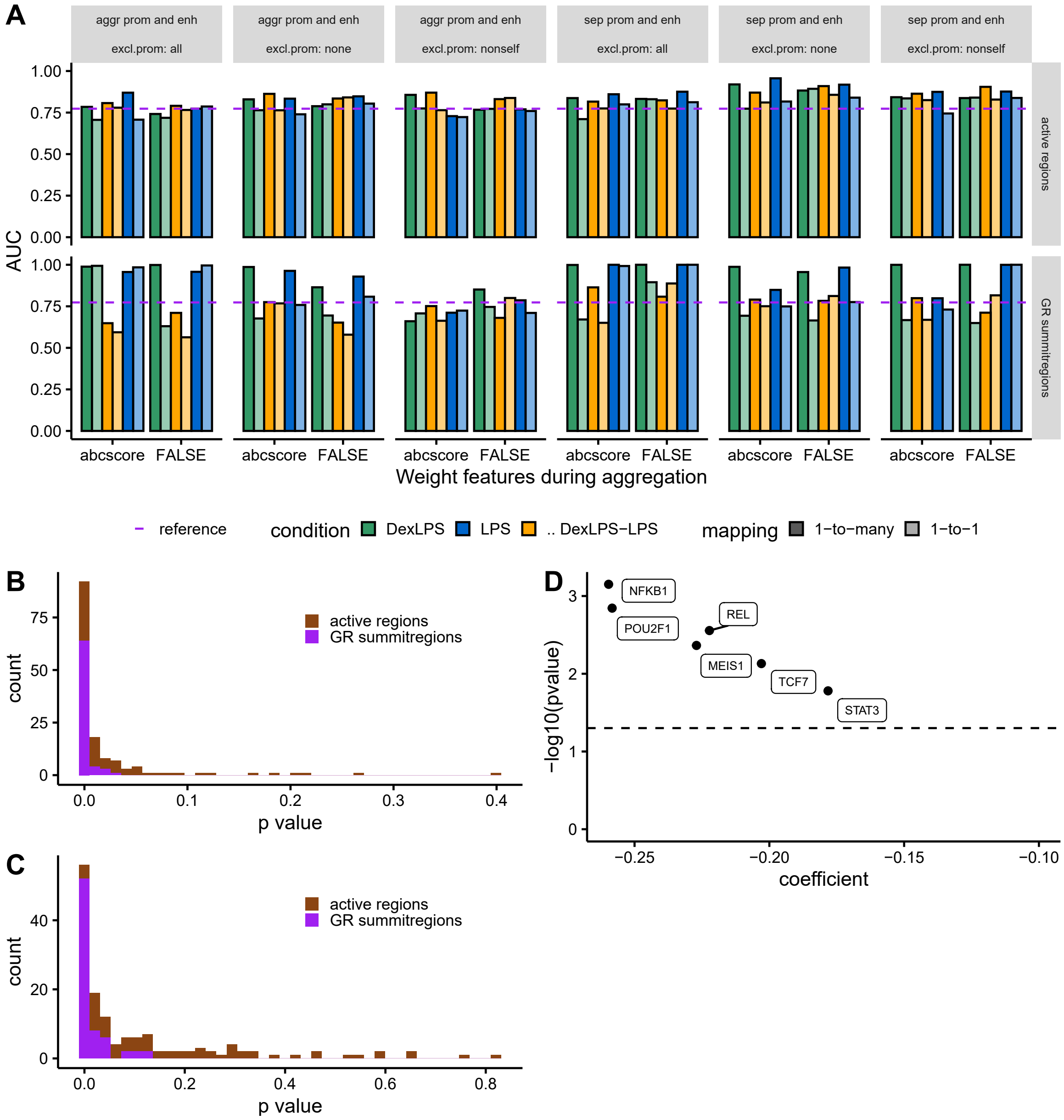

Figure A.4.: **Follow-up analyses to GLM results.** (A) Model performance on the training set. (B+C) Testing differences in model performance through DeLong's method, followed by Storey's method to assess the number of true null hypotheses. (B) Comparing performance of the best performing model pairwise with all other models. (C) Comparing performance of reference model pairwise with all other models. (D) Bivariate analyses of potential repressors identified in the reference model predicting expression change to confirm polarity and compare magnitude of coefficients. Figure and legend taken from [2].

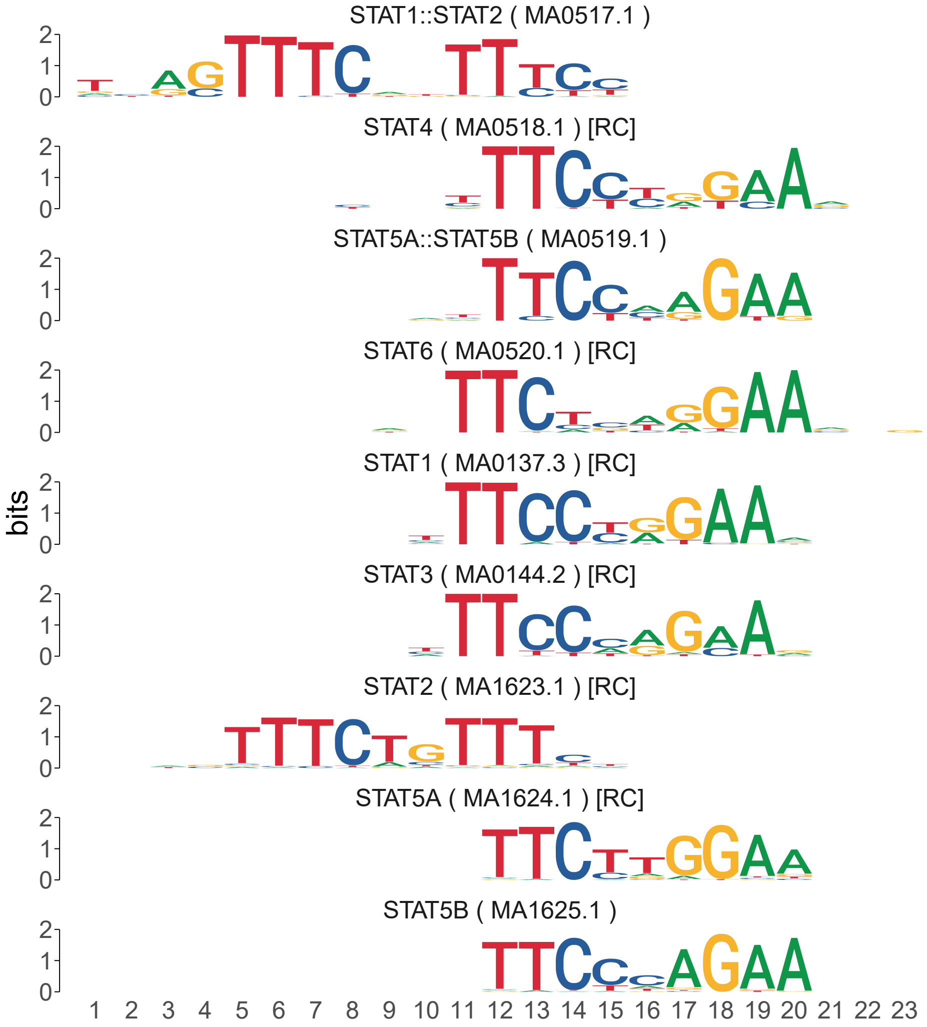

Figure A.5.: **STAT motifs.** Motifs of STAT family members retrieved from JASPARdb. [RC] indicates that the depicted motif is the reverse complement of the one deposited in the database. Figure and legend taken from [2].

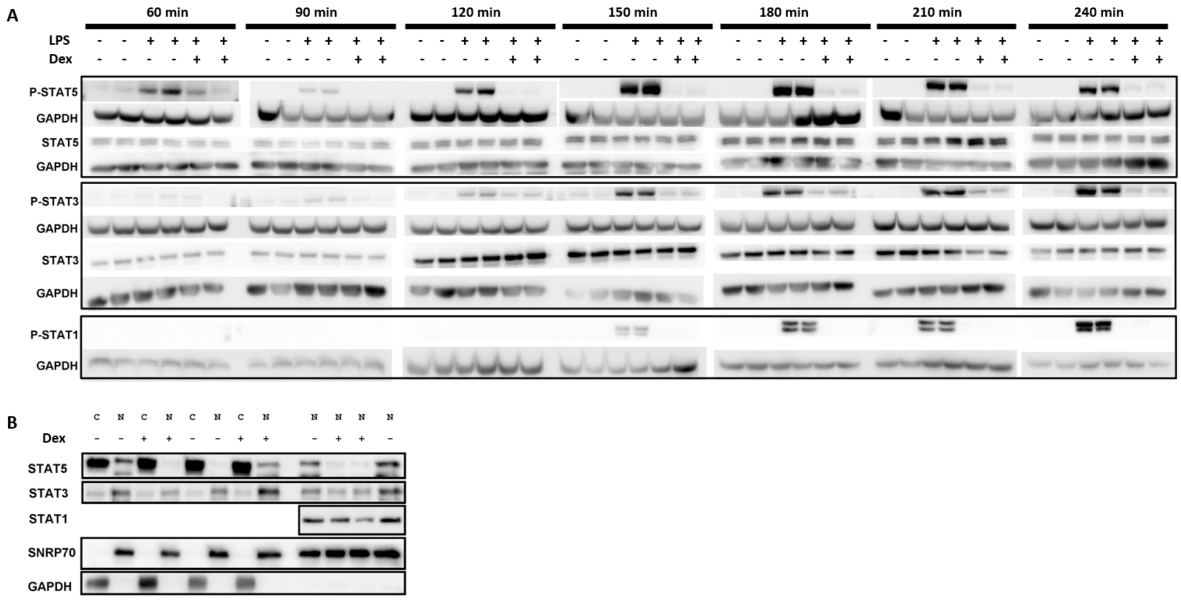

Figure A.6.: **Western blots showing STAT activity and localization.** (A) Phosphorylation of STAT proteins in BMDMs treated with LPS, Dex+LPS or untreated controls. (B) STAT5 and STAT3 levels in BMDMs split into nuclear (N) and cytoplasmic (C) fraction. Figure and legend taken from [2].

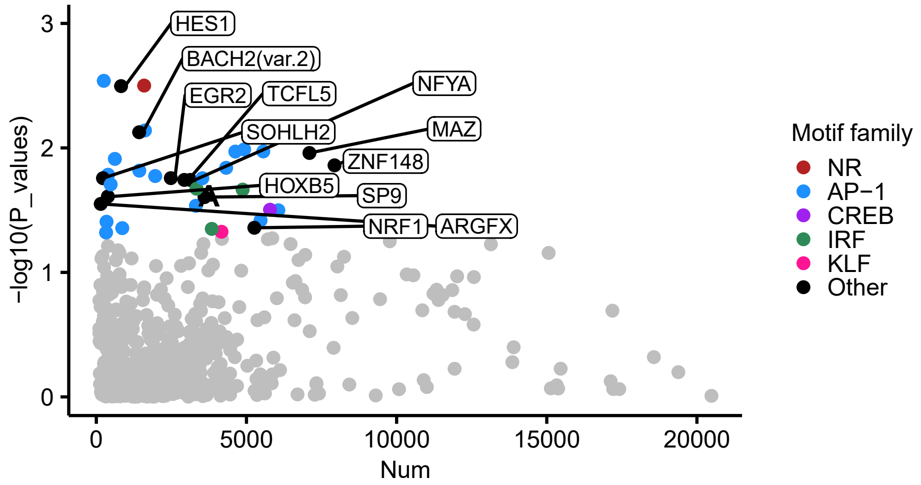

Figure A.7.: **Footprinting.** Differential footprinting results from comparing Dex+LPS and LPS ATACseq data. Figure and legend taken from [2].

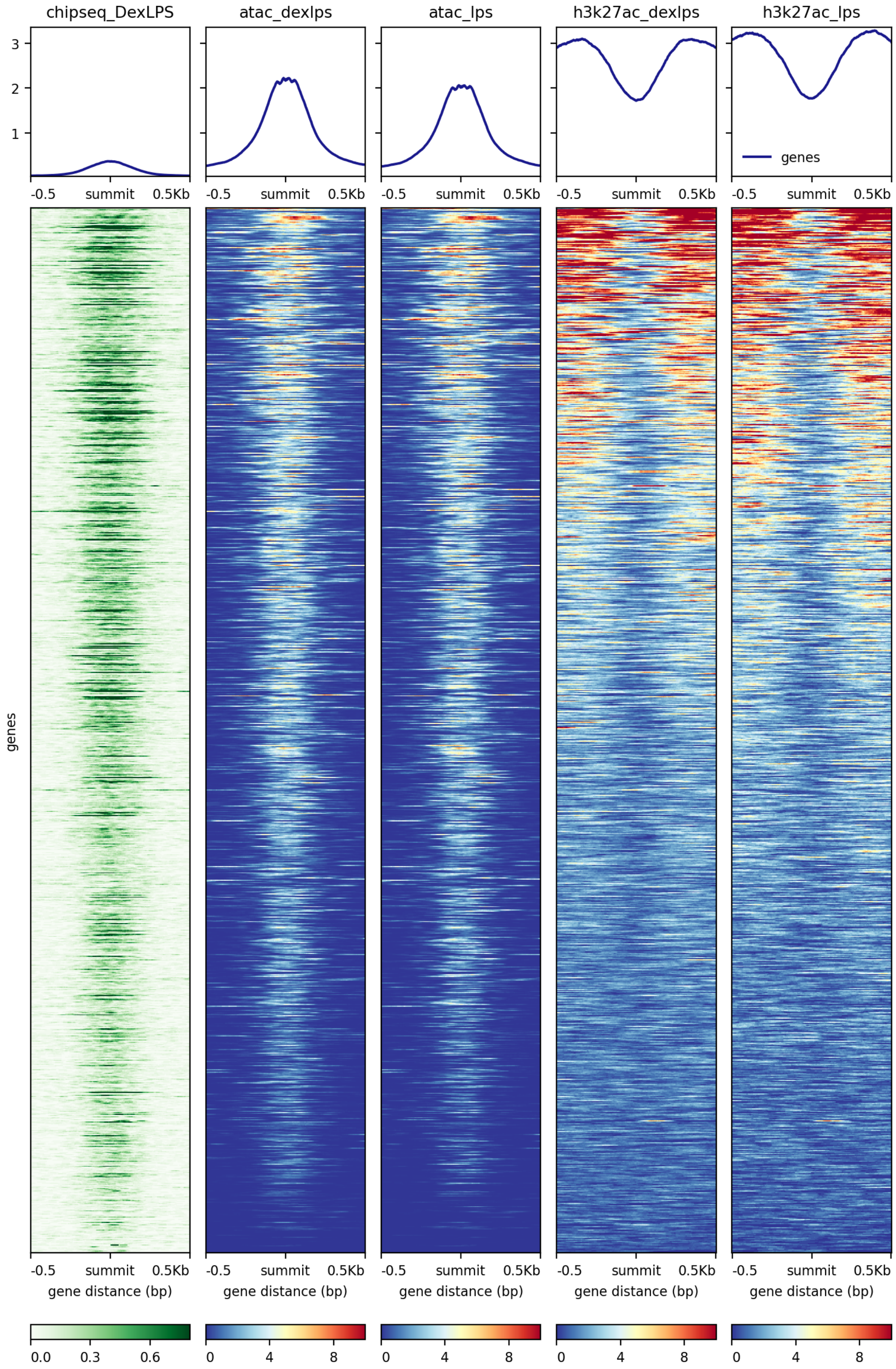

Figure A.8.: **Bigwig tracks at 100bp summitregions with STAT3 hit.** Figure and legend taken from [2].

B. Supplements for Chapter 4

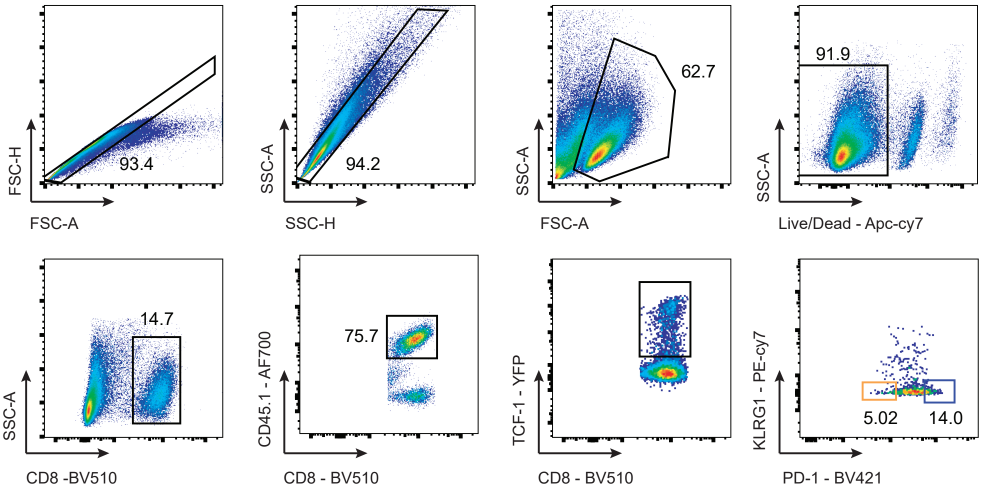

Figure B.1.: **Gating strategy for PD1hi and PD1lo populations.** TCF1+ progenitor populations high or low for PD-1 were submitted for bulk RNA-seq. Figure created by Dr. Talyn Chu.

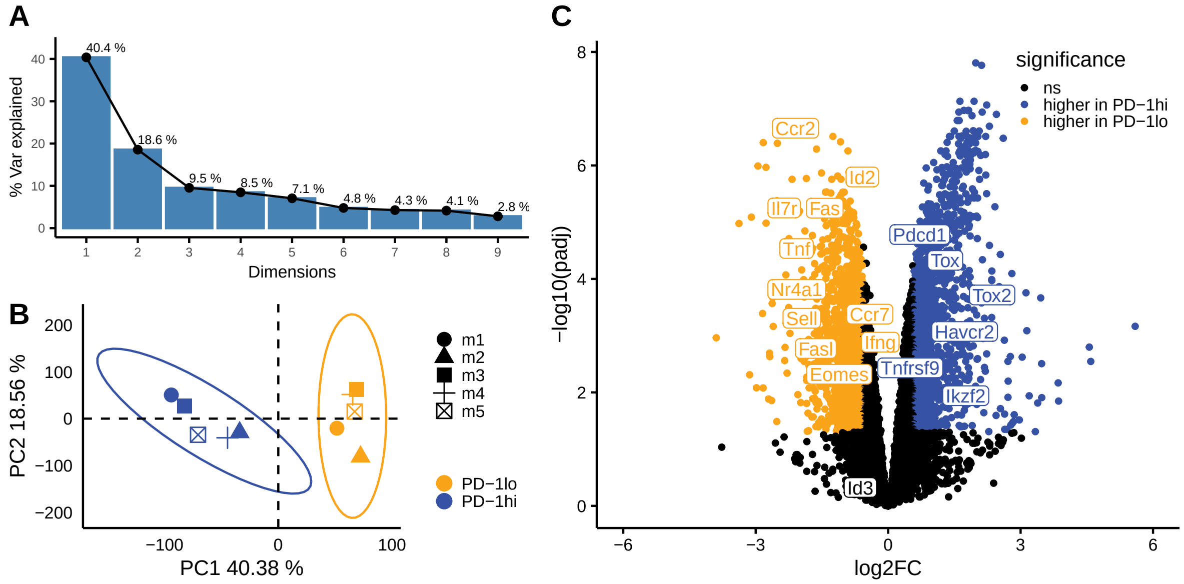

Figure B.2.: **Transcriptional differences between PD1hi and PD1lo progenitors.** (A) Elbow plot of most important principal components. (B) Samples projected into PC space showing PC1 by PC2. (C) Volcano plot of differential expression between PD1hi and PD1lo samples. Genes passing significance threshold of adj. p-value < 0.05 and absolute log2FC > 0.58 are coloured in blue and yellow.

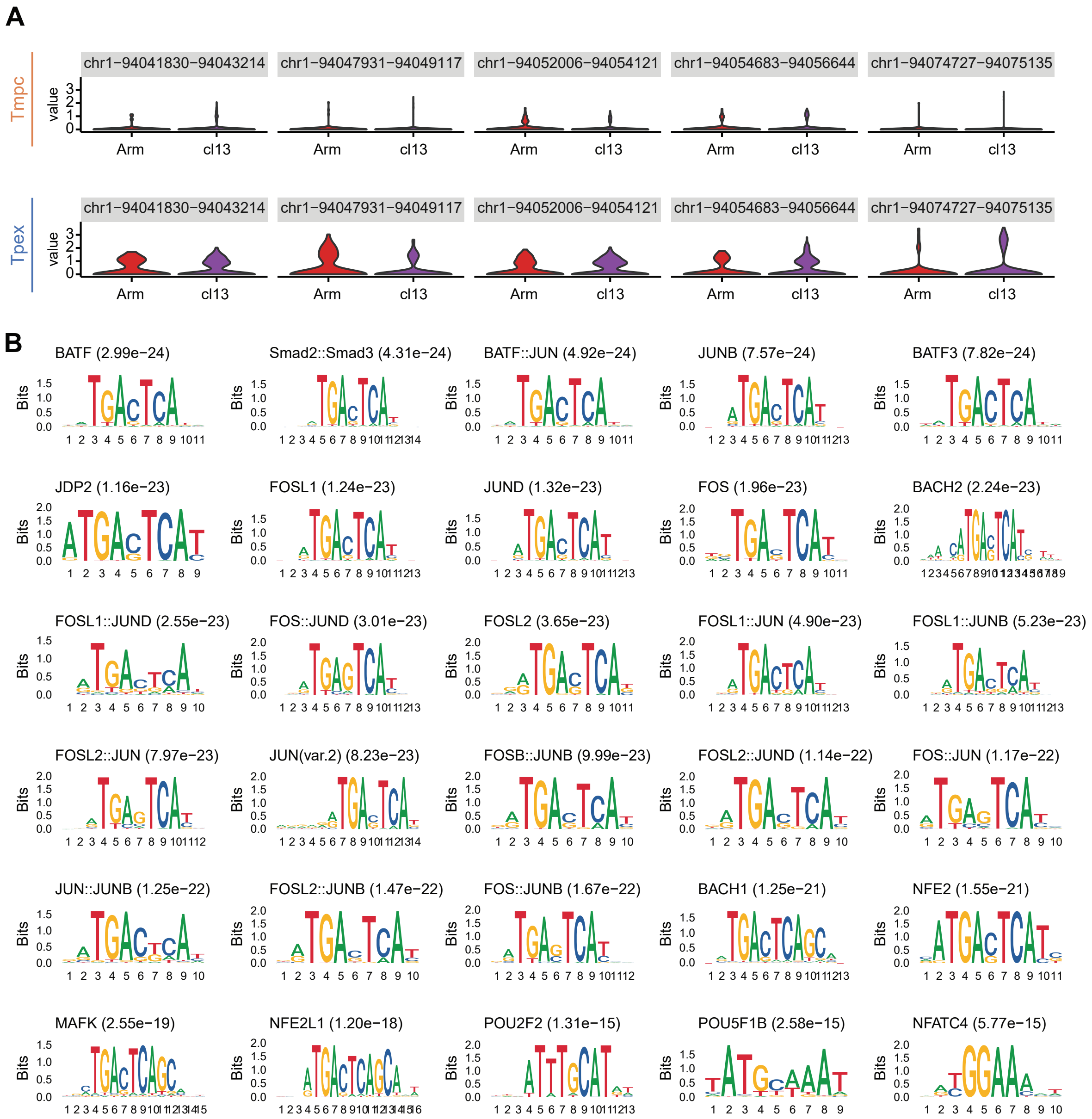

Figure B.3.: **Accessibility comparison of regions proximal to the *Pdcd1* locus and motif analysis.** (A) Accessibility levels of Tpex marker regions at *Pdcd1* locus directly comparing Tpex and Tmpc clusters of Armstrong and clone13 infections. (B) Top 30 most differentially active transcription factors comparing Tmpc and Tpex clusters of Armstrong infection.

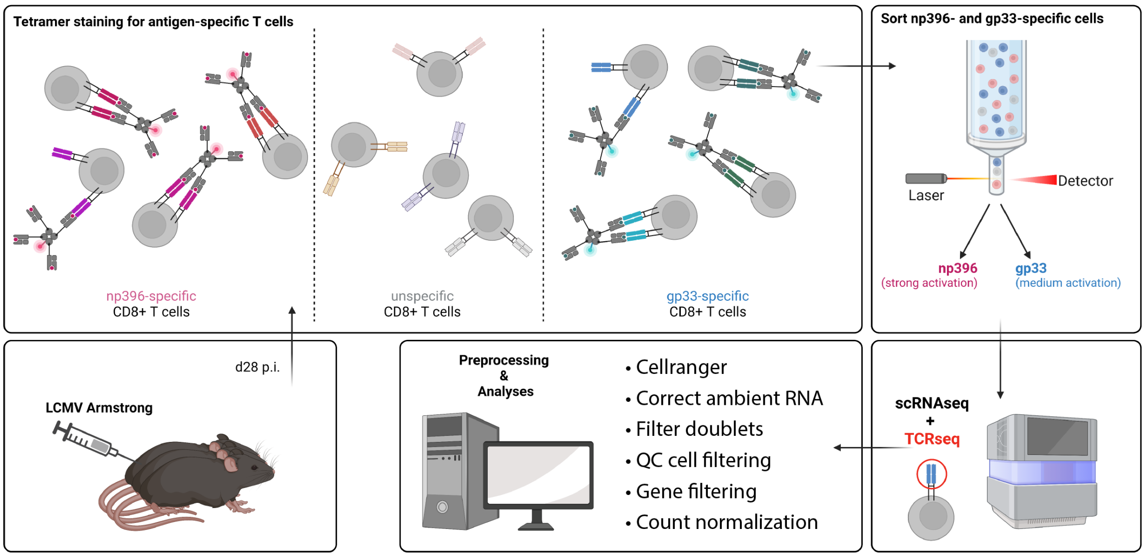

Figure B.4.: Schematic for scRNA-seq and scTCR-seq data generation of epitope-specific CD8+ T cells at late timepoint.

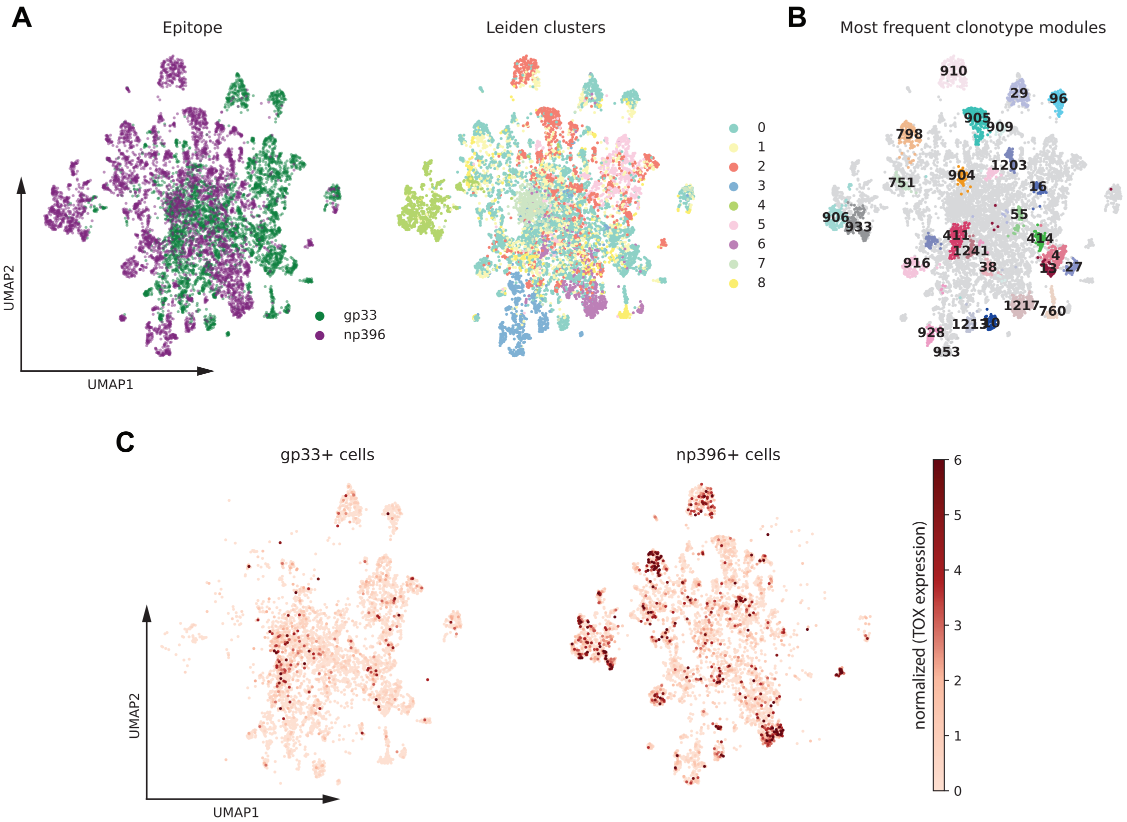

Figure B.5.: **np396 specific T cells express higher levels of exhaustion markers.** gp33 and np396-specific TCF1+ progenitors were purified from Tcf1 reporter mice infected with LCMV 4 weeks post infection. Embedding of transcriptome and TCR-seq data yields joint representation of both data modalities. Results shown for all mice (n=3). (A) UMAP colored by epitope (left) and mouse ID (right). (C) UMAP with most frequent clonotype modules (size >=50) highlighted. (D) Normalized Tox expression shown separately for np396+ and gp33+ cells.

Table B.1.: Hyperparameter tuning for expression prediction

parameter	range or categories	stepsize	further info
pooling approach	avg pool, max pool, conv		
n channels of conv output	4 - 16	1	only relevant if pooling is conv
n FC layers in expression head	1 - 2	1	
dropout rate	0.1 - 0.5	0.05	
loss	bag loss, mix of bag and instance loss		
freeze weights	none, only batchnorm, all		"all" did not include new layers
hidden nodes in attention network	20 - 40	5	
n attention heads	1 - 10	1	
batchnorm after weighted sum	yes or no		
act. function after weighted sum	tanh, leakyrelu		
init learning rate	$5e^{-5}$ - 5^{e-3}		log = True

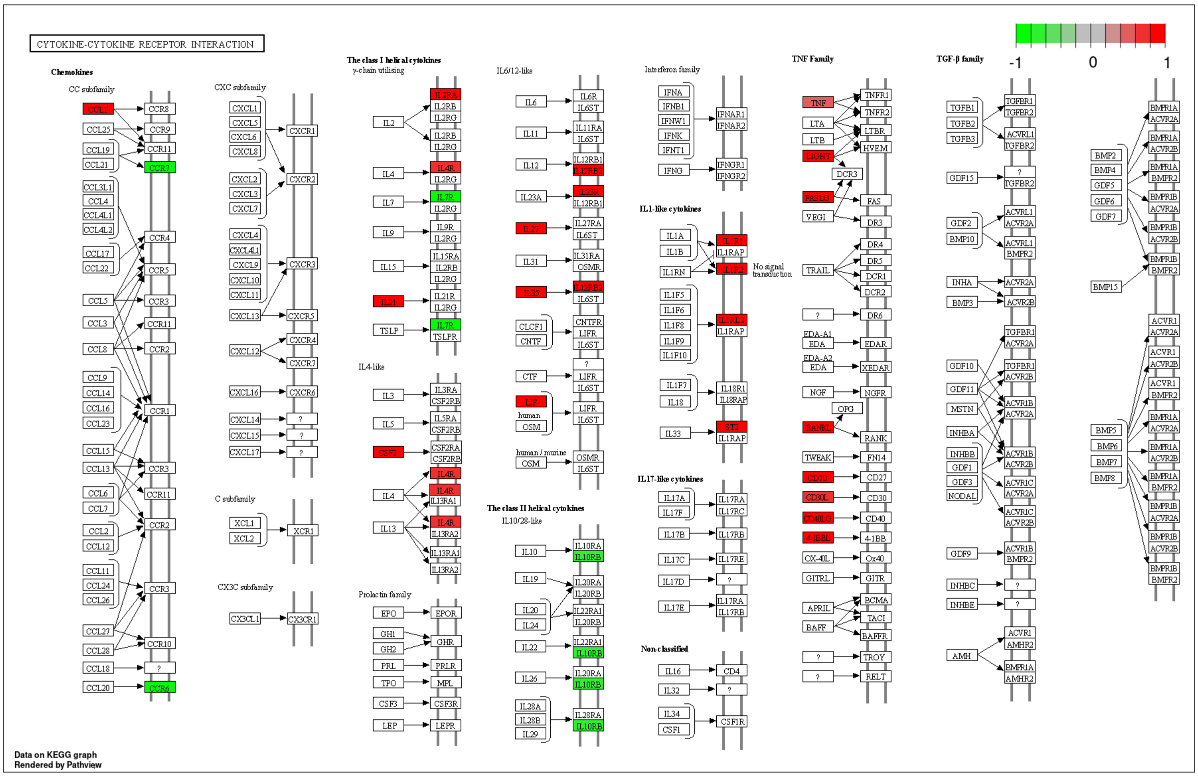

Figure B.6.: **Cytokine-cytokine receptor interaction KEGG pathway.** Genes differentially expressed at day 7 are coloured based on their log2FC comparing ToxKO to WT progenitors with genes higher expressed in WT shown in green and displayed on KEGG pathway mmu04060.

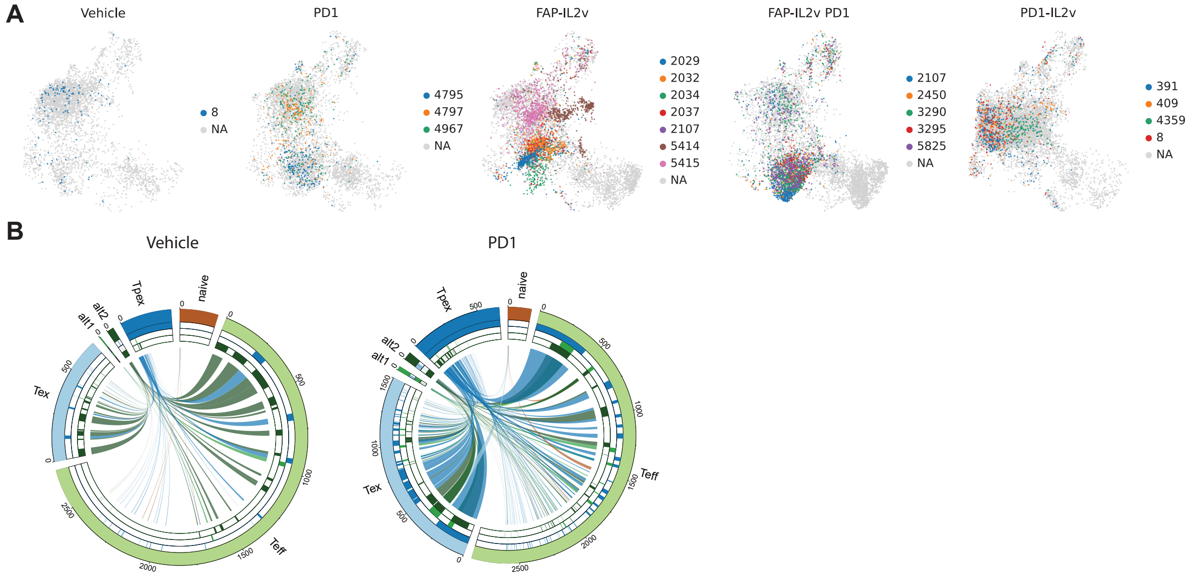

Figure B.7.: **Most abundant clones by condition.** (A) Cells coloured by clonotype for those with clone size of at least 200. (B) Circos plots as shown in Figure 4.10.

List of Figures

List of Tables

Abbreviations

4sU-seq 4-thiouridine labelling followed by sequencing.

AP-1 activator protein 1.

APC antigen presenting cell.

ATAC-seq Assay for Transposase-Accessible Chromatin using sequencing.

BCR B cell receptor.

BMDM bone-marrow derived macrophage.

CAGE-seq Cap-analysis gene expression followed by sequencing.

CAR chimeric antigen receptors.

ChIP-seq chromatin immunoprecipitation followed by sequencing.

CITE-seq Cellular Indexing of Transcriptomes and Epitopes by sequencing.

CNN convolutional neural network.

DA differentially accessible.

DEGs differentially expressed genes.

Dex Dexamethasone.

FACS Fluorescence-activated Cell Sorting.

FDR false discovery rate.

FWER family-wise error rate.

GLM Generalized Linear Model.

GR glucocorticoidreceptor.

GRE GR response element.

IDR irreproducible discovery rate.

IFN interferon.

IL interleukin.

LCMV *lymphocytic choriomeningitis virus*.

LPS Lipopolysaccharide.

MAD median absolute deviations.

MIL multiple instance learning.

MLE maximum likelihood estimation.

mRNA messenger RNA.

NF-κB nuclear factor-kappa B.

NGS next-generation sequencing.

PCA principal Component Analysis.

RNA-seq ribonucleic acid sequencing.

scATAC-seq single-cell ATAC-seq.

scRNA-seq single-cell RNA-seq.

scTCR-seq single-cell TCR-seq.

STAT signal transducer and activator of transcription.

TCR T cell receptor.

TCR-seq TCR sequencing.

TF transcription factor.

Tmpc memory precursor T cells.

TNF tumor necrosis factor.

TOX thymocyte selection-associated high mobility group box.

Tpex progenitors of exhausted T cells.

TSS transcription start site.

UMAP Uniform Manifold Approximation and Projection.

Bibliography

[1] B. Höllbacher, K. Balázs, M. Heinig, and N. H. Uhlenhaut. “Seq-ing answers: Current data integration approaches to uncover mechanisms of transcriptional regulation”. In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 1330–1341. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2020.05.018.

[2] B. Höllbacher, B. Strickland, F. Greulich, N. H. Uhlenhaut, and M. Heinig. “Machine learning reveals STAT motifs as predictors for GR-mediated gene repression”. In: *Computational and Structural Biotechnology Journal* (2023). ISSN: 2001-0370. DOI: 10.1016/j.csbj.2023.02.015.

[3] M. Wu, B. Hoellbacher, C. Wurmser, J. Berner, L. Donhauser, L. Bongers, F. Toppeta, P. Strobl, M. Heinig, T. Chu, and D. Zehn. “Precursors of exhausted T-cells are preemptively formed regardless of the outcome of infection”. In: *Under submission* (2023).

[4] K. T. Schmid, B. Höllbacher, C. Cruceanu, A. Böttcher, H. Lickert, E. B. Binder, F. J. Theis, and M. Heinig. “scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies”. In: *Nature Communications* 12.1 (2021), p. 6625. ISSN: 2041-1723. DOI: 10.1038/s41467-021-26779-7.

[5] A. K. Abbas, A. H. Lichtman, and S. Pillai. *Cellular and Molecular Immunology*. Eighth. Philadelphia, PA: Elsevier/Saunders, 2015. ISBN: 978-0-323-22275-4.

[6] R. Medzhitov and C. Janeway. “Innate immunity”. In: *The New England Journal of Medicine* 343.5 (2000), pp. 338–344. ISSN: 0028-4793. DOI: 10.1056/NEJM200008033430506.

[7] E. Mass, F. Nimmerjahn, K. Kierdorf, and A. Schlitzer. “Tissue-specific macrophages: how they develop and choreograph tissue biology”. In: *Nature Reviews. Immunology* (2023), pp. 1–17. ISSN: 1474-1741. DOI: 10.1038/s41577-023-00848-y.

[8] C. A. Janeway. “Approaching the Asymptote? Evolution and Revolution in Immunology”. In: *Cold Spring Harbor Symposia on Quantitative Biology* 54 (1989), pp. 1–13. ISSN: 0091-7451, 1943-4456. DOI: 10.1101/SQB.1989.054.01.003.

[9] C. A. Janeway and R. Medzhitov. “Innate immune recognition”. In: *Annual Review of Immunology* 20 (2002), pp. 197–216. ISSN: 0732-0582. DOI: 10.1146/annurev.immunol.20.083001.084359.

[10] K. Newton and V. M. Dixit. “Signaling in innate immunity and inflammation”. In: *Cold Spring Harbor Perspectives in Biology* 4.3 (2012), a006049. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a006049.

[11] RECOVERY Collaborative Group, P. Horby, W. S. Lim, et al. “Dexamethasone in Hospitalized Patients with Covid-19”. In: *The New England Journal of Medicine* 384.8 (2021), pp. 693–704. ISSN: 1533-4406. DOI: `10.1056/NEJMoa2021436`.

[12] C. Hua, F. Buttgereit, and B. Combe. “Glucocorticoids in rheumatoid arthritis: current status and future studies”. In: *RMD Open* 6.1 (2020), e000536. ISSN: 2056-5933. DOI: `10.1136/rmdopen-2017-000536`.

[13] P. J. Barnes. “Corticosteroids: The drugs to beat”. In: *European Journal of Pharmacology*. The Pharmacology of the Respiratory Tract 533.1 (2006), pp. 2–14. ISSN: 0014-2999. DOI: `10.1016/j.ejphar.2005.12.052`.

[14] J. M. Ehrchen, J. Roth, and K. Barczyk-Kahlert. “More Than Suppression: Glucocorticoid Action on Monocytes and Macrophages”. In: *Frontiers in Immunology* 10 (2019). ISSN: 1664-3224.

[15] Z. Werb, R. Foley, and A. Munck. “Interaction of glucocorticoids with macrophages. Identification of glucocorticoid receptors in monocytes and macrophages”. In: *The Journal of experimental medicine* 147.6 (1978). ISSN: 0022-1007. DOI: `10.1084/jem.147.6.1684`.

[16] A. Vegiopoulos and S. Herzig. “Glucocorticoids, metabolism and metabolic diseases”. In: *Molecular and Cellular Endocrinology* 275.1-2 (2007), pp. 43–61. ISSN: 0303-7207. DOI: `10.1016/j.mce.2007.05.015`.

[17] P. Rogliani, B. L. Ritondo, E. Puxeddu, G. Pane, M. Cazzola, and L. Calzetta. “Experimental Glucocorticoid Receptor Agonists for the Treatment of Asthma: A Systematic Review”. In: *Journal of Experimental Pharmacology* 12 (2020), pp. 233–254. ISSN: 1179-1454. DOI: `10.2147/JEP.S237480`.

[18] T. E. Reddy, F. Pauli, R. O. Sprouse, N. F. Neff, K. M. Newberry, M. J. Garabedian, and R. M. Myers. “Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation”. In: *Genome Research* 19.12 (2009), pp. 2163–2171. ISSN: 1549-5469. DOI: `10.1101/gr.097022.109`.

[19] Y.-H. Chen, J. H. Kim, and M. R. Stallcup. “GAC63, a GRIP1-Dependent Nuclear Receptor Coactivator”. In: *Molecular and Cellular Biology* 25.14 (2005), pp. 5965–5972. ISSN: 0270-7306. DOI: `10.1128/MCB.25.14.5965-5972.2005`.

[20] J. Dobrovolna, Y. Chinenov, M. A. Kennedy, B. Liu, and I. Rogatsky. “Glucocorticoid-Dependent Phosphorylation of the Transcriptional Coregulator GRIP1”. In: *Molecular and Cellular Biology* 32.4 (2012), pp. 730–739. ISSN: 0270-7306. DOI: `10.1128/MCB.06473-11`.

[21] D. A. Rollins, J. B. Kharlyngdoh, M. Coppo, B. Tharmalingam, S. Mimouna, Z. Guo, M. A. Sacta, M. A. Pufall, R. P. Fisher, X. Hu, Y. Chinenov, and I. Rogatsky. “Glucocorticoid-induced phosphorylation by CDK9 modulates the coactivator functions of transcriptional cofactor GRIP1 in macrophages”. In: *Nature Communications* 8 (2017), p. 1739. ISSN: 2041-1723. DOI: `10.1038/s41467-017-01569-2`.

[22] A. E. Wallberg, K. E. Neely, A. H. Hassan, J.-Å. Gustafsson, J. L. Workman, and A. P. H. Wright. "Recruitment of the SWI-SNF Chromatin Remodeling Complex as a Mechanism of Gene Activation by the Glucocorticoid Receptor \texttau1 Activation Domain". In: *Molecular and Cellular Biology* 20.6 (2000), pp. 2004–2013. ISSN: 0270-7306.

[23] A. N. Gerber, R. Newton, and S. K. Sasse. "Repression of transcription by the glucocorticoid receptor: A parsimonious model for the genomics era". In: *Journal of Biological Chemistry* 296 (2021), p. 100687. ISSN: 0021-9258. DOI: 10.1016/j.jbc.2021.100687.

[24] L. Escoter-Torres, F. Greulich, F. Quagliarini, M. Wierer, and N. H. Uhlenhaut. "Anti-inflammatory functions of the glucocorticoid receptor require DNA binding". In: *Nucleic Acids Research* 48.15 (2020), pp. 8393–8407. ISSN: 1362-4962. DOI: 10.1093/nar/gkaa565.

[25] W. H. Hudson, C. Youn, and E. A. Ortlund. "The structural basis of direct glucocorticoid-mediated transrepression". In: *Nature Structural & Molecular Biology* 20.1 (2013), pp. 53–58. ISSN: 1545-9985. DOI: 10.1038/nsmb.2456.

[26] S. R. Starick, J. Ibn-Salem, M. Jurk, C. Hernandez, M. I. Love, H.-R. Chung, M. Vingron, M. Thomas-Chollier, and S. H. Meijsing. "ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors". In: *Genome Research* 25.6 (2015), pp. 825–835. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.185157.114.

[27] N. H. Uhlenhaut, G. D. Barish, R. T. Yu, M. Downes, M. Karunasiri, C. Liddle, P. Schwalie, N. Hübner, and R. M. Evans. "Insights into Negative Regulation by the Glucocorticoid Receptor from Genome-wide Profiling of Inflammatory Cistromes". In: *Molecular Cell* 49.1 (2013), pp. 158–171. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2012.10.013.

[28] F. Frank, X. Liu, and E. A. Ortlund. "Glucocorticoid receptor condensates link DNA-dependent receptor dimerization and transcriptional transactivation". In: *Proceedings of the National Academy of Sciences* 118.30 (2021), e2024685118. DOI: 10.1073/pnas.2024685118.

[29] B. A. Strickland, S. A. Ansari, W. Dantoft, and N. H. Uhlenhaut. "How to tame your genes: mechanisms of inflammatory gene repression by glucocorticoids". In: *FEBS letters* 596.20 (2022), pp. 2596–2616. ISSN: 1873-3468. DOI: 10.1002/1873-3468.14409.

[30] L. Wang, T. G. Oh, J. Magida, G. Estepa, S. M. B. Obayomi, L.-W. Chong, J. Gatchalian, R. T. Yu, A. R. Atkins, D. Hargreaves, M. Downes, Z. Wei, and R. M. Evans. "Bromodomain containing 9 (BRD9) regulates macrophage inflammatory responses by potentiating glucocorticoid receptor activity". In: *Proceedings of the National Academy of Sciences of the United States of America* 118.35 (2021), e2109517118. ISSN: 1091-6490. DOI: 10.1073/pnas.2109517118.

[31] J. Charles A Janeway, P. Travers, M. Walport, and M. J. Shlomchik. "Generation of lymphocytes in bone marrow and thymus". In: *Immunobiology: The Immune System in Health and Disease. 5th edition*. Garland Science, 2001.

[32] K. Murphy. *Janeway's immunobiology*. Ninth edition. New York, NY, USA: Garland Science, Taylor & Francis Group, LLC, 2017. ISBN: 978-0-8153-4505-3.

[33] J. C. Ribot, N. Lopes, and B. Silva-Santos. "$\gamma\delta$ T cells in tissue physiology and surveillance". In: *Nature Reviews. Immunology* 21.4 (2021), pp. 221–232. ISSN: 1474-1741. DOI: 10.1038/s41577-020-00452-4.

[34] F. Sallusto. "Heterogeneity of Human CD4(+) T Cells Against Microbes". In: *Annual Review of Immunology* 34 (2016), pp. 317–334. ISSN: 1545-3278. DOI: 10.1146/annurev-immunol-032414-112056.

[35] D. G. Schatz and Y. Ji. "Recombination centres and the orchestration of V(D)J recombination". In: *Nature Reviews Immunology* 11.4 (2011), pp. 251–263. ISSN: 1474-1741. DOI: 10.1038/nri2941.

[36] J. Nikolich-Zugich, M. K. Slifka, and I. Messaoudi. "The many important facets of T-cell repertoire diversity". In: *Nature Reviews. Immunology* 4.2 (2004), pp. 123–132. ISSN: 1474-1733. DOI: 10.1038/nri1292.

[37] J. A. Pai and A. T. Satpathy. "High-throughput and single-cell T cell receptor sequencing technologies". In: *Nature Methods* 18.8 (2021), pp. 881–892. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01201-8.

[38] D. Redmond, A. Poran, and O. Elemento. "Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq". In: *Genome Medicine* 8.1 (2016), p. 80. ISSN: 1756-994X. DOI: 10.1186/s13073-016-0335-7.

[39] U. H. von Andrian and T. R. Mempel. "Homing and cellular traffic in lymph nodes". In: *Nature Reviews Immunology* 3.11 (2003), pp. 867–878. ISSN: 1474-1741. DOI: 10.1038/nri1222.

[40] N. Pishesha, T. J. Harmand, and H. L. Ploegh. "A guide to antigen processing and presentation". In: *Nature Reviews. Immunology* 22.12 (2022), pp. 751–764. ISSN: 1474-1741. DOI: 10.1038/s41577-022-00707-2.

[41] J. W. J. van Heijst, C. Gerlach, E. Swart, D. Sie, C. Nunes-Alves, R. M. Kerkhoven, R. Arens, M. Correia-Neves, K. Schepers, and T. N. M. Schumacher. "Recruitment of antigen-specific CD8+ T cells in response to infection is markedly efficient". In: *Science (New York, N.Y.)* 325.5945 (2009), pp. 1265–1269. ISSN: 1095-9203. DOI: 10.1126/science.1175455.

[42] S. M. Kaech and E. J. Wherry. "Heterogeneity and cell-fate decisions in effector and memory CD8+ T cell differentiation during viral infection". In: *Immunity* 27.3 (2007), pp. 393–405. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2007.08.007.

[43] D. Raghu, H.-H. Xue, and L. A. Mielke. "Control of Lymphocyte Fate, Infection, and Tumor Immunity by TCF-1". In: *Trends in Immunology* 40.12 (2019), pp. 1149–1162. ISSN: 14714906. DOI: 10.1016/j.it.2019.10.006.

[44] E. J. Wherry and R. Ahmed. “Memory CD8 T-cell differentiation during viral infection”. In: *Journal of Virology* 78.11 (2004), pp. 5535–5545. ISSN: 0022-538X. DOI: 10.1128/JVI.78.11.5535-5545.2004.

[45] B. M. Sullivan, S. F. Emonet, M. J. Welch, A. M. Lee, K. P. Campbell, J. C. de la Torre, and M. B. Oldstone. “Point mutation in the glycoprotein of lymphocytic choriomeningitis virus is necessary for receptor binding, dendritic cell infection, and long-term persistence”. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.7 (2011), pp. 2969–2974. ISSN: 0027-8424. DOI: 10.1073/pnas.1019304108.

[46] E. J. Wherry and M. Kurachi. “Molecular and cellular insights into T cell exhaustion”. In: *Nature reviews. Immunology* 15.8 (2015), pp. 486–499. ISSN: 1474-1733. DOI: 10.1038/nri3862.

[47] D. L. Barber, E. J. Wherry, D. Masopust, B. Zhu, J. P. Allison, A. H. Sharpe, G. J. Freeman, and R. Ahmed. “Restoring function in exhausted CD8 T cells during chronic viral infection”. In: *Nature* 439.7077 (2006), pp. 682–687. ISSN: 1476-4687. DOI: 10.1038/nature04444.

[48] F. Alfei, K. Kanev, M. Hofmann, M. Wu, H. E. Ghoneim, P. Roelli, D. T. Utzschneider, M. von Hoesslin, J. G. Cullen, Y. Fan, V. Eisenberg, D. Wohlleber, K. Steiger, D. Merkler, M. Delorenzi, P. A. Knolle, C. J. Cohen, R. Thimme, B. Youngblood, and D. Zehn. “TOX reinforces the phenotype and longevity of exhausted T cells in chronic viral infection”. In: *Nature* 571.7764 (2019), pp. 265–269. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1326-9.

[49] A. C. Scott, F. Dündar, P. Zumbo, et al. “TOX is a critical regulator of tumour-specific T cell differentiation”. In: *Nature* 571.7764 (2019), pp. 270–274. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1324-y.

[50] M. Hashimoto, K. Araki, M. A. Cardenas, et al. “PD-1 combination therapy with IL-2 modifies CD8+ T cell exhaustion program”. In: *Nature* 610.7930 (2022), pp. 173–181. ISSN: 1476-4687. DOI: 10.1038/s41586-022-05257-0.

[51] E. B. Garon, N. A. Rizvi, R. Hui, et al. “Pembrolizumab for the treatment of non-small-cell lung cancer”. In: *The New England Journal of Medicine* 372.21 (2015), pp. 2018–2028. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1501824.

[52] S. J. Im, M. Hashimoto, M. Y. Gerner, J. Lee, H. T. Kissick, M. C. Burger, Q. Shan, J. S. Hale, J. Lee, T. H. Nasti, A. H. Sharpe, G. J. Freeman, R. N. Germain, H. I. Nakaya, H.-H. Xue, and R. Ahmed. “Defining CD8+ T cells that provide the proliferative burst after PD-1 therapy”. In: *Nature* 537.7620 (2016), pp. 417–421. ISSN: 1476-4687. DOI: 10.1038/nature19330.

[53] B. C. Miller, D. R. Sen, R. A. Abosy, et al. “Subsets of exhausted CD8+ T cells differentially mediate tumor control and respond to checkpoint blockade”. In: *Nature immunology* 20.3 (2019), pp. 326–336. ISSN: 1529-2908. DOI: 10.1038/s41590-019-0312-6.

[54] C. U. Blank, W. N. Haining, W. Held, P. G. Hogan, A. Kallies, E. Lugli, R. C. Lynn, M. Philip, A. Rao, N. P. Restifo, A. Schietinger, T. N. Schumacher, P. L. Schwartzberg, A. H. Sharpe, D. E. Speiser, E. J. Wherry, B. A. Youngblood, and D. Zehn. "Defining 'T cell exhaustion'". In: *Nature Reviews. Immunology* 19.11 (2019), pp. 665–674. ISSN: 1474-1741. DOI: 10.1038/s41577-019-0221-9.

[55] B. J. Schneider, J. Naidoo, B. D. Santomasso, et al. "Management of Immune-Related Adverse Events in Patients Treated With Immune Checkpoint Inhibitor Therapy: ASCO Guideline Update". In: *Journal of Clinical Oncology* 39.36 (2021), pp. 4073–4126. ISSN: 0732-183X. DOI: 10.1200/JCO.21.01440.

[56] G. C. Sim and L. Radvanyi. "The IL-2 cytokine family in cancer immunotherapy". In: *Cytokine & Growth Factor Reviews* 25.4 (2014), pp. 377–390. ISSN: 1879-0305. DOI: 10.1016/j.cytogfr.2014.07.018.

[57] O. Boyman and J. Sprent. "The role of interleukin-2 during homeostasis and activation of the immune system". In: *Nature Reviews Immunology* 12.3 (2012), pp. 180–190. ISSN: 1474-1741. DOI: 10.1038/nri3156.

[58] L. Khoryati, M. N. Pham, M. Sherve, S. Kumari, K. Cook, J. Pearson, M. Bogdani, D. J. Campbell, and M. A. Gavin. "An IL-2 mutein engineered to promote expansion of regulatory T cells arrests ongoing autoimmunity in mice". In: *Science Immunology* 5.50 (2020), eaba5264. ISSN: 2470-9468. DOI: 10.1126/sciimmunol.aba5264.

[59] C. Klein, I. Waldhauer, V. G. Nicolini, et al. "Cergutuzumab amunaleukin (CEA-IL2v), a CEA-targeted IL-2 variant-based immunocytokine for combination cancer immunotherapy: Overcoming limitations of aldesleukin and conventional IL-2-based immunocytokines". In: *Oncoimmunology* 6.3 (2017), e1277306. ISSN: 2162-4011. DOI: 10.1080/2162402X.2016.1277306.

[60] I. Waldhauer, V. Gonzalez-Nicolini, A. Freimoser-Grundschober, et al. "Simlukafusp alfa (FAP-IL2v) immunocytokine is a versatile combination partner for cancer immunotherapy". In: *mAbs* 13.1 (2021), p. 1913791. ISSN: 1942-0870. DOI: 10.1080/19420862.2021.1913791.

[61] L. Codarri Deak, V. Nicolini, M. Hashimoto, et al. "PD-1-cis IL-2R agonism yields better effectors from stem-like CD8+ T cells". In: *Nature* 610.7930 (2022), pp. 161–172. ISSN: 1476-4687. DOI: 10.1038/s41586-022-05192-0.

[62] O. Khan, J. R. Giles, S. McDonald, et al. "TOX transcriptionally and epigenetically programs CD8+ T cell exhaustion". In: *Nature* 571.7764 (2019), pp. 211–218. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1325-x.

[63] K. E. Pauken, M. A. Sammons, P. M. Odorizzi, S. Manne, J. Godec, O. Khan, A. M. Drake, Z. Chen, D. R. Sen, M. Kurachi, R. A. Barnitz, C. Bartman, B. Bengsch, A. C. Huang, J. M. Schenkel, G. Vahedi, W. N. Haining, S. L. Berger, and E. J. Wherry. "Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade". In: *Science (New York, N.Y.)* 354.6316 (2016), pp. 1160–1165. ISSN: 1095-9203. DOI: 10.1126/science.aaf2807.

[64] H. Seo, J. Chen, E. González-Avalos, D. Samaniego-Castruita, A. Das, Y. H. Wang, I. F. López-Moyado, R. O. Georges, W. Zhang, A. Onodera, C.-J. Wu, L.-F. Lu, P. G. Hogan, A. Bhandoola, and A. Rao. "TOX and TOX2 transcription factors cooperate with NR4A transcription factors to impose CD8+ T cell exhaustion". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.25 (2019), pp. 12410–12415. ISSN: 1091-6490. DOI: 10.1073/pnas.1905675116.

[65] F. Crick. "Central Dogma of Molecular Biology". In: *Nature* 227.5258 (1970), pp. 561–563. ISSN: 1476-4687. DOI: 10.1038/227561a0.

[66] F. H. Crick. "On protein synthesis". In: *Symposia of the Society for Experimental Biology* 12 (1958), pp. 138–163. ISSN: 0081-1386.

[67] S. G. Landt, G. K. Marinov, A. Kundaje, et al. "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia". In: *Genome Research* 22.9 (2012), pp. 1813–1831. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.136184.111.

[68] Y. Chu and D. R. Corey. "RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation". In: *Nucleic Acid Therapeutics* 22.4 (2012), pp. 271–274. ISSN: 2159-3337. DOI: 10.1089/nat.2012.0367.

[69] Z. Wang, M. Gerstein, and M. Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature reviews. Genetics* 10.1 (2009), pp. 57–63. ISSN: 1471-0056. DOI: 10.1038/nrg2484.

[70] J. Buenrostro, B. Wu, H. Chang, and W. Greenleaf. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". In: *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* 109 (2015), pp. 21.29.1–21.29.9. ISSN: 1934-3639. DOI: 10.1002/0471142727.mb2129s109.

[71] J. R. Hughes, N. Roberts, S. McGowan, D. Hay, E. Giannoulatou, M. Lynch, M. De Gobbi, S. Taylor, R. Gibbons, and D. R. Higgs. "Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment". In: *Nature Genetics* 46.2 (2014), pp. 205–212. ISSN: 1546-1718. DOI: 10.1038/ng.2871.

[72] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. "Comprehensive mapping of long range interactions reveals folding principles of the human genome". In: *Science (New York, N.Y.)* 326.5950 (2009), pp. 289–293. ISSN: 0036-8075. DOI: 10.1126/science.1181369.

[73] B. Hwang, J. H. Lee, and D. Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental & Molecular Medicine* 50.8 (2018), pp. 1–14. ISSN: 2092-6413. DOI: 10.1038/s12276-018-0071-8.

[74] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. "Simultaneous epitope and transcriptome measurement in single cells". In: *Nature Methods* 14.9 (2017), pp. 865–868. ISSN: 1548-7105. DOI: 10.1038/nmeth.4380.

[75] V. G. Allfrey, R. Faulkner, and A. E. Mirsky. "Acetylation and Methylation of Histones and their Possible Role in the Regulation of RNA Synthesis". In: *Proceedings of the National Academy of Sciences of the United States of America* 51.5 (1964), pp. 786–794. ISSN: 0027-8424.

[76] R. Marmorstein. "Protein modules that manipulate histone tails for chromatin regulation". In: *Nature Reviews Molecular Cell Biology* 2.6 (2001), pp. 422–432. ISSN: 1471-0080. DOI: 10.1038/35073047.

[77] S. L. Berger. "The complex language of chromatin regulation during transcription". In: *Nature* 447.7143 (2007), pp. 407–412. ISSN: 1476-4687. DOI: 10.1038/nature05915.

[78] T. Jenuwein and C. D. Allis. "Translating the Histone Code". In: *Science* 293.5532 (2001), pp. 1074–1080. DOI: 10.1126/science.1063127.

[79] A. M. Bolger, M. Lohse, and B. Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (2014), pp. 2114–2120. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu170.

[80] M. Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1 (2011), pp. 10–12. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200.

[81] B. Langmead and S. L. Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4 (2012), pp. 357–359. ISSN: 1548-7091. DOI: 10.1038/nmeth.1923.

[82] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. "Model-based analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (2008), R137. ISSN: 1474-760X. DOI: 10.1186/gb-2008-9-9-r137.

[83] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein. "PeakSeq: Systematic Scoring of ChIP-Seq Experiments Relative to Controls". In: *Nature biotechnology* 27.1 (2009), pp. 66–75. ISSN: 1087-0156. DOI: 10.1038/nbt.1518.

[84] Q. Li, J. B. Brown, H. Huang, and P. J. Bickel. "Measuring reproducibility of high-throughput experiments". In: *The Annals of Applied Statistics* 5.3 (2011), pp. 1752–1779. ISSN: 1932-6157. DOI: 10.1214/11-AOAS466.

[85] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent. "The UCSC Genome Browser Database". In: *Nucleic Acids Research* 31.1 (2003), pp. 51–54. ISSN: 0305-1048. DOI: 10.1093/nar/gkg129.

[86] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. "Integrative Genomics Viewer". In: *Nature biotechnology* 29.1 (2011), pp. 24–26. ISSN: 1087-0156. DOI: 10.1038/nbt.1754.

[87] G. Yu, L.-G. Wang, and Q.-Y. He. "ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization". In: *Bioinformatics* 31.14 (2015), pp. 2382–2383. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv145.

[88] S. Schoenfelder, B.-M. Javierre, M. Furlan-Magaril, S. W. Wingett, and P. Fraser. "Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions". In: *Journal of Visualized Experiments : JoVE* 136 (2018). ISSN: 1940-087X. DOI: 10.3791/57320.

[89] F. Zambelli, G. Pesole, and G. Pavesi. "Motif discovery and transcription factor binding sites before and after the next-generation sequencing era". In: *Briefings in Bioinformatics* 14.2 (2013), pp. 225–237. ISSN: 1467-5463. DOI: 10.1093/bib/bbs016.

[90] P. Machanick and T. L. Bailey. "MEME-ChIP: motif analysis of large DNA datasets". In: *Bioinformatics* 27.12 (2011), pp. 1696–1697. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr189.

[91] J. A. Castro-Mondragon, R. Riudavets-Puig, I. Rauluseviciute, et al. "JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles". In: *Nucleic Acids Research* 50.D1 (2022), pp. D165–D173. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1113.

[92] G. Felsenfeld and M. Groudine. "Controlling the double helix". In: *Nature* 421.6921 (2003), pp. 448–453. ISSN: 1476-4687. DOI: 10.1038/nature01411.

[93] L. Song and G. E. Crawford. "DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells". In: *Cold Spring Harbor Protocols* 2010.2 (2010), pdb.prot5384. ISSN: 1940-3402, 1559-6095. DOI: 10.1101/pdb.prot5384.

[94] W. S. Reznikoff. "Transposon Tn5". In: *Annual Review of Genetics* 42.1 (2008), pp. 269–286. DOI: 10.1146/annurev.genet.42.110807.091656.

[95] Z. Li, M. H. Schulz, T. Look, M. Begemann, M. Zenke, and I. G. Costa. "Identification of transcription factor binding sites using ATAC-seq". In: *Genome Biology* 20.1 (2019), p. 45. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1642-2.

[96] A. P. Boyle, L. Song, B.-K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford, and T. S. Furey. "High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells". In: *Genome Research* 21.3 (2011), pp. 456–464. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.112656.110.

[97] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". In: *Nature Biotechnology* 28.5 (2010), pp. 511–515. ISSN: 1546-1696. DOI: 10.1038/nbt.1621.

[98] B. J. Haas, A. Papanicolaou, M. Yassour, et al. "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis". In: *Nature Protocols* 8.8 (2013), pp. 1494–1512. ISSN: 1750-2799. DOI: 10.1038/nprot.2013.084.

[99] B. Rädle, A. J. Rutkowski, Z. Ruzsics, C. C. Friedel, U. H. Koszinowski, and L. Dölken. "Metabolic labeling of newly transcribed RNA for high resolution gene expression profiling of RNA synthesis, processing and decay in cell culture". In: *Journal of Visualized Experiments: JoVE* 78 (2013), p. 50195. ISSN: 1940-087X. DOI: 10.3791/50195.

[100] M. Kanamori-Katayama, M. Itoh, H. Kawaji, T. Lassmann, S. Katayama, M. Kojima, N. Bertin, A. Kaiho, N. Ninomiya, C. O. Daub, P. Carninci, A. R. R. Forrest, and Y. Hayashizaki. "Unamplified cap analysis of gene expression on a single-molecule sequencer". In: *Genome Research* 21.7 (2011), pp. 1150–1159. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.115469.110.

[101] W. Zhao, X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, and C. M. Perou. "Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling". In: *BMC Genomics* 15.1 (2014), p. 419. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-419.

[102] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts635.

[103] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biology* 14.4 (2013), R36. ISSN: 1474-760X. DOI: 10.1186/gb-2013-14-4-r36.

[104] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nature Methods* 14.4 (2017), pp. 417–419. ISSN: 1548-7105. DOI: 10.1038/nmeth.4197.

[105] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology* 34.5 (2016), pp. 525–527. ISSN: 1546-1696. DOI: 10.1038/nbt.3519.

[106] Y. Liao, G. K. Smyth, and W. Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics (Oxford, England)* 30.7 (2014), pp. 923–930. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt656.

[107] Y. Benjamini and T. P. Speed. "Summarizing and correcting the GC content bias in high-throughput sequencing". In: *Nucleic Acids Research* 40.10 (2012), e72. ISSN: 0305-1048. DOI: 10.1093/nar/gks001.

[108] M. D. Robinson and A. Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome Biology* 11.3 (2010), R25. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-3-r25.

[109] S. Anders and W. Huber. "Differential expression analysis for sequence count data". In: *Genome Biology* 11.10 (2010), R106. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-10-r106.

[110] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015), e47. ISSN: 1362-4962. DOI: 10.1093/nar/gkv007.

[111] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics (Oxford, England)* 26.1 (2010), pp. 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616.

[112] L. Larsson, J. Frisén, and J. Lundeberg. "Spatially resolved transcriptomics adds a new dimension to genomics". In: *Nature Methods* 18.1 (2021), pp. 15–18. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01038-7.

[113] F. A. Wolf, P. Angerer, and F. J. Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome Biology* 19.1 (2018), p. 15. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0.

[114] Y. Hao, S. Hao, E. Andersen-Nissen, et al. "Integrated analysis of multimodal single-cell data". In: *Cell* 184.13 (2021), 3573–3587.e29. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2021.04.048.

[115] T. Stuart, A. Srivastava, S. Madad, C. A. Lareau, and R. Satija. "Single-cell chromatin state analysis with Signac". In: *Nature Methods* 18.11 (2021), pp. 1333–1341. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01282-5.

[116] M. D. Young and S. Behjati. "SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data". In: *GigaScience* 9.12 (2020), giaa151. ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa151.

[117] M. D. Luecken and F. J. Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular Systems Biology* 15.6 (2019), e8746. ISSN: 1744-4292. DOI: 10.15252/msb.20188746.

[118] P.-L. Germain, A. Lun, C. G. Meixide, W. Macnair, and M. D. Robinson. *Doublet identification in single-cell sequencing data using scDblFinder*. Tech. rep. 10:979. F1000Research, 2022. DOI: 10.12688/f1000research.73600.2.

[119] L. Heumos, A. C. Schaar, C. Lance, A. Litinetskaya, F. Drost, L. Zappia, M. D. Lücken, D. C. Strobl, J. Henao, F. Curion, Single-cell Best Practices Consortium, H. B. Schiller, and F. J. Theis. "Best practices for single-cell analysis across modalities". In: *Nature Reviews. Genetics* (2023), pp. 1–23. ISSN: 1471-0064. DOI: 10.1038/s41576-023-00586-w.

[120] C. Ahlmann-Eltze and W. Huber. "Comparison of transformations for single-cell RNA-seq data". In: *Nature Methods* 20.5 (2023), pp. 665–672. ISSN: 1548-7105. DOI: 10.1038/s41592-023-01814-1.

[121] C. Hafemeister and R. Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome Biology* 20.1 (2019), p. 296. ISSN: 1474-760X. DOI: `10.1186/s13059-019-1874-1`.

[122] P.-L. Germain, A. Sonrel, and M. D. Robinson. "pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools". In: *Genome Biology* 21.1 (2020), p. 227. ISSN: 1474-760X. DOI: `10.1186/s13059-020-02136-7`.

[123] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Probability and mathematical statistics. London ; New York: Academic Press, 1979. ISBN: 978-0-12-471252-2.

[124] L. McInnes, J. Healy, and J. Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. DOI: `10.48550/arXiv.1802.03426`.

[125] V. A. Traag, L. Waltman, and N. J. van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". In: *Scientific Reports* 9.1 (2019), p. 5233. ISSN: 2045-2322. DOI: `10.1038/s41598-019-41695-z`.

[126] A. Natarajan, G. G. Yardımcı, N. C. Sheffield, G. E. Crawford, and U. Ohler. "Predicting cell-type–specific gene expression from regions of open chromatin". In: *Genome Research* 22.9 (2012), pp. 1711–1722. ISSN: 1088-9051, 1549-5469. DOI: `10.1101/gr.135129.111`.

[127] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis. "Deep learning: new computational modelling techniques for genomics". In: *Nature Reviews Genetics* 20.7 (2019), pp. 389–403. ISSN: 1471-0064. DOI: `10.1038/s41576-019-0122-6`.

[128] D. Quang and X. Xie. "FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data". In: *Methods* 166 (2019), pp. 40–47. ISSN: 10462023. DOI: `10.1016/j.ymeth.2019.03.020`.

[129] Ž. Avsec, M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, R. Fropf, C. McAnany, J. Gagneur, A. Kundaje, and J. Zeitlinger. "Base-resolution models of transcription-factor binding reveal soft motif syntax". In: *Nature Genetics* 53 (2021), pp. 1–13. DOI: `10.1038/s41588-021-00782-6`.

[130] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya. "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk". In: *Nature Genetics* 50.8 (2018), pp. 1171–1179. ISSN: 1546-1718. DOI: `10.1038/s41588-018-0160-6`.

[131] V. Agarwal and J. Shendure. "Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks". In: *Cell Reports* 31.7 (2020), p. 107663. ISSN: 2211-1247. DOI: `10.1016/j.celrep.2020.107663`.

[132] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[133] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome". In: *Bioinformatics* 37.15 (2021), pp. 2112–2120. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btab083`.

[134] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley. "Effective gene expression prediction from sequence by integrating long-range interactions". In: *Nature Methods* 18.10 (2021), pp. 1196–1203. ISSN: 1548-7105. DOI: `10.1038/s41592-021-01252-x`.

[135] D. Ragab, H. Salah Eldin, M. Taeimah, R. Khattab, and R. Salem. "The COVID-19 Cytokine Storm; What We Know So Far". In: *Frontiers in Immunology* 11 (2020). ISSN: 1664-3224.

[136] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. 3rd ed. Morgan Kaufmann Publishers, 2011. ISBN: 978-0-12-381480-7.

[137] J. A. Nelder and R. W. M. Wedderburn. "Generalized Linear Models". In: *Journal of the Royal Statistical Society: Series A (General)* 135.3 (1972), pp. 370–384. ISSN: 2397-2327. DOI: `10.2307/2344614`.

[138] M. I. Love, W. Huber, and S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12 (2014), pp. 1–21.

[139] H. Zou and T. Hastie. "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320. ISSN: 1369-7412.

[140] T. Fawcett. "ROC Graphs: Notes and Practical Considerations for Researchers". In: 2007.

[141] M. Andrychowicz, M. Denil, S. Gómez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. de Freitas. "Learning to learn by gradient descent by gradient descent". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016.

[142] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. DOI: `10.48550/arXiv.1412.6980`.

[143] I. Loshchilov and F. Hutter. *Decoupled Weight Decay Regularization*. 2019. DOI: `10.48550/arXiv.1711.05101`.

[144] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. ISSN: 1558-2256. DOI: `10.1109/5.726791`.

[145] A. Farcomeni. "A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion". In: *Statistical Methods in Medical Research* 17.4 (2008), pp. 347–388. ISSN: 0962-2802. DOI: `10.1177/0962280206079046`.

[146] Y. Benjamini and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 289–300. ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[147] H. Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of Educational Psychology* 24.6 (1933), pp. 417–441. ISSN: 1939-2176, 0022-0663. DOI: 10.1037/h0071325.

[148] J. B. Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1 (1964), pp. 1–27. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/BF02289565.

[149] L. van der Maaten and G. Hinton. "Viualizing data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[150] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. "Integrating single-cell transcriptomic data across different conditions, technologies, and species". In: *Nature Biotechnology* 36.5 (2018), pp. 411–420. ISSN: 1546-1696. DOI: 10.1038/nbt.4096.

[151] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature Biotechnology* 37.1 (2019), pp. 38–44. ISSN: 1546-1696. DOI: 10.1038/nbt.4314.

[152] F. Greulich, K. A. Bielefeld, R. Scheundel, A. Mechtidou, B. Strickland, and N. H. Uhlenhaut. "Enhancer RNA Expression in Response to Glucocorticoid Treatment in Murine Macrophages". In: *Cells* 11.1 (2021), p. 28. ISSN: 2073-4409. DOI: 10.3390/cells11010028.

[153] M. A. Sacta, B. Tharmalingam, M. Coppo, D. A. Rollins, D. K. Deochand, B. Benjamin, L. Yu, B. Zhang, X. Hu, R. Li, Y. Chinenov, and I. Rogatsky. "Gene-specific mechanisms direct glucocorticoid-receptor-driven repression of inflammatory response genes in macrophages". In: *eLife* 7 (2018), e34864. ISSN: 2050-084X. DOI: 10.7554/eLife.34864.

[154] A. Mechtidou, F. Greulich, B. A. Strickland, C. Jouffe, F. M. Cernilogar, G. Schotta, and N. H. Uhlenhaut. *BRG1 defines a genomic subset of inflammatory genes transcriptionally controlled by the glucocorticoid receptor*. 2021. DOI: 10.1101/2021.12.13.472398.

[155] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble. "The MEME Suite". In: *Nucleic Acids Research* 43.W1 (2015), W39–W49. ISSN: 0305-1048. DOI: 10.1093/nar/gkv416.

[156] L. Grøntved, S. John, S. Baek, Y. Liu, J. R. Buckley, C. Vinson, G. Aguilera, and G. L. Hager. "C/EBP maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements". In: *The EMBO Journal* 32.11 (2013), pp. 1568–1583. ISSN: 0261-4189. DOI: 10.1038/emboj.2013.106.

[157] S. C. Biddie, S. John, P. J. Sabo, R. E. Thurman, T. A. Johnson, R. L. Schiltz, T. B. Miranda, M.-H. Sung, S. Trump, S. L. Lightman, C. Vinson, J. A. Stamatoyannopoulos, and G. L. Hager. "Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding". In: *Molecular Cell* 43.1 (2011), pp. 145–155. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2011.06.016.

[158] F. Quagliarini, A. A. Mir, K. Balazs, M. Wierer, K. A. Dyar, C. Jouffe, K. Makris, J. Hawe, M. Heinig, F. V. Filipp, G. D. Barish, and N. H. Uhlenhaut. "Cistromic Reprogramming of the Diurnal Glucocorticoid Hormone Response by High-Fat Diet". In: *Molecular Cell* 76.4 (2019), 531–545.e5. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2019.10.007.

[159] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young, and R. Jaenisch. "Histone H3K27ac separates active from poised enhancers and predicts developmental state". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.50 (2010), pp. 21931–21936. ISSN: 1091-6490. DOI: 10.1073/pnas.1016071107.

[160] K.-S. Oh, H. Patel, R. A. Gottschalk, W. S. Lee, S. Baek, I. D. Fraser, G. L. Hager, and M.-H. Sung. "Anti-Inflammatory Chromatinscape Suggests Alternative Mechanisms of Glucocorticoid Receptor Action". In: *Immunity* 47.2 (2017), 298–309.e5. ISSN: 10747613. DOI: 10.1016/j.immuni.2017.07.012.

[161] C. P. Fulco, J. Nasser, T. R. Jones, G. Munson, D. T. Bergman, V. Subramanian, S. R. Grossman, R. Anyoha, B. R. Doughty, T. A. Patwardhan, T. H. Nguyen, M. Kane, E. M. Perez, N. C. Durand, C. A. Lareau, E. K. Stamenova, E. L. Aiden, E. S. Lander, and J. M. Engreitz. "Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations". In: *Nature Genetics* 51.12 (2019), pp. 1664–1669. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0538-0.

[162] M. R. Mumbach, J. M. Granja, R. A. Flynn, C. M. Roake, A. T. Satpathy, A. J. Rubin, Y. Qi, Z. Jiang, S. Shams, B. H. Louie, J. K. Guo, D. G. Gennert, M. R. Corces, P. A. Khavari, M. K. Atianand, S. E. Artandi, K. A. Fitzgerald, W. J. Greenleaf, and H. Y. Chang. "HiChIRP reveals RNA-associated chromosome conformation". In: *Nature Methods* 16.6 (2019), pp. 489–492. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0407-x.

[163] F. Greulich, M. Wierer, A. Mechtidou, O. Gonzalez-Garcia, and N. H. Uhlenhaut. "The glucocorticoid receptor recruits the COMPASS complex to regulate inflammatory transcription at macrophage enhancers". In: *Cell Reports* 34.6 (2021), p. 108742. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2021.108742.

[164] J. E. Darnell. "STATs and gene regulation". In: *Science (New York, N.Y.)* 277.5332 (1997), pp. 1630–1635. ISSN: 0036-8075. DOI: 10.1126/science.277.5332.1630.

[165] J. N. Ihle. "The Stat family in cytokine signaling". In: *Current Opinion in Cell Biology* 13.2 (2001), pp. 211–217. ISSN: 0955-0674. DOI: 10.1016/S0955-0674(00)00199-X.

[166] D. T. Utzschneider, S. S. Gabriel, D. Chisanga, R. Gloury, P. M. Gubser, A. Vasanthakumar, W. Shi, and A. Kallies. "Early precursor T cells establish and propagate T cell exhaustion in chronic infection". In: *Nature immunology* 21.10 (2020), pp. 1256–1266.

[167] D. Zehn, R. Thimme, E. Lugli, G. P. de Almeida, and A. Oxenius. "'Stem-like' precursors are the fount to sustain persistent CD8+ T cell responses". In: *Nature Immunology* 23.6 (2022), pp. 836–847. ISSN: 1529-2916. DOI: 10.1038/s41590-022-01219-w.

[168] Y. Pritykin, J. van der Veeken, A. R. Pine, Y. Zhong, M. Sahin, L. Mazutis, D. Pe'er, A. Y. Rudensky, and C. S. Leslie. "A unified atlas of CD8 T cell dysfunctional states in cancer and infection". In: *Molecular Cell* (2021).

[169] A. N. Schep, B. Wu, J. D. Buenrostro, and W. J. Greenleaf. "chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data". In: *Nature Methods* 14.10 (2017), pp. 975–978. ISSN: 1548-7105. DOI: 10.1038/nmeth.4401.

[170] M. Shakiba, P. Zumbo, G. Espinosa-Carrasco, et al. "TCR signal strength defines distinct mechanisms of T cell dysfunction and cancer evasion". In: *The Journal of Experimental Medicine* 219.2 (2022), e20201966. ISSN: 1540-9538. DOI: 10.1084/jem.20201966.

[171] R. G. van der Most, K. Murali-Krishna, J. L. Whitton, C. Oseroff, J. Alexander, S. Southwood, J. Sidney, R. W. Chesnut, A. Sette, and R. Ahmed. "Identification of Db- and Kb-restricted subdominant cytotoxic T-cell responses in lymphocytic choriomeningitis virus-infected mice". In: *Virology* 240.1 (1998), pp. 158–167. ISSN: 0042-6822. DOI: 10.1006/viro.1997.8934.

[172] F. Drost, Y. An, L. M. Dratva, R. G. Lindeboom, M. Haniffa, S. A. Teichmann, F. Theis, M. Lotfollahi, and B. Schubert. *Integrating T-cell receptor and transcriptome for large-scale single-cell immune profiling analysis*. 2022. DOI: 10.1101/2021.06.24.449733.

[173] D. T. Utzschneider, A. Legat, S. A. Fuertes Marraco, L. Carrié, I. Luescher, D. E. Speiser, and D. Zehn. "T cells maintain an exhausted phenotype after antigen withdrawal and population reexpansion". In: *Nature Immunology* 14.6 (2013), pp. 603–610. ISSN: 1529-2916. DOI: 10.1038/ni.2606.

[174] L. M. McLane, M. S. Abdel-Hakeem, and E. J. Wherry. "CD8 T cell exhaustion during chronic viral infection and cancer". In: *Annual review of immunology* 37 (2019), pp. 457–495.

[175] N. Page, B. Klimek, M. De Roo, K. Steinbach, H. Soldati, S. Lemeille, I. Wagner, M. Kreutzfeldt, G. Di Liberto, I. Vincenti, T. Lingner, G. Salinas, W. Brück, M. Simons, R. Murr, J. Kaye, D. Zehn, D. D. Pinschewer, and D. Merkler. "Expression of the DNA-Binding Factor TOX Promotes the Encephalitogenic Potential of Microbe-Induced Autoreactive CD8+ T Cells". In: *Immunity* 48.5 (2018), 937–950.e8. ISSN: 1097-4180. DOI: 10.1016/j.immuni.2018.04.005.

[176] E. O'Flaherty and J. Kaye. "TOX defines a conserved subfamily of HMG-box proteins". In: *BMC Genomics* 4 (2003), p. 13. ISSN: 1471-2164. DOI: 10.1186/1471-2164-4-13.

[177] R. Browaeys, W. Saelens, and Y. Saeys. "NicheNet: modeling intercellular communication by linking ligands to target genes". In: *Nature Methods* 17.2 (2020), pp. 159–162. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0667-5.

[178] P. A. Ewels, A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M. U. Garcia, P. Di Tommaso, and S. Nahnsen. "The nf-core framework for community-curated bioinformatics pipelines". In: *Nature Biotechnology* 38.3 (2020), pp. 276–278. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0439-x.

[179] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame. "Nextflow enables reproducible computational workflows". In: *Nature Biotechnology* 35.4 (2017), pp. 316–319. ISSN: 1546-1696. DOI: `10.1038/nbt.3820`.

[180] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics (Oxford, England)* 25.16 (2009), pp. 2078–2079. ISSN: 1367-4811. DOI: `10.1093/bioinformatics/btp352`.

[181] L. Wang, S. Wang, and W. Li. "RSeQC: quality control of RNA-seq experiments". In: *Bioinformatics (Oxford, England)* 28.16 (2012), pp. 2184–2185. ISSN: 1367-4811. DOI: `10.1093/bioinformatics/bts356`.

[182] K. Okonechnikov, A. Conesa, and F. García-Alcalde. "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data". In: *Bioinformatics (Oxford, England)* 32.2 (2016), pp. 292–294. ISSN: 1367-4811. DOI: `10.1093/bioinformatics/btv566`.

[183] P. Ewels, M. Magnusson, S. Lundin, and M. Käller. "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19 (2016), pp. 3047–3048.

[184] G. Tan and B. Lenhard. "TFBSTools: an R/bioconductor package for transcription factor binding site analysis". In: *Bioinformatics (Oxford, England)* 32.10 (2016), pp. 1555–1556. ISSN: 1367-4811. DOI: `10.1093/bioinformatics/btw024`.

[185] P. Aliahmad and J. Kaye. "Development of all CD4 T lineages requires nuclear factor TOX". In: *The Journal of Experimental Medicine* 205.1 (2008), pp. 245–256. ISSN: 0022-1007. DOI: `10.1084/jem.20071944`.

[186] Ž. Avsec, R. Kreuzhuber, J. Israeli, N. Xu, J. Cheng, A. Shrikumar, A. Banerjee, D. S. Kim, T. Beier, L. Urban, A. Kundaje, O. Stegle, and J. Gagneur. "The Kipoi repository accelerates community exchange and reuse of predictive models for genomics". In: *Nature Biotechnology* 37.6 (2019), pp. 592–600. ISSN: 1546-1696. DOI: `10.1038/s41587-019-0140-0`.

[187] M. Ilse, J. M. Tomczak, and M. Welling. "Attention-based Deep Multiple Instance Learning". In: *arXiv:1802.04712 [cs, stat]* (2018).

[188] A. Sadafi, A. Makhro, A. Bogdanova, N. Navab, T. Peng, S. Albarqouni, and C. Marr. "Attention based Multiple Instance Learning for Classification of Blood Cell Disorders". In: *arXiv:2007.11641 [cs, eess]* (2020).

[189] A. Shrikumar, K. Tian, Ž. Avsec, A. Shcherbina, A. Banerjee, M. Sharmin, S. Nair, and A. Kundaje. *Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5*. 2020. DOI: `10.48550/arXiv.1811.00416`.

[190] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. "clusterProfiler: an R package for comparing biological themes among gene clusters". In: *Omics: A Journal of Integrative Biology* 16.5 (2012), pp. 284–287. ISSN: 1557-8100. DOI: `10.1089/omi.2011.0118`.

[191] Z. Gu, L. Gu, R. Eils, M. Schlesner, and B. Brors. "circlize implements and enhances circular visualization in R". In: *Bioinformatics* 30.19 (2014), pp. 2811–2812. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu393.

[192] I. M. Kramer. "Chapter 8 - Nuclear Receptors". In: *Signal Transduction (Third Edition)*. Ed. by I. M. Kramer. Boston: Academic Press, 2016, pp. 477–527. ISBN: 978-0-12-394803-8. DOI: 10.1016/B978-0-12-394803-8.00008-5.

[193] C.-Z. Song, X. Tian, and T. D. Gelehrter. "Glucocorticoid receptor inhibits transforming growth factor-β signaling by directly targeting the transcriptional activation function of Smad3". In: *Proceedings of the National Academy of Sciences of the United States of America* 96.21 (1999), pp. 11776–11781. ISSN: 0027-8424.

[194] C. Jonat, H. J. Rahmsdorf, K. K. Park, A. C. Cato, S. Gebel, H. Ponta, and P. Herrlich. "Antitumor promotion and antiinflammation: down-modulation of AP-1 (Fos/Jun) activity by glucocorticoid hormone". In: *Cell* 62.6 (1990), pp. 1189–1204. ISSN: 0092-8674. DOI: 10.1016/0092-8674(90)90395-u.

[195] W. H. Hudson, I. M. S. d. Vera, J. C. Nwachukwu, E. R. Weikum, A. G. Herbst, Q. Yang, D. L. Bain, K. W. Nettles, D. J. Kojetin, and E. A. Ortlund. "Cryptic glucocorticoid receptor-binding sites pervade genomic NF-κB response elements". In: *Nature Communications* 9.1 (2018), p. 1337. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03780-1.

[196] E. R. Weikum, I. M. S. de Vera, J. C. Nwachukwu, W. H. Hudson, K. W. Nettles, D. J. Kojetin, and E. A. Ortlund. "Tethering not required: the glucocorticoid receptor binds directly to activator protein-1 recognition motifs to repress inflammatory genes". In: *Nucleic Acids Research* 45.14 (2017), pp. 8596–8608. ISSN: 0305-1048. DOI: 10.1093/nar/gkx509.

[197] L. C. Platanias. "Mechanisms of type-I- and type-II-interferon-mediated signalling". In: *Nature Reviews Immunology* 5.5 (2005), pp. 375–386. ISSN: 1474-1741. DOI: 10.1038/nri1604.

[198] Z. Zhang, S. Jones, J. S. Hagood, N. L. Fuentes, and G. M. Fuller. "STAT3 Acts as a Co-activator of Glucocorticoid Receptor Signaling*". In: *Journal of Biological Chemistry* 272.49 (1997), pp. 30607–30610. ISSN: 0021-9258. DOI: 10.1074/jbc.272.49.30607.

[199] D. Langlais, C. Couture, A. Balsalobre, and J. Drouin. "The Stat3/GR Interaction Code: Predictive Value of Direct/Indirect DNA Recruitment for Transcription Outcome". In: *Molecular Cell* 47.1 (2012), pp. 38–49. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2012.04.021.

[200] J. S. Yi, M. A. Cox, and A. J. Zajac. "T-cell exhaustion: characteristics, causes and conversion". In: *Immunology* 129.4 (2010), pp. 474–481. ISSN: 0019-2805. DOI: 10.1111/j.1365-2567.2010.03255.x.

[201] E. J. Wherry. "T cell exhaustion". In: *Nature Immunology* 12.6 (2011), pp. 492–499. ISSN: 1529-2916. DOI: 10.1038/ni.2035.

[202] D. Zehn, S. Y. Lee, and M. J. Bevan. “Complete but curtailed T-cell response to very low-affinity antigen”. In: *Nature* 458.7235 (2009), pp. 211–214. ISSN: 1476-4687. DOI: 10.1038/nature07657.

[203] D. E. Speiser, D. T. Utzschneider, S. G. Oberle, C. Münz, P. Romero, and D. Zehn. “T cell differentiation in chronic infection and cancer: functional adaptation or exhaustion?” In: *Nature Reviews. Immunology* 14.11 (2014), pp. 768–774. ISSN: 1474-1741. DOI: 10.1038/nri3740.

[204] D. R. Sen, J. Kaminski, R. A. Barnitz, M. Kurachi, U. Gerdemann, K. B. Yates, H.-W. Tsao, J. Godec, M. W. LaFleur, F. D. Brown, P. Tonnerre, R. T. Chung, D. C. Tully, T. M. Allen, N. Frahm, G. M. Lauer, E. J. Wherry, N. Yosef, and W. N. Haining. “The epigenetic landscape of T cell exhaustion”. In: *Science (New York, N.Y.)* 354.6316 (2016), pp. 1165–1169. ISSN: 1095-9203. DOI: 10.1126/science.aae0491.

[205] S. S. Chin, E. Guillen, L. Chorro, S. Achar, K. Ng, S. Oberle, F. Alfei, D. Zehn, G. Altan-Bonnet, F. Delahaye, and G. Lauvau. “T cell receptor and IL-2 signaling strength control memory CD8+ T cell functional fitness via chromatin remodeling”. In: *Nature Communications* 13.1 (2022), p. 2240. ISSN: 2041-1723. DOI: 10.1038/s41467-022-29718-2.

[206] K. S. Schluns and L. Lefrançois. “Cytokine control of memory T-cell development and survival”. In: *Nature Reviews Immunology* 3.4 (2003), pp. 269–279. ISSN: 1474-1741. DOI: 10.1038/nri1052.

[207] A. K. Abbas, E. Trotta, D. R Simeonov, A. Marson, and J. A. Bluestone. “Revisiting IL-2: Biology and therapeutic prospects”. In: *Science Immunology* 3.25 (2018), eaat1482. ISSN: 2470-9468. DOI: 10.1126/sciimmunol.aat1482.

[208] B. Kwon. “The two faces of IL-2: a key driver of CD8+ T-cell exhaustion”. In: *Cellular and Molecular Immunology* 18.7 (2021), pp. 1641–1643. ISSN: 1672-7681. DOI: 10.1038/s41423-021-00712-w.