Technische Universität München
TUM School of Life Sciences

Charting targets for CAR T cell therapy and investigating species-specific neural differentiation: Towards novel cellular therapies at the single cell level

Moritz Thomas, M.Sc.

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Markus List

Prüfende der Dissertation:

1. Prof. Dr. Dr. Fabian J. Theis

2. Prof. Dr. Wolfgang Enard

Die Dissertation wurde am 21.03.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 18.07.2024 angenommen.

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to Carsten Marr for granting me the opportunity to pursue my PhD in his group and allowing me to work on such fascinating projects with numerous outstanding collaborators. Carsten's mentorship has been invaluable and his ideas, guidance and positive attitude throughout all our scientific discussions helped me grow both professionally and personally. His impact on me cannot be understated. I could not have asked for a better supervisor.

I am also deeply thankful to Fabian Theis for enabling my PhD, for his unwavering support throughout my time as a master and PhD student and for creating such an amazing, open and thriving research environment at the Computational Health Center in Helmholtz Munich.

I would further like to thank all the collaborators I had the pleasure of working with, in particular Sebastian Kobold, Christian Schröter, Micha Drukker, and Katharina Götze, for their valuable guidance and support during our meetings. Additionally, I would like to thank Wolfgang Enard and Maria-Elena Torres Padilla for being part of my thesis advisory committee, providing valuable ideas and support to my projects.

The ICB and the AIH have been a wonderful place to conduct research, surrounded by fantastic and inspiring individuals. I feel incredibly lucky to have met so many remarkable people across the Marr, Theis, and Peng labs. I would like to thank Dominik Waibel, Melanie Schulz, Benedikt Mairhörmann, Ruben Brabenec, Alexandra de la Porte, Julia Schröder, Ali Boushehri, Lea Schuh, Minas Schwager, Michael Sterr, Sophia Wagner, Lorenz Lamm, and Valentin Koch for engaging in stimulating scientific discussions, contributing their ideas to my projects, and offering continuous support. A special thanks goes to Sophie Tritschler for introducing me to the world of computational biology and providing valuable guidance at the start of my master's thesis and PhD. Equally important, I am grateful to all members of the Computational Health Center for fostering such an enjoyable working atmosphere.

To my family, I cannot express enough gratitude for your unwavering support throughout this journey leading to my thesis. You all made this journey a very enjoyable and memorable experience. I could not have done it without you. To Isabel, Jakob, Martin, Patrick, and Anabel, thank you for being by my side at all times.

And thank you, Zuleika, for showing me the meaning of success.

Abstract

Breakthroughs in single cell omics, a discipline that focuses on the analysis of individual cells, have revolutionized our understanding of complex cell biology in recent years. Technologies such as single cell RNA sequencing (scRNA Seq) and single cell sequencing for DNA accessibility (scATAC Seq), routinely allow sequencing thousands of individual cells. This generates massive amounts of data that, with the help of sophisticated computational techniques, provide unprecedented insights into the molecular underpinnings of cellular identity and function. Cellular therapies have gained increasing attention as a promising approach for the treatment of a variety of diseases, yet persistent challenges involve refining their performance while ensuring widespread scalability. In this thesis, I explore and analyze scRNA and scATAC data with the goal of contributing to the safety and efficacy of cellular therapies employing chimeric antigen receptor (CAR) T cells or pluripotent stem cells (PSCs).

A promising form of therapy using immune cells is the genetic modification of a patient's T cells to generate so-called CAR T cells. These cells can bind to cancer-specific targets and thus have an increased potential to recognize and fight cancer cells. Despite remarkable progress in the treatment of B cell lymphomas, broad applicability of CAR T cells is still hindered by a lack of cancer-specific and thus safe targets, resulting in a range of sometimes severe and life threatening side effects. I introduce a computational approach to identify potential targets for CAR T cell therapy using large scale scRNA Seq data. Applying this approach to acute myeloid leukemia, an aggressive blood cancer, led to the discovery of two previously unidentified targets that demonstrated robust efficacy with minimal toxicity, providing a strong rationale for further clinical development. In addition, I investigate global gene expression profiles of targets that are currently being tested in clinical trials. I interpret target expression in tumor cells and healthy tissues in conjunction with clinical outcome data obtained from patients who underwent CAR T cell therapy and propose novel targets that exhibit a favorable gene expression pattern across malignant and healthy cells.

PSCs have enormous potential for regenerative medicine, given their ability to differentiate into specific cell types that are required to repair damaged or destroyed tissues. As a result, they have garnered significant attention as a potential remedy for the global shortage of transplantable tissues and organs. However, clinical use is still limited due to slow and inefficient differentiation protocols of these cells and immature properties of the derived cell types. Acceleration of stem cell differentiation requires a better understanding of the basic mechanisms that determine cellular developmental time. However, these developmental times vary between species and remain poorly understood. Therefore, I use timelapsed multiomic sequencing, a method that allows us to simultaneously observe changes in gene expression and DNA accessibility over time, to characterize differentiation of PSCs from three species to neural progenitor cells. I explore distinct species-specific variations in gene expression, global differentiation rates, cell cycle phases, DNA

accessibility, transcription factor activity, and potential factors driving differentiation.

This thesis presents strategies for handling extensive transcriptomic and timelapsed multiomic data, demonstrating their potential for driving progress in cellular therapies. The findings of this thesis facilitate the identification and clinical development of novel CAR T cell therapeutics and establish groundwork to develop improved and more effective protocols for stem cell differentiation in the field of regenerative medicine.

Zusammenfassung

Bahnbrechende Fortschritte in der Einzelzell-Omik, einer Disziplin, die sich mit der Analyse einzelner Zellen befasst, haben in den letzten Jahren unser Verständnis der komplexen Zellbiologie revolutioniert. Mit Technologien wie der Einzelzell-RNA-Sequenzierung (scRNASeq) und der Einzelzell-Sequenzierung für DNA-Zugänglichkeit (scATAC Seq) können routinemäßig Tausende individueller Zellen sequenziert werden. Dadurch entstehen riesige Datenmengen, die mit Hilfe hochentwickelter rechnerischer Verfahren beispiellose Einblicke in die molekularen Grundlagen der zellulären Identität und Funktion bieten. Zelluläre Therapien haben als vielversprechender Ansatz für die Behandlung einer Vielzahl von Krankheiten zunehmend an Aufmerksamkeit gewonnen, doch die Herausforderungen bestehen weiterhin darin, ihre Leistung zu verbessern und gleichzeitig eine breite Skalierbarkeit zu gewährleisten. In dieser Arbeit untersuche und analysiere ich scRNA und scATAC Daten mit dem Ziel, zur Effektivität und Sicherheit zellulärer Therapien beizutragen, bei denen chimäre Antigenrezeptor (CAR) T Zellen oder pluripotente Stammzellen (PSZ) eingesetzt werden.

Eine vielversprechende Form der Therapie mit Immunzellen ist die genetische Modifizierung von T Zellen eines Patienten, um daraus sogenannte CAR T Zellen zu erzeugen. Diese Zellen können an krebsspezifische Zielstrukturen binden und besitzen dadurch ein erhöhtes Potenzial zur Erkennung und Bekämpfung von Krebszellen. Trotz bemerkenswerter Fortschritte bei der Behandlung von B-Zell-Lymphomen wird eine breite Anwendbarkeit von CAR T Zellen durch einen Mangel an krebsspezifischen und dadurch sicheren Zielstrukturen behindert, was sich in einer Reihe von mitunter schwerwiegenden und lebensbedrohlichen Nebenwirkungen äußert. Ich stelle einen computergestützten Ansatz zur Identifizierung potenzieller Zielstrukturen für die CAR T Zelltherapie unter Verwendung von umfangreichen scRNA Seq Daten vor. Die Anwendung dieses Ansatzes auf die akute myeloische Leukämie, ein aggressiver Blutkrebs, führte zur Entdeckung von zwei bisher unerkannten Zielstrukturen, die eine robuste Wirksamkeit bei minimaler Toxizität zeigten, was eine solide Grundlage für die weitere klinische Entwicklung darstellt. Darüber hinaus untersuche ich globale Genexpressionsprofile von Zielstrukturen, die derzeit in klinischen Studien getestet werden. Ich interpretiere die Expression dieser Zielstrukturen in Tumorzellen und gesundem Gewebe in Verbindung mit klinischen Daten von Patienten, die mit CAR T Zellen behandelt wurden, und schlage neue Zielstrukturen vor, die ein vorteilhaftes Genexpressionsmuster in malignen und gesunden Zellen aufweisen.

PSZ verfügen über ein enormes Potenzial für die regenerative Medizin, da sie in der Lage sind, sich in spezifische Zelltypen zu differenzieren, die für die Reparatur von geschädigtem oder zerstörtem Gewebe erforderlich sind. Als potenzielles Mittel gegen den weltweiten Mangel an transplantierbaren Geweben und Organen haben sie dadurch große Aufmerksamkeit erregt. Der klinische Einsatz ist jedoch aufgrund langsamer und ineffizienter Differenzierungsprotokolle dieser Zellen, sowie unausgereifter Eigenschaften der gewonnenen Zelltypen begrenzt. Die Beschleunigung der Stamm-

zelldifferenzierung erfordert ein besseres Verständnis der grundlegenden Mechanismen, welche die Entwicklungszeit der Zellen bestimmen. Diese Entwicklungszeiten sind jedoch für jede Spezies unterschiedlich und bislang nur unzureichend bekannt. Daher nutze ich Zeitraffer-Multiomik Sequenzierung, eine Methode, die es ermöglicht, gleichzeitig Veränderungen in der Genexpression und der DNA-Zugänglichkeit im Laufe der Zeit zu beobachten, um die Differenzierung von PSZ aus drei Spezies zu neuralen Vorläuferzellen zu charakterisieren. Ich untersuche ausgeprägte speziesspezifische Variationen in der Genexpression, den globalen Differenzierungsraten, den Zellzyklusphasen, der DNA-Zugänglichkeit, der Aktivität von Transkriptionsfaktoren und potenziellen Einflussfaktoren auf die Differenzierung.

In dieser Arbeit werden Strategien für den Umgang mit umfangreichen transkriptomischen und Zeitraffer-Multiomik Daten vorgestellt und ihr Potenzial für den Fortschritt bei Zelltherapien demonstriert. Die Ergebnisse dieser Arbeit erleichtern die Identifizierung und klinische Entwicklung neuartiger CAR T Zelltherapeutika und bilden die Grundlage für die Entwicklung verbesserter und effektiverer Protokolle für die Stammzelldifferenzierung im Bereich der regenerativen Medizin.

Contents

1 Introduction

Cellular therapies refer to the use of living cells for the treatment of various diseases and medical conditions. The earliest records of cellular therapy are found in the Ebers Papyrus, an ancient Egyptian medical document dating back to 1550 BC that describes the use of plant extracts to treat various ailments such as diarrhea, skin diseases, and even cancer. Similarly, blood transfusions have been practiced for centuries, with the earliest evidence dating back to ancient Greece, where Ovid mentioned the removal and replacement of "bad blood" in the seventh book of the Metamorphoses in 43 BC [1].

In more modern times, the first actual recorded blood transfusion was performed by Richard Lower in 1665 between two dogs, and between an animal and a human the following year [2, 3]. However, the practice was abandoned for many centuries, since safe transfusions required understanding and recognition of blood types, which was not achieved until the early 20th century [4]. Then, rather than simply administering blood transfusions, researchers began to show a growing interest in completely replacing a patients' damaged blood-forming cells with healthy cells. This procedure, known as bone marrow transplantation, was first employed in the treatment of leukemia, a progressive malignancy that results in the production of large amounts of abnormal or immature white blood cells from the bone marrow. The first successful bone marrow transplantation was performed by E. Donnall Thomas in 1957 on identical twins, whose cells had the same set of genetic markers [5].

However, allogeneic transplantation, i.e. transplantation of cells or tissue from a donor to a non-genetically identical host recipient posed a challenge, as the patient's immune system was a hindrance to transplantation by recognizing matter from other bodies as foreign and attempting to reject it [6]. One approach to circumvent this issue was to perform an initial eradication of both malignant and healthy blood cells using high doses of chemotherapy and radiation, destroying the functioning marrow, before replacing it with healthy donor bone marrow cells [7].

Despite this workaround and early successes in transplantations, patients would often suffer from anemia, infections, or experience cancer relapse, because the transplanted cells failed to engraft, meaning they didn't successfully migrate to the host's bone marrow and establish a new blood cell production system [8]. Furthermore, in certain cases, the donor cells themselves exhibited an immune response against the patient's body, a phenomenon known as graft-versus-host disease [9]. However, it is interesting to note that patients who suffered from graft-versus-host disease also showed the least signs of cancer relapse [10, 11], indicating that foreign engrafted immune cells effectively targeted residual cancer cells in the patient. This realization that a patient's immune system could effectively be redesigned to subsequently recognize and attack cancer led to what is now known today as the concept of adoptive cell therapy. It explores the possibility of reengineering or training immune cells to recognize and attack cancerous cells without harming healthy

cells from the host [12]. Adoptive cell therapy completely transformed the perception of drugs from synthetic or purified small molecules to living cells, expanding the boundaries of medicine as well as therapeutic manufacturing and commercialization [13].

In recent years, cellular therapies have gained increasing attention as a promising approach for the treatment of a variety of diseases, including cancer, autoimmune disorders, cardiovascular diseases and HIV [14–16]. The global cell therapy market size was valued at 21.6 billion USD in 2022 and is expected to grow at a compound annual growth rate of 14.15% from 2023 to 2030 [17]. Compared to traditional drug-based treatments, cellular therapies offer the promise of targeted and personalized therapies with potential long-lasting effects, and reduced side effects. However, the efficacy of cellular therapies is heavily dependent on the cells' ability to properly integrate with the host tissue and perform their intended function [14]. Thus, a comprehensive understanding of the biology and behavior of the cells used, as well as the development of appropriate delivery methods, is essential. Despite significant progress in the field, many challenges, such as improving the safety and efficacy of these therapies and facilitating scaling for widespread use, remain to be solved. Nevertheless, the potential benefits of cellular therapies are significant and their continued development holds great promise for the treatment of various diseases and conditions [18].

Cellular therapies can be broadly classified into autologous and allogeneic therapies. Autologous therapies use cells that are derived from the same individual receiving the treatment, whereas allogeneic therapies use cells that are derived from a usually genetically similar, but not identical, donor. Autologous therapies have the advantage of avoiding immune rejection, as the cells used are from the patient's own body, thus minimizing the need for immunosuppressive drugs and reducing the risk of graft-versus-host disease. However, autologous therapies may be limited by the availability and quality of a patient's own cells, as well as the potential for contamination or mutation during processing [14, 19, 20]. A promising approach for autologous treatments involves the use of immune cells, specifically T cells, due to their potential to recognize and target specific cells, such as cancer cells or cells infected with a virus [21]. Notably, despite commonly observed acute toxicities [22], these autologous treatments are commonly employed in the treatment of hematological malignancies such as leukemia and lymphoma, demonstrating highly encouraging results [23, 24].

On the other hand, allogeneic therapies have the potential for greater scalability and availability, as cells can be derived from healthy donors, thus offering "off-the-shelf" supply for a wider range of patients. However, this approach requires careful matching to avoid immune rejection and graft-versus-host disease [14, 19, 20]. For allogeneic cellular therapies, pluripotent stem cells, which can differentiate into various cell types, including those involved in tissue repair and regeneration, have shown great potential, offering scalable and cost-effective treatments [25]. Nevertheless, significant challenges persist, notably the potential for tumor formation, which may occur if reprogramming factors remain active or when immature cells persist in the final cell

product [25]. Despite these hurdles, ongoing clinical investigations involving pluripotent stem cells for allogeneic therapies span various clinical domains, underscoring their potential to address an array of diseases and injuries. Notable applications include age-related macular degeneration, spinal cord injury, Parkinson's disease, and type 1 diabetes [25].

In the upcoming sections, I will present the utilization of T cells for autologous cellular therapies, as well as pluripotent stem cells for allogeneic cellular therapies.

1.1 Somatic cell-based cellular therapies for cancer treatment

Non-stem cell-based therapies encompass a wide range of treatments that utilize somatic non-stem cells, i.e. cells that make up the body of an organism and are not involved in reproduction, such as immune cells, pancreatic islets cells, or fibroblasts. These cells are typically obtained either from healthy donors or from the patient themselves, and are then subjected to various forms of manipulation or treatment before being administered to the patient [14]. Of the cell types used in non-stem cell-based therapies, pancreatic islet cell transplantation has shown particular promise as a treatment for patients with type 1 diabetes [26, 27].

Another emerging area of non-stem cell therapies is adoptive cell therapy, which involves the modification of immune cells, such as T cells, followed by their infusion into patients to elicit an immunologic response against tumors [28, 29]. The most prominent form of adoptive cell therapy is chimeric antigen receptor (CAR) T cell therapy, which I will describe in more detail below.

1.1.1 CAR T cell therapy for treatment of hematological malignancies

The recognition of foreign proteins by mature T cells is mediated by receptors located on their cell surface, which typically suffices to eradicate malignant cells or viruses [30]. However, cancer cells are characterized by their ability to evade, suppress and exploit the host immune system, enabling them to grow and spread rapidly [31]. The genetic modification of a patient's T cells to express a CAR has evolved as a promising strategy for enhancing the immune response against cancer cells. These engineered CARs then redirect the T cells' specificity and function, enabling them to recognize and selectively target tumor-specific antigens. Following ex vivo expansion, these CAR T cells are infused back into the patient, where they selectively migrate to sites of tumor growth and exert their cytotoxic effects by recognizing and killing cancer cells [21] (Figure 1.1a).

CAR T cell therapy has undergone significant improvements over the years, as evident from the successive generations of CAR structures themselves, resulting in substantial enhancements in the efficacy and safety of CAR T cells [32]. Characterized by improved CAR T cell expansion, modulation of the antigen density threshold for activation, and increased overall persistency, these advancements have led to the development of "living drugs" that form the backbone of modern CAR T cell therapy today [32, 33].

Figure 1.1b illustrates the components of a typical CAR. At its core, a baseline or prototype CAR comprises an ectodomain consisting of a single-chain variable fragment (scFv) to specifically recognize and bind to the tumor antigen. This is followed by a hinge or spacer domain that provides conformational flexibility and projects the scFv away from the cell surface. The transmembrane domain anchors the receptor in the T cell surface. Costimulatory and CD3ζ signaling domains, which promote CAR T cell expansion and persistence, and provide downstream activation signals, have been shown to be significant contributors to the overall efficacy of CAR T cell therapy [32].

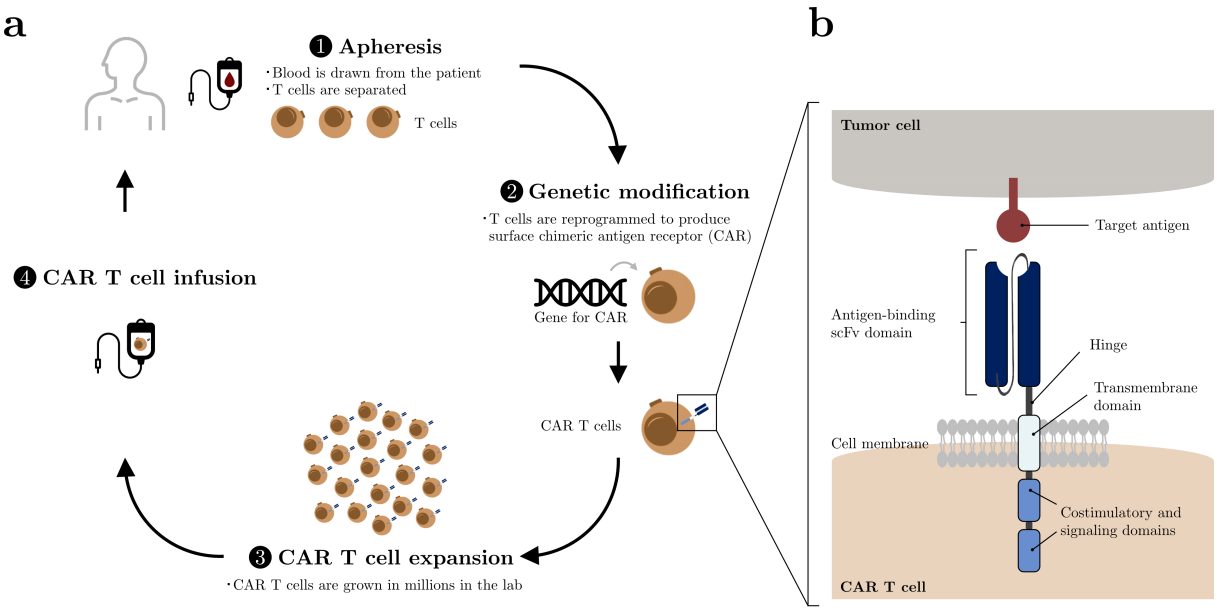

Figure 1.1: An overview of CAR T cell therapy. a) The workflow of CAR T cell therapy: Patient's T cells are removed from the blood (1) and genetically engineered to express the gene for a chimeric antigen receptor (CAR) (2). CAR T cells are subsequently expanded in the laboratory (3), before being infused back into the patient (4). **b)** Prototype structure of a CAR including the antigen-binding single chain variable fragment (scFv) domain, a hinge domain, a transmembrane domain and intracellular costimulatory and signaling domains.

Since their inception in 1987 [34], the field of CAR T cell therapy has rapidly evolved from a promising immuno-oncology approach in preclinical models to over 1000 clinical trials testing CAR T cells (registered on http://www.clinicaltrials.gov as of March 2023). Despite these advancements, the substantial cost, reaching several hundred thousand USD per patient, remains a formidable barrier to the broader administration of CAR T cell therapy [35, 36]. Nevertheless, six FDA-approved and commercialized CAR T cells targeting B cell lineage antigens CD19 or B cell maturation antigen (BCMA) have demonstrated clinical efficacy in patients suffering from various B cell malignancies, including B cell lymphoma, B cell acute lymphoblastic leukemia, and multiple myeloma [21, 32, 37–40].

1.1.2 Lack of safe targets prevents broad clinical application of CAR T cells

Despite the remarkable progress made in the treatment of B cell lymphomas, CAR T cell therapy is associated with acute toxicities. While some toxicities were expected, such as B cell aplasia, occurring when B cell-specific antigens like CD19 or BCMA are targeted [21], many unforeseen toxicities emerged, which, importantly, were retrospective findings, discovered only after treatment inception [41–43]. The two most commonly observed acute adverse effects of CAR T cell therapy are cytokine release syndrome and neurologic toxicity.

Cytokine release syndrome is a systemic inflammatory response caused by the release of large amounts of proinflammatory cytokines by activated CAR T cells [33]. Upon encountering their target cells, CAR T cells release cytokines, such as interleukin-6 (IL-6), tumor necrosis factor-alpha (TNF-α), and interferon-gamma (IFN-γ), which activate other immune cells, leading to the release of additional cytokines. This cytokine cascade can quickly spiral out of control [44], resulting in a wide range of symptoms from minor constitutional symptoms, such as fever, to long-term and potentially life threatening consequences such as multi-organ failure [42]. The severity of cytokine release syndrome varies widely between patients and is dependent on several factors, such as the extent of tumor burden and the dose of CAR T cells administered [21, 23].

Neurotoxicity, on the other hand, despite being reported many times [23, 45, 46], remains poorly understood. It is believed to involve the disruption of the blood-brain barrier due to inflammation caused by increased cytokine levels [47–49], which can lead to a wide range of symptoms such as confusion, delirium, expressive aphasia or seizures [43].

The origins and management of these toxicities are not yet well understood in the context of CAR T cell therapy, although it is commonly believed that these adverse effects are a consequence of the interaction between CAR T cells and the cells they target. As the efficacy of these therapies heavily depends on targeting tumor-specific antigens, these side effects have been partially linked to target expression and availability [43].

An ideal target antigen serves as a vital survival signal for the tumor clone and is restricted to the tumor cell [43, 50]. Conversely, shared expression and subsequent engagement of target antigens on nonpathogenic tissues is commonly believed to increase the risk for these so-called on-target, off-tumor toxicities [51–53]. For instance, the use of CAIX (CA9)-targeting CAR T cells for the treatment of renal cell carcinoma resulted in hepatobiliary toxicity, likely due to shared expression of CAIX on bile duct epithelium [54, 55]. Similarly, administration of CAR T cells targeting HER2 in a patient with colorectal cancer led to cardiorespiratory failure due to cross-reactivity with pulmonary and cardiac tissue [56, 57].

Identifying and selecting safe target antigens is thus pivotal to translate the vast potential of CAR T cell therapy to non-B cell malignancies and optimize long-term patient management and outcomes.

1.2 Pluripotent stem cells for regenerative medicine

Stem cells are undifferentiated cells capable of self-renewal and differentiation into various cell types depending on their developmental potency [14, 58]. Pluripotent stem cells (PSCs) are a type of stem cell that can differentiate into cells of all three germ layers, i.e. layers of cells formed during embryonic development that develop into different parts of the body. PSCs can therefore give rise to almost all cell types in the body, except for extraembryonic structures like the placenta [58]. PSCs can be derived from different stages of biological development [59] (Figure 1.2).

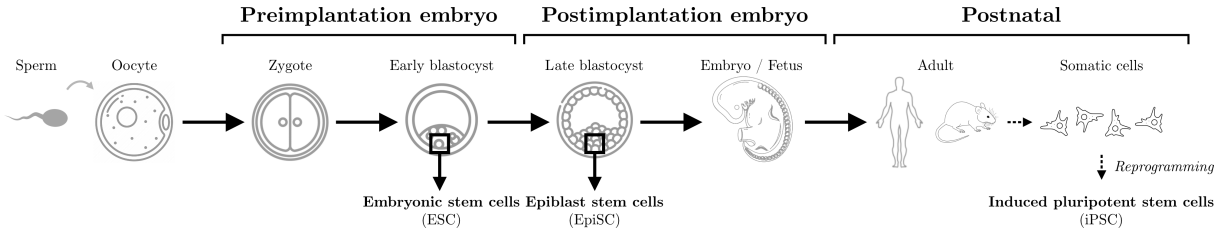

Figure 1.2: The origin of PSCs during development. Pluripotent cells of various types can be obtained by explanting or reprogramming cells at different stages of development, ranging from early embryonic stages to postnatal adulthood.

Embryonic stem cells (ESCs) are derived from the early inner blastocyst cell mass of preimplantation embryos. ESCs express key pluripotency genes and have the ability to differentiate into any cell type in the body, providing an unlimited supply of cells for therapeutic purposes. In contrast, epiblast stem cells (EpiSCs) are derived from the epiblast layer of epithelial cells in postimplantation embryos and exhibit a slightly more limited developmental potential compared to ESCs [59]. However, the use of embryonic and epiblast stem cells is controversial due to ethical concerns related to the use of human embryos and endangering fetal lives [60].

Given these concerns, a major breakthrough occurred in 2006, when researchers demonstrated that mature and differentiated cells could be reprogrammed back into stem cells [61, 62]. These reprogrammed cells are known as induced PSCs (iPSCs) and can be generated from various adult somatic cells in vitro. iPSCs exhibit similar characteristics to ESCs and provide a promising alternative, with the added advantage of not raising ethical concerns [59, 63]. Since then, iPSCs are becoming increasingly attractive for their potential to generate personalized therapies, with many different iPSC lines customized for specific applications [64].

Given their ability to differentiate into various cell types of the human body, including specific cell types that are required to repair damaged or destroyed tissues, PSCs hold enormous potential for regenerative medicine, providing a way for cell differentiation or tissue repair and have garnered significant attention as a potential remedy for the global shortage of transplantable tissues and organs [64]. Steady advancements in cell culture and differentiation techniques have expanded the scope of PSCs in cellular therapies and their potential impact on human health [25, 65],

facilitating promising treatments for various diseases, including type 1 diabetes, where PSCs can be differentiated into insulin-producing beta cells, potentially replacing dysfunctional insulin-secreting beta cells [66–68], heart failure [26, 69], where PSCs can be differentiated into mature cardiomyocytes, Parkinson's and Alzheimer's disease or strokes [70].

The generation of patient-specific iPSC lines has revolutionized stem cell research by providing a customizable and biocompatible platform, thus become indispensable for disease modeling and high-throughput screening for drug discovery and toxicity tests [71–73]. However, the clinical use of PSCs for cell differentiation or tissue repair is still limited to investigational regenerative medicine, as their translational potential is hindered by certain limitations, such as the slow and inefficient differentiation process, as well as the immature characteristics of derived cell types [59, 63, 65, 71, 74, 75]. The yield of functional cells is usually quite low, with major yield loss often being reported during later stages of the differentiation process. For example, for cell types with complex functions, such as insulin-producing beta cells, the majority of yield loss is often encountered at later stages, impacting the efficiency and success of the therapeutic outcome [76, 77]. The overall efficiency of the differentiation process can also heavily be influenced by its duration, as PSC differentiation into specific cell lineages often involves multiple steps, and each transition may introduce inefficiencies that contribute to yield loss. Differentiation of PSCs into functional cell types often takes several weeks [59, 75]. Consequently, there is a growing interest in accelerating the differentiation of PSCs to produce functional cell types more rapidly and efficiently [14, 75].

The complex process of PSC differentiation can be better understood by studying mammalian developmental timescales underlying cellular differentiation. Exploring beyond human timescales is essential, as comparative studies across species can provide insights into evolutionarily conserved and species-specific aspects of PSC differentiation. Analysis of factors that influence cellular development can facilitate a more efficient and expeditious differentiation of PSCs into desired cell types, contributing to the development of more robust and universally applicable therapies. As such, studying developmental timescales is critical for realizing the full potential of PSCs in cellular therapies [78].

1.2.1 Species-specific differentiation timescales

Mammalian embryonic development follows a defined sequence of events, although the duration of these events varies widely among different species. For instance, while it takes 13 days for human embryos to proceed from oocyte fertilization to gastrulation, mice reach this stage in just six days [79]. Furthermore, the timescales of organogenesis, as well as neuronal differentiation in the peripheral nervous system and midbrain are notably longer in humans than in mice [80].

In vivo developmental speed is reflected by the pace of in vitro differentiation of PSCs and remains remarkably species-specific [78, 81]. In addition, the segmentation clock, a molecular

process governing the rhythmic generation of repeating structures, such as precursor cells in the spine, oscillates at a 2.5 and 5-hour periodicity in mouse and human PSC differentiation models, respectively [82]. These findings suggest that cell differentiation timescales have a genetic basis that is intrinsic to the cells themselves.

However, this intrinsic timescale can be modified by extrinsic factors. For example, reducing the number of cells in mouse embryos through pharmacological or mechanical methods slows down differentiation compared to proliferation during gastrulation [83, 84]. Additionally, diapause allows many mammals to halt development when environmental conditions are unsuitable for prenatal or postnatal survival [85]. Moreover, the creation of interspecies chimeras from pluripotent cells of different species suggests that external signals can coordinate cell differentiation timescales [75, 86, 87].

Taken together, these observations indicate that genetically determined cell-intrinsic mechanisms establish species-specific developmental timescales that can be influenced by both evolutionary and environmental factors. Recent studies propose that differences in biochemical reaction rates may contribute to interspecies differences between mouse and human [78, 81, 88]. However, despite the growing attention in this field, the molecular mechanisms governing these timescales, particularly beyond human and mouse development, remain largely elusive. PSCs represent an invaluable experimental tool for elucidating the mechanisms underlying species-specific differentiation timelines. The epiblast-like primed pluripotent state, which is shared by human ESCs, mouse EpiSCs, and iPSCs derived from non-human and human primates [89, 90], provides a common starting point for comparative studies. By inducing differentiation of primed PSCs in vitro to generate developmental intermediates and potentially somatic-like cells, it becomes feasible to compare the dynamics of cell state transitions across diverse species.

Following the extensive discussion of two prominent forms of cellular therapies, it becomes evident that understanding cellular dynamics at a molecular level is paramount. Transitioning from the exploration of PSCs and interspecies developmental variations, the following section now shifts focus to the fundamental unit of life: the cell.

1.3 The flow of genetic information within a cell

The concept of a cell as the fundamental unit of life was first introduced in the mid-17th century by Robert Hooke. Using a primitive microscope, he observed tiny structures resembling small rooms in a cork slice and named them "cells", derived from the Latin word "cellula" [91]. The classical cell theory was subsequently extended to the animal kingdom by Theodor Schwann in 1839, who proposed cells as the basic units of life, further stating that all organisms are made of cells and that these cells come from preexisting cells that have multiplied [91]. Animal cells, or eukaryotic cells, have a complex structure that enables them to perform various functions (Figure 1.3a). At the core of every animal cell is the nucleus, which contains the genetic material in the form of DNA, which is organized into structures called chromosomes. As it regulates gene expression and is involved in various cellular processes such as DNA replication and repair, the nucleus is considered the control center of the cell [92]. The nucleolus, a subregion within the nucleus, is involved in the production of ribosomes, the cellular structures responsible for protein synthesis. Surrounding the nucleus is the cytoplasm, a gel-like substance that fills the cell and contains various organelles, such as mitochondria, which are responsible for generating energy through cellular respiration. Mitochondria contain their own DNA and are able to replicate independently of the cell. The plasma or cell membrane is a flexible, selectively permeable barrier that encloses the cell, separates its internal environment from the external environment and regulates the movement of substances in and out of the cell [92].

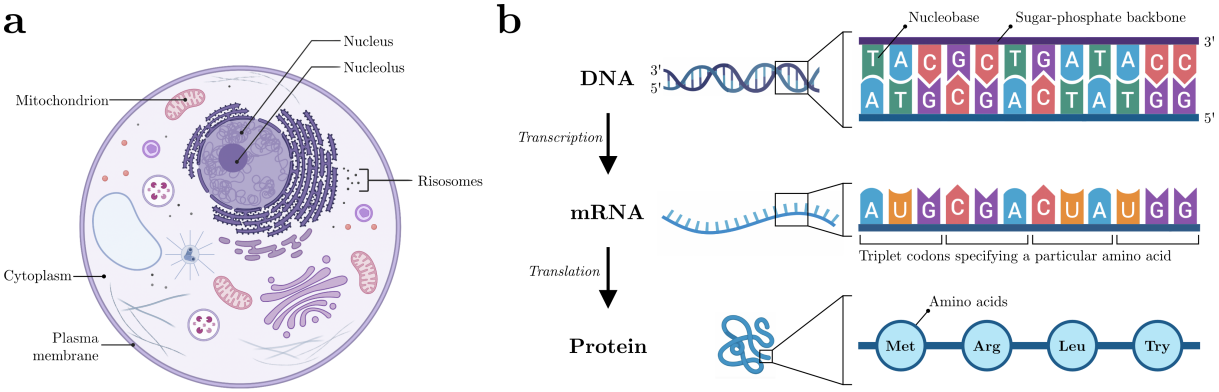

Figure 1.3: The molecular basis of life. **a)** Simplified structure on an animal cell. **b)** The central biological dogma illustrating the flow of genetic information from DNA to mRNA to protein. Adapted from the templates "Structural Overview of an Animal Cell" and "Central Dogma" from BioRender.com, retrieved from https://app.biorender.com/biorender-templates (2023).

In the 19th century, Gregor Mendel introduced the concept of genes as the mechanism of inheritance through his experiments with pea plants, discovering that specific traits were inherited in a predictable manner. Mendel's work on genetics laid the foundation for our understanding of heredity and the transmission of traits from one generation to the next and set the stage

for further investigations into the nature and function of genes. However, the discovery of the molecular structure of DNA as a double stranded helix in 1953 by Watson and Crick [93] kicked in the true golden age of molecular biology. The DNA double helix is composed of strands of nucleotide molecules linked by phosphate groups. Each nucleotide consists of a deoxyribose sugar, a phosphate group and a nucleobase. In DNA, there are four different types of nucleobases: adenine (A), cytosine (C), guanine (G) and thymine (T). Hydrogen bonds between matching nucleotides (A-T and G-C) hold together the two strands, hence DNA is usually measured in base-pairs (bp).

In the central dogma of molecular biology [92], the flow of information within a cell is traditionally displayed from double stranded DNA to single stranded RNA to protein (Figure 1.3b). The DNA hereby serves as a blueprint for gene and protein expression and stores long-term genetic information in the cell's nucleus and is constantly being renewed, a process called DNA replication. Transcribing regions of DNA gives rise to single stranded RNA. While there are many types of RNA molecules of different length, properties and function [94], the messenger RNA (mRNA) ultimately gets translated into a protein. Before protein production however, immature mRNA molecules are subject to post-transcriptional modifications. These include a polyA extension at the end of the molecule and removing regions of RNA that do not code for proteins, leading to a mature, functional RNA molecule that can then be transported out of the nucleus to the ribosomes, where protein synthesis takes place.

Discoveries in cellular complexity and function in the last 50 years ushered in a new understanding of cellular and molecular complexity. Sequencing technologies have allowed researchers to accurately identify the key nucleotide sequences within DNA, providing a fundamental attribute by which all life forms can be delineated and distinguished from one another. Widespread sequencing technologies gained popularity in the late 1970s with the Sanger Sequencing protocol, utilizing the process of DNA amplification using modified, labeled nucleotides, enabling accurate characterization of each base in the DNA [95, 96]. While facilitating a then revolutionary understanding of molecular biology, the Sanger sequencing method was limited in throughput due to its labor-intensive nature, requiring multiple rounds of labeling, sequencing, and separation of DNA fragments, making it difficult to sequence large numbers of samples quickly and efficiently [97].

Since the early 2000s, steady progress in both development and subsequent analysis of cellular sequencing, largely impacted by the Human Genome Project [98], has given rise to next generation sequencing (NGS) assays [99, 100]. These technologies perform millions of parallel sequencing reactions on a micrometer scale, thus heavily decreasing the sequencing cost per base, and in turn facilitating high throughput sequencing of larger genomes [100]. Although these assays typically produce shorter read fragments (usually up to 500 bases) at an increased error rate, their high throughput allows for massive sequence coverage of millions of short DNA fragment reads to construct a best-fit consensus sequence of DNA.

This 'genomics revolution' [100] driven predominantly by these advancements DNA sequencing technology, has significantly transformed the cost and accessibility of DNA sequencing. The expense associated with sequencing the human genome witnessed a remarkable descent, plummeting from an estimated 10 million USD in 2007 to 10,000 USD in 2012, and currently standing at less than 1,000 USD [101, 102]. DNA sequencing capability and accessibility have increased at a rate even higher than the computing revolution popularly described in Moore's law, doubling roughly every five months between 2004 and 2010 [101].

1.4 Understanding identity and function of single cells

Traditionally, these sequencing assays were performed in bulk, averaging the obtained readout from millions of cells in a sample, thereby masking cellular heterogeneity and molecular programs [103–105]. Conversely, single cell omics technologies measure a variety of possible features, such as gene expression or DNA accessibility at a single cell level, enabling the study of cellular identity, heterogeneity, development and disease with unprecedented resolution [105]. In recent years, rapid technological advances have enabled genome-wide profiling of RNA, DNA, and protein at the single cell level, with an ever-increasing scale, now routinely covering thousands or even millions of single cells [106, 107].

With the emergence of high throughput single cell omics technologies, it is now feasible to create reference maps of human tissues, known as atlases [108]. These atlases act as comprehensive catalogs of cellular profiles and provide valuable insights into the diversity and function of different cell populations within tissues and organs [104], enabling an accurate dissection of cellular heterogeneity and identification of novel cell types [109, 110].

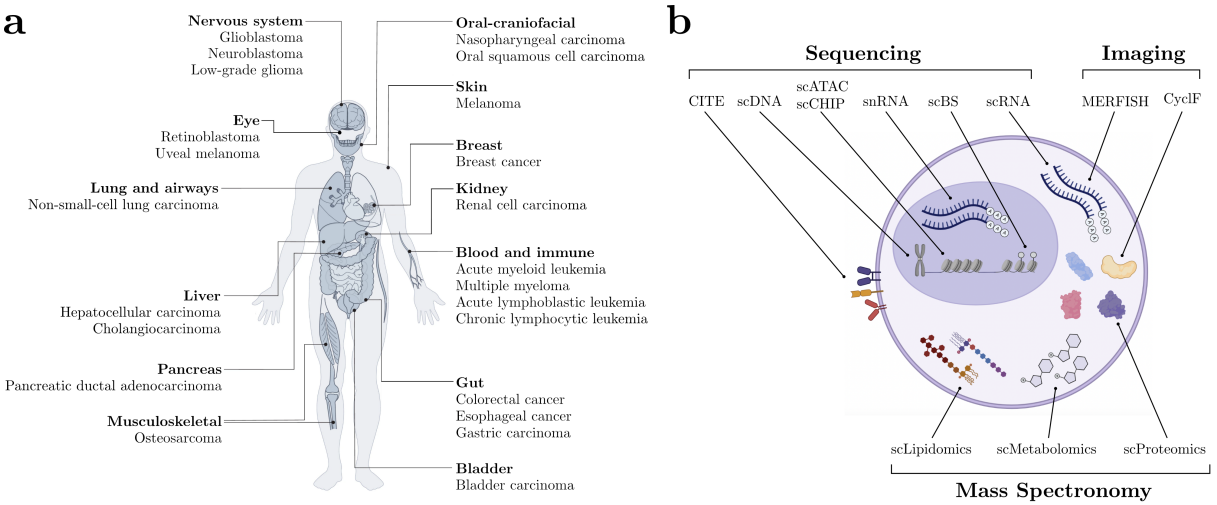

Figure 1.4: Advanced single cell technologies provide insights into the molecular complexity of human health and disease. a) Key organs and biological systems, as well as associated cancer types, for which the Human Cell Atlas initiative and related studies have compiled large-scale single cell atlases in both healthy and diseased patients. **b)** Various technologies and data modalities are employed to approximate cellular identity, leveraging distinct molecular phenotypes at the single cell level. Adapted from Rood et al. [104] and Jackson & Vogel [111].

Additionally, single cell atlases have played a crucial role in advancing our understanding of disease biology. By analyzing cell atlases from patients with diseases in comparison to healthy references, researchers can gain insights into the intricate biological complexities either on a cellular population level [112, 113] or by identifying disease-related genes and gene programs in which they participate [114, 115], further enabling discovery of biomarkers for disease diagnosis and

prognosis [104] (Figure 1.4a). In the context of cellular therapies, this allows for the identification of regenerative mechanisms that are utilized as therapeutic targets or for improving engineered cell therapies by defining and comparing the desired target state from human-derived models to healthy human tissue [104].

The advent of atlases and the availability of numerous single cell datasets have significantly transformed our understanding of cellular identity and function. Traditionally, cells have been classified into cell types using phenotypic measurements such as gene or protein expression or morphology [116]. However, recent technological advances in sequencing, imaging, and mass spectrometry have allowed the characterization of various aspects of cellular identity and function, including genomic variations (such as single nucleotide variation and copy number variation), transcriptomic measurements (such as transcript sequence and abundance), epigenomic measurements (such as chromatin accessibility and conformation, DNA methylation, and histone modification), and proteomic measurements (such as protein sequence and abundance) [116] (Figure 1.4b). Importantly, these different omics layers are connected with each other. For instance, mRNA expression only has measurable functional potential when it is ultimately translated into a protein. Furthermore, mRNA expression is heavily regulated by epigenetic mechanisms such as histone modifications, which modulate DNA accessibility.

Therefore, approaches that integrate these diverse sources of information, so-called multiomic technologies, offer a more comprehensive understanding of cellular identity, function, and disease-related shifts, facilitating new prospects of causal and mechanistic biology, ultimately approximating cellular identity and function in much higher detail and resolution [116]. Through this expanded the range of features for cell type classification, traditional classification schemes that delineate distinct cell types shifted towards a more continuous representation of cellular states [117, 118]. Consequently, molecular cell identity should not be perceived as a fixed set of characteristics obtained from a single measurement, but rather as a combination of different omics [118].

The subsequent sections will provide an overview of two widely employed single cell omics techniques: single cell RNA sequencing (scRNA Seq), which characterizes gene expression profiles at the single cell level, and single cell assay for transposase-accessible chromatin sequencing (scATAC Seq), which captures chromatin accessibility patterns in individual cells.

1.4.1 Profiling gene expression at the single cell level

Since all cells in our body share nearly identical genetic material, knowledge of an organism's DNA sequence and regulatory element positions provides limited insight into the intricate and dynamic operations that occur within a cell. Consequently, researchers have increasingly focused on profiling the cellular transcriptome, which offers greater insights into cellular protein production. The first transcriptomic sequencing of a single cell was reported in 2009 [119]. Since then,

significant technological advances have reduced the required input volume and increased both the number of cells and the sensitivity of readouts [120, 121].

Although numerous scRNA Seq platforms and protocols have been developed in the meantime [122, 123], the fundamental steps remain similar. Initially, cells are isolated and lysed, followed by capture of cellular mRNA and its reverse transcription into double-stranded cDNA while simultaneously barcoding each molecule with a unique cell identifier. Certain protocols permit further barcoding of each individual molecule using a unique molecular identifier (UMI). Finally, cDNA is amplified and the labeled cells and/or molecules are pooled to construct a library for next-generation sequencing. In the subsequent section, I will provide an overview of the most widely used commercial droplet-based scRNA Seq platform from 10X Genomics (Figure 1.5).

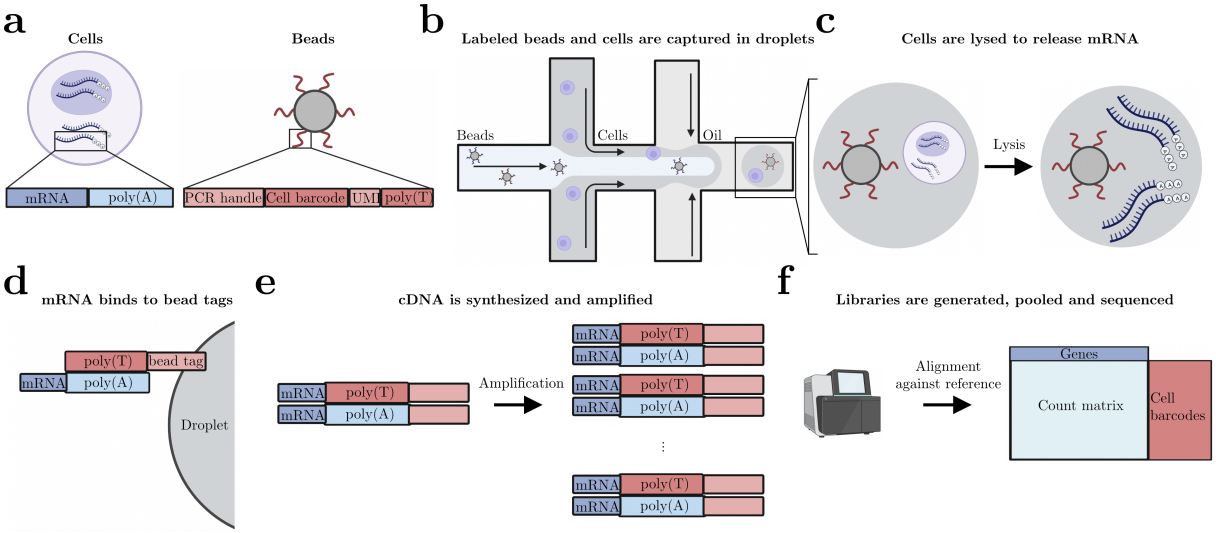

Figure 1.5: An overview of droplet-based single cell RNA sequencing from 10X Genomics. a) Cells and tagged microbeads as input to droplet-based sequencing protocols. **b-f)** Experimental workflow: Labeled beads and cells are captured in droplets **(b)**. Cell lysis releases mRNA, which binds to oligonucleotides on the microbeads **(c-d)**. cDNA is synthesized from mRNA and subsequently amplified before libraries are generated, pooled and sequenced **(e-f)**. Adapted from the template "CITE-seq Workflow" from BioRender.com, retrieved from https://app.biorender.com/biorender-templates (2023).

The scRNA Seq platform from 10X Genomics employs microfluidic technology to partition individual cells into nanoliter-sized droplets that capture cells along with a bead coated with oligonucleotides. These oligonucleotides consist of a distinct cell barcode, UMIs, and a poly(T) sequence that captures mRNA molecules (Figure 1.5a). Upon lysis of captured cells within the droplet, mRNA binds to the poly(T) sequence of the oligonucleotides on the bead (Figure 1.5b-d). Within each droplet, mRNA is reverse transcribed into cDNA. After amplification of the cDNA, droplets that contain cDNA from each cell are pooled together and subjected to NGS sequencing (Figure 1.5e-f). The use of unique cell barcodes and UMIs enables precise identification and

quantification of gene expression in individual cells, even at low sequencing depth.

In order to understand the intricate landscape of cellular dynamics and its implications for cellular therapies, the integration of additional omics approaches becomes pivotal. While traditional single cell techniques, such as scRNA Seq, provide valuable insights into gene expression, they fall short in capturing the comprehensive regulatory mechanisms governing cellular behavior. Therefore, scATAC Seq, the most popular approach to profile chromatin accessibility at the single cell level, will briefly be outlined below.

1.4.2 Profiling chromatin accessibility at the single cell level

Cellular heterogeneity does not solely arise from gene expression information, but also from the regulation of gene expression through epigenetic modifications of the genome. As all cells start with the same genetic background, chromatin must be highly adaptable to facilitate changes in the accessible regions of DNA for flexible transcription. Accessibility of sequences within the genome can therefore be viewed as a mark of genomic activity, representing the expression of particular genes or the openness of specific sequences, including enhancers or transcription factor binding sites [124, 125].

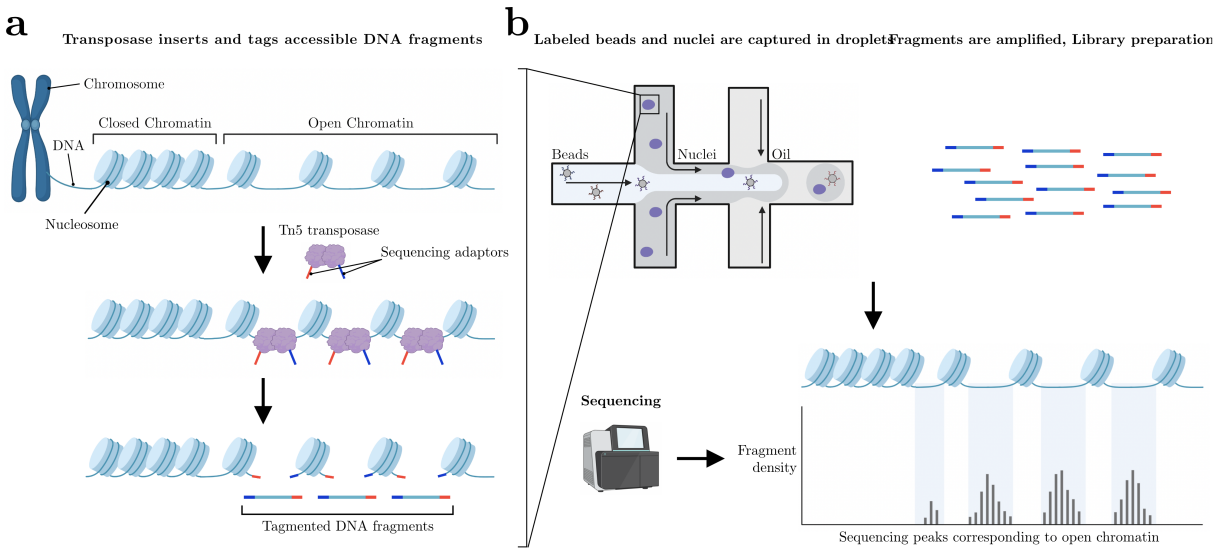

Figure 1.6: Profiling chromatin accessibility with droplet-based single cell ATAC sequencing from 10X Genomics. a) Transposase enzyme inserts and tags accessible regions of DNA. **b)** Experimental workflow: Labeled beads and nuclei are captured in droplets and subject to transposase tagmentation. Fragments are amplified, a sequencing library is prepared and sequenced. Resulting fragments represent regions of open chromatin. Adapted from the templates "ATAC Sequencing" and "CITE-seq Workflow" from BioRender.com, retrieved from https://app.biorender.com/biorender-templates (2023).

In recent years, researchers have utilized transposase enzyme activity to profile chromatin accessi-

bility across single cells, which randomly fragments and tags chromatin DNA and subsequently inserting adapters for sequencing (Figure 1.6a). As certain regions of DNA are wrapped around histone proteins, these regions are not physically accessible for the enzyme and cannot be fragmented. Therefore, resulting fragments represent accessible regions in chromatin DNA. Since its conception in 2015 [126], numerous variations of the original protocol have emerged. Similar to scRNA Seq, I will briefly outline the commercially available scATAC Seq protocol from 10X Genomics (Figure 1.6b).

The scATAC Seq workflow begins with cell isolation, followed by the release of the nuclei via cell lysis. The isolation of nuclei is critical to ensure the capture of chromatin DNA and avoid extraneous sources of DNA, such as mitochondrial DNA. Next, nuclei are incubated with a Tn5 transposase enzyme containing sequencing adapters that cuts and inserts sequencing adapters into accessible regions of chromatin DNA, resulting in tagged fragments of chromatin DNA. These fragments are then encapsulated into droplets along with barcoded oligonucleotides that contain a cell barcode and a UMI. Similar to scRNA Seq, each droplet contains a unique combination of chromatin fragments that can be attributed to a specific cell. The droplets are subsequently pooled and PCR amplified to create a sequencing library for subsequent NGS sequencing.

Single cell sequencing technologies, such as scRNA Seq and scATAC Seq, have revolutionized our ability to investigate cellular heterogeneity. Simultaneous profiling of gene expression and chromatin accessibility on the same cells has enabled a more comprehensive characterization of the entire cell state (see chapter 5).These approaches have not only expanded our fundamental understanding of cellular biology but also hold great promise in improving the efficacy and safety of cellular therapies.

1.5 Research questions and contributions

The advent of single cell technologies has revolutionized the field of cellular biology, offering unprecedented opportunities to probe the molecular underpinnings of cellular identity and function in health and disease.

CAR T cell therapy, while promising in B cell malignancies, faces challenges in translating its success to other malignancies due to a lack of safe targets, evident from the observed toxicities and limited efficacy of CAR T cells (see section 1.1.2). Current target identification strategies suffer from their retrospective nature, where safety considerations often follow efficacy assessments. In addition, with a rapid increase in clinical trials, careful consideration of the safety and potential risks associated with target antigen selection is essential before initiating clinical testing.

Despite offering a customizable and biocompatible platform, the translational potential of PSCs for cell differentiation or tissue repair remains constrained by limitations such as slow and inefficient differentiation processes and immature characteristics of derived cell types (see section 1.2). Accelerating the differentiation of PSCs to yield functional cell types more rapidly and efficiently is of great interest, but requires a fundamental understanding of developmental timescales underlying cellular differentiation.

The present thesis addresses the following research questions:

1. Can we use single cell RNA sequencing data to identify new targets for CAR T cell therapy?
2. Can we examine the gene expression profiles of current CAR targets as a means of guiding target selection before clinical translations?
3. Can we profile gene expression and chromatin accessibility from single PSCs from diverse mammalian species undergoing neural progenitor differentiation to characterize species-specific developmental timing?

In the context of cellular therapies, I used scRNA Seq data across healthy and malignant cell populations to identify and evaluate putative therapeutic targets for CAR T cell therapy. Additionally, I characterize features of species-specific developmental timing during neural progenitor differentiation, thus establishing groundwork for optimizing stem cell differentiation protocols for regenerative medicine applications (Figure 1.7b).

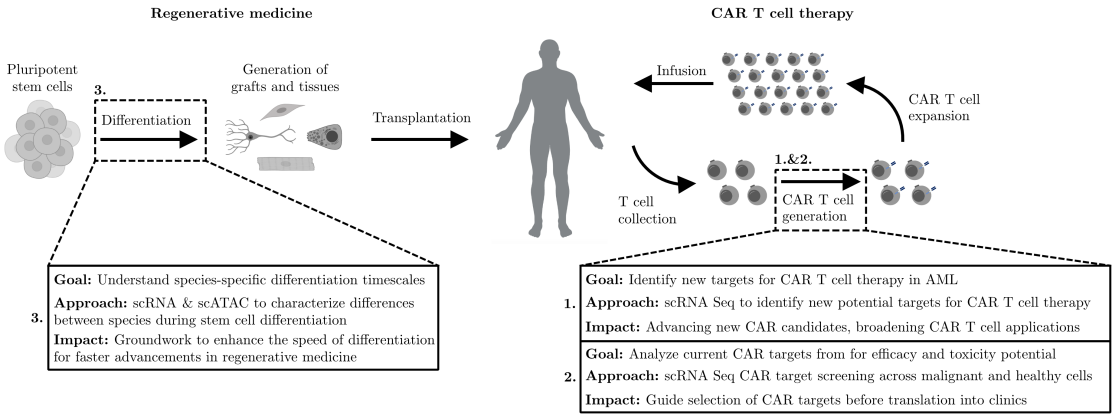

Figure 1.7: Overview of thesis contributions to cellular therapies. This thesis presents a novel approach for identifying new targets for CAR T cell therapy (1), analysis of current CAR targets on malignant and healthy cell populations (2), and characterization of gene expression and chromatin accessibility during PSC differentiation (3).

To address the first question, I make the following contributions:

1.1 I present a computational methodology to identify potential targets for CAR T cell therapy utilizing large-scale single cell gene expression data from malignant and healthy cell populations. Addressing a critical gap in the field, the approach overcomes the scarcity of safe targets for CAR T cell therapy, thus facilitating broader therapeutic application.

1.2 I validate the approach using acute myeloid leukemia (AML) data, which led to the discovery of two previously unidentified target epitopes for CAR T cell therapy in AML: CSF1R and CD86. The discovery underscores the benefit of a criteria-based target identification approach.

1.3 I benchmarked the two identified target antigens CSF1R and CD86 on a transcriptomic level to the reference genes CD123 and CD33 using two independent AML cohorts consisting of over 100,000 single cells from 20 individuals and over 500,000 single cells from vital healthy human tissues.

1.4 In collaboration with Adrian Gottschlich and Ruth Grünmeier under supervision of Prof. Sebastian Kobold, functional validation of these CAR T cells was carried out in vivo and in vitro. Established CAR T cells demonstrate robust in vitro and in vivo efficacy with minimal off-target toxicity toward relevant healthy human tissues, providing a strong rationale for further clinical development.

The corresponding manuscript is published in *Nature Biotechnology* [127] *(shared first authorship).*

To address the second question, I make the following contributions:

2.1 In collaboration with Ruben Brabenec and Lisa Gregor under supervision of Prof. Sebastian Kobold, we screened current clinical trials and scientific literature to obtain approved and investigational CAR targets in clinical trials as well as clinical outcome data obtained from patients who underwent CAR T cell therapy.

2.2 To assess the specificity of CAR targets on tumor cells for CAR-T cell therapy, I characterize CAR target expression in over 300,000 single cells obtained from patients suffering from follicular lymphoma, multiple myeloma, and acute lymphoblastic leukemia between cell populations and individual patients.

2.3 To identify potential patterns that might indicate potential risks for on-target, off-tumor toxicities, I analyze differences in global CAR target expression using a transcriptional atlas of over 3 million single cells from 35 healthy tissues throughout the human body.

2.4 In conjunction with clinical outcome data obtained from patients who underwent CAR T cell therapy, I interpret differences between CAR target expression profiles on malignant and healthy cell populations.

2.5 I identify novel CAR targets for the treatment of follicular lymphoma, multiple myeloma, and acute lymphoblastic leukemia that exhibit a favorable gene expression pattern across both malignant and healthy cell populations using a computational screening approach based on single cell gene expression data.

The corresponding manuscript is in preparation *(shared first authorship)*.

To address the third question, I make the following contributions:

3.1 In collaboration with Alexandra de la Porte under supervision of Prof. Micha Drukker, we generated PSCs from human, mouse, and cynomolgus monkey (macaca), established standardized cell culture conditions and a common protocol for neural progenitor cell differentiation for all species. RT-PCR and immunostaining results performed by Alexandra de la Porte and Julia Schröder under supervision of Dr. Christian Schröter, respectively, confirmed successful differentiation of PSCs across all three species.

3.2 To characterize global species-wide differences during neural progenitor cell differentiation, I analyzed timelapsed single cell multiomic sequencing data from over 75,000 differentiating PSCs from human, mouse, and macaca over a ten-day differentiation course utilizing combined single cell gene expression and single cell chromatin accessibility profiling.

3.3 I characterize and compare species-specific timescales underlying neural progenitor differentiation on various levels, including gene expression, cell cycle distribution, chromatin accessibility and transcription factor activity.

3.4 I characterize species-specific variations in cellular stages and global differentiation rates during neural progenitor development using single cell gene expression data and identify TF regulators that exhibit similar dynamics between gene expression and TF motif accessibility, serving as putative drivers of cellular differentiation.

The corresponding manuscript is in preparation *(shared first authorship).*

As parts of my contributions have already been published in peer-reviewed journals or are in the process of review or submission, sections 3, 4 and 5 of this thesis correspond to or are to some degree identical to the following publications:

Core publications (note that "*" marks an equal contribution):

1. Gottschlich, A.*, **Thomas, M.***, Grünmeier, R.*, ... Marr, C. and Kobold, S., 2023. Single-cell transcriptomic atlas-guided development of CAR-T cells for the treatment of acute myeloid leukemia. *Nature Biotechnology, pp.1-15. Shared first authorship.*

2. **Thomas, M.***, Brabenec, R.*, ... Kobold, S. and Marr, C., 2023. Single-cell gene expression screening for efficacy and safety prediction of CAR T cell therapy targets. *In preparation. Shared first authorship.*

3. **Thomas, M.***, de la Porte, A.*, Schröder, J.*, ... Drukker, M., Schröter, C. and Marr, C., 2023. Single-cell multiomic comparison of differentiating pluripotent stem cells from three mammalian species. *In preparation. Shared first authorship.*

My individual contributions to the first publication are as follows:

Carsten Marr, Sebastian Kobold and I had the idea to use single cell gene expression data to identify novel targets for CAR T cell therapy. Together with Carsten Marr and Sebastian Kobold, I developed the outline of the computational approach. I designed and wrote the entire code and created the complete computational framework to develop an approach to identify targets for CAR T cell therapy based on gene expression data. I applied the approach to AML data and identified two previously unrecognized targets for CAR T cell therapy in AML. I performed all subsequent computational analyses of these targets in tumor and healthy tissues, including assessing and benchmarking gene expression profiles of these targets against reference targets across malignant and healthy tissues. Together with Adrian Gottschlich, I coordinated the project,

interpreted results, created the figures and wrote the manuscript in collaboration with Adrian Gottschlich, Carsten Marr and Sebastian Kobold.

My individual contributions to the second publication are as follows:

Carsten Marr, Sebastian Kobold and I had the idea to examine single cell gene expression profiles of CAR targets derived from clinical trials to identify potential differences that could account for the observed toxicities in clinical settings. Together with Carsten Marr and Sebastian Kobold, I developed the outline of the project and the computational framework. Together with Ruben Brabenec, I screened clinical trials and scientific literature to obtain approved and investigational CAR targets and processed large-scale single cell gene expression data from malignant and healthy tissues. Lisa Gregor was responsible for screening clinical trials to obtain patient outcome data. I performed all subsequent computational analyses of CAR targets across malignant and healthy cell populations, including analyzing CAR target expression between malignant and healthy cell populations and between individual patients. I interpreted differences between CAR target expression profiles on malignant and healthy cell populations in conjunction with clinical outcome data obtained from patients who underwent CAR T cell therapy. I identified novel CAR targets for the treatment of follicular lymphoma, multiple myeloma, and acute lymphoblastic leukemia using the previously explained computational screening approach based on single cell gene expression data. I coordinated the project, interpreted results, created the figures and wrote the manuscript in collaboration with Carsten Marr and Sebastian Kobold.

My individual contributions to the third publication are as follows:

Carsten Marr, Micha Drukker and Christian Schröter had the initial idea to profile differentiating PSCs from multiple species to investigate species-specific developmental timescales. Together, Carsten Marr, Micha Drukker, Christian Schröter, Alexandra de la Porte, Julia Schröder and I refined the scope of this project. Together with Carsten Marr, I developed the outline of the computational approach. I designed and wrote the entire code and created the complete computational framework to analyze timelapsed multiomic data from PSCs from three species during neural progenitor cell differentiation. I performed all subsequent computational analyses of differentiating PSCs from three species and characterized species-specific developmental timescales on gene expression, cell cycle distribution, chromatin accessibility and transcription factor activity levels. I coordinated the project, interpreted results, created the figures and wrote the manuscript in collaboration with Carsten Marr, Micha Drukker, Christian Schröter, Alexandra de la Porte and Julia Schröder.

Furthermore, my doctoral research contributed to the following publications, which are not included in this thesis:

4. Caulier, B, Joaquina, S., ... **Thomas, M.**, ... Wälchli, S., 2023. CD37 is a safe Chimeric Antigen Receptor target to treat acute myeloid leukaemia. *In preparation.*.

5. Gottschlich, A.* ,Grünmeier, R.*, ... **Thomas, M.**, ... Kobold, S., 2023. Multimodal dissection of single-cell landscapes enables the development of dual-targeting chimeric-antigen receptor T cells for the treatment of Hodgkin's lymphoma. *In preparation.*.

6. Benmebarek, R.*, Märkl, F.*, ... **Thomas, M.**, ... Kobold, S., 2023. Bispecific antibodies redirect synthetic agonistic receptor modified T cells against melanoma. *Journal for ImmunoTherapy of Cancer*, 11(5), pp.1-14.

7. Tritschler, S., **Thomas, M.**, ... Lickert, H. and Theis, F.J., 2022. A transcriptional cross species map of pancreatic islet cells. *Molecular Metabolism*, 66, pp.1-18.

8. Lesch, S., Blumenberg, V., ... **Thomas, M.**, ... Kobold, S., 2021. T cells armed with C-X-C chemokine receptor type 6 enhance adoptive cell therapy for pancreatic tumours. *Nature Biomedical Engineering*, 5(11), pp.1246–1260.

9. Cadilha, B.L.*, Benmebarek, M.R.*, ... **Thomas, M.**, ... Kobold, S., 2021. Combined tumor-directed recruitment and protection from immune suppression enable CAR T cell efficacy in solid tumors. *Science Advances*, 7(24), pp.1-12.

2 Computational methods for single cell sequencing data

The advent of single cell sequencing technologies has revolutionized the study of cellular heterogeneity and gene regulation by enabling high resolution profiling of individual cells. However, this high throughput data presents computational challenges due to its inherent dimensionality and sparsity, which increases as the scale and complexity of single cell sequencing experiments continue to grow [128]. To keep pace with ever-developing technologies and corresponding biological applications, computational models must adapt and evolve. Because the computational analysis workflow heavily depends on the experimental setup and applied sequencing protocols, there is no gold standard for processing and analyzing single cell data. With the increasing interest in single cell sequencing technologies (Figure 2.1a), researchers have dedicated significant efforts to inferring biological knowledge from single cell omics data, leading to the development of over 1400 published bioinformatic tools for single cell RNA sequencing (scRNA Seq) alone [129] (Figure 2.1b-c).

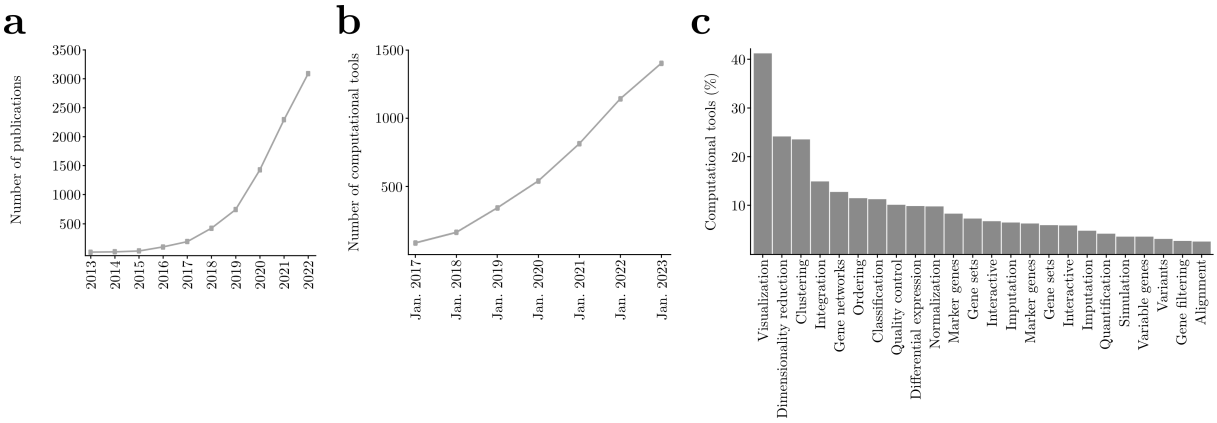

Figure 2.1: Growing interest in single cell RNA sequencing. a) Increasing number of single cell RNA sequencing (scRNA Seq) publications. Data extracted from https://pubmed.ncbi.nlm.nih.gov/ as of 04-05-2023, using the search term 'single cell RNA sequencing' in title, abstract, or text words. **b-c)** The number and functions of computational tools developed for the analysis of scRNA Seq. Data extracted from https://www.scrna-tools.org/ as of 04-05-2023.

This chapter of the thesis reviews computational approaches to extract biological insights from the immense amount of data produced by scRNA Seq and single cell assay for transposase-accessible chromatin sequencing (scATAC) Seq experiments. I will review a comprehensive workflow that begins with a count matrix for scRNA Seq or DNA fragments for scATAC Seq and elaborate on each step of data preprocessing and analysis that I employed throughout this thesis in subsequent sections.

2.1 Single cell gene expression

The analysis of scRNA Seq data typically commences with the generation of a count matrix, which represents the number of transcripts detected from each gene for each individual cell. The count matrix is then subject to various preprocessing and analysis steps, which will be the main focus of this section of this thesis.

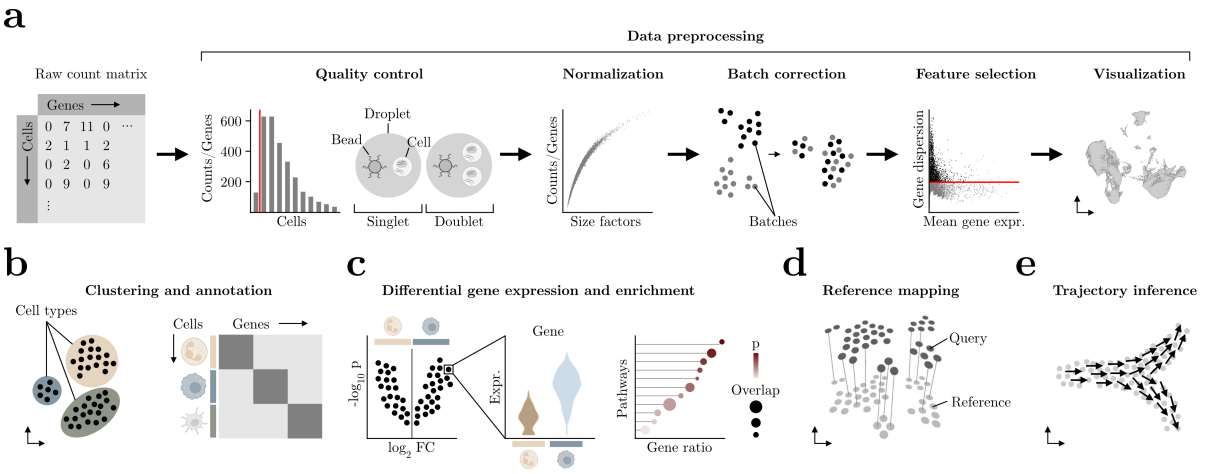

Figure 2.2: Workflow for single cell RNA sequencing data preprocessing and analysis. **a)** Starting with a count matrix, data preprocessing steps are critical for analyzing single cell RNA sequencing (scRNA Seq) data, as they help to correct for technical variability and improve the data quality prior to downstream analysis. Data preprocessing includes quality control, normalization, batch correction, feature selection and visualization in a low dimensional embedding. **b)** Clustering is used to group cells based on similarities in gene expression profiles, allowing identification of distinct cell populations. **c)** Differential gene expression analysis can be used to identify genes that are significantly enriched between different cell populations. Subsequent gene enrichment analysis on differentially expressed genes identifies pathways and processes that are enriched in specific cell types. **d)** Leveraging existing annotated reference datasets can improve the accuracy of scRNA Seq cell annotations and facilitate comparative analyses across datasets, tissues and conditions. **e)** Trajectory inference methods use gene expression profiles to reconstruct developmental paths and identify lineage relationships, providing insights into cell fate decisions in development, disease, and regeneration. Adapted from Heumos et al. [130] using elements from BioRender.com (2023).

Ready-made analysis pipelines, such as *Cellranger* from 10X Genomics, provide a convenient way to obtain count matrices from raw scRNA Seq data. The *Cellranger* software automatically processes and aligns raw single cell data from 10X Genomics protocols and assigns each Unique Molecular Identifier (UMI) read to a specific cell, a crucial process called demultiplexing. However, in some cases, cells from multiple samples, for example from different donors or species are pooled together in one single cell experiment to reduce costs or limit technical batch effects. Additional demultiplexing methods, such as *souporcell* [131] or *demuxlet* [132] can be utilized to confidently assign each of these cells to their respective sample of origin. These methods typically cluster

cells by their genotype by calling and counting single nucleotide polymorphisms - variants in DNA sequence that differentiate people and species - resulting in a count matrix where cells are unambiguously assigned to their sample of origin.

Figure 2.2 displays the basic preprocessing and selected subsequent analysis steps that I used for this thesis, starting with a count matrix. Below, the individual steps of data preprocessing and analysis along with computational methods employed for this thesis are discussed in more detail. For the analysis of scRNA Seq, several ready-made analysis platforms provide easy access to a combination of various tools and pipelines, from data preprocessing to high level downstream analysis. The most widely used platforms are *Seurat* [133], written in R, and *Scanpy* [134], part of the Python-based *scverse* ecosystem [135], which I predominantly used for scRNA Seq analysis during this thesis.

2.1.1 Quality control

In scRNA Seq data preprocessing, the initial step often involves identifying and removing low quality cells. What qualifies a cell of being of poor quality is usually a matter of definition and depends on various factors such as the experimental design, sequencing depth, and comparison to other sequenced cells. Cells are typically inspected based on their total distributions of UMIs and genes [136]. Cells with a comparatively low total amount of captured UMIs (counts) or genes are removed, as these may indicate failed reverse transcription or amplification, or empty oil droplets with background mRNA. Upper and lower thresholds for filtering cells according to captured UMI counts are defined as C_{min} and C_{max}. Upper and lower thresholds for filtering cells according to captured genes are defined as G_{min} and G_{max}.

Cells that are metabolically active tend to have a higher proportion of mitochondrial gene-derived mRNA molecules. However, cells with an unusually high proportion of mitochondria-derived genes, in conjunction with other indicators of low quality, may suggest a stressed or dying cell with broken membranes, and are therefore typically excluded from further analysis [137, 138]. Upper thresholds for filtering cells according to their maximum fraction of mitochondria-derived genes are defined as MT_{max}.

It is important to note that there is no universally accepted standard for these quality control metrics, and the specific thresholds are often determined by manual inspection of relevant plots. Optimal thresholds are determined by various experimental factors, such as the tissue source, dissociation protocol, and data quality, and should be set permissively to avoid inadvertently filtering out viable cell populations [128, 138].

In droplet-based single cell protocols, two or multiple cells can be captured in a single oil droplet, leading to the generation of so-called doublets or multiplets that are sequenced together, masked

as one single cell. These doublets or multiplets need to be accurately identified and removed to prevent confounding downstream analysis and biological interpretation [139]. The likelihood of doublet formation in droplet-based single cell experiments is related to cell concentration and the probability of capturing multiple cells within a single droplet [132, 140].

As such, the probability of doublet formation increases when higher cell concentrations are used to profile a larger number of single cells. While some quality control metrics can help to identify potential doublets in scRNA Seq data, utilizing count and gene metrics alone is often inadequate for precise doublet identification [141]. To address this issue, sophisticated doublet identification tools, such as *scrublet* [142] or *DoubletFinder* [143], simulate artificial doublets from the given dataset and use a nearest neighbor classifier to distinguish doublets from single cells, thus estimating a doublet probability for each cell. Upper thresholds for filtering cells according to this estimated doublet probability are defined as D_{max}.

After quality control and doublet removal, uninformative genes of low overall expression across cells are typically filtered out to reduce noise and improve the statistical power of downstream analysis [138]. Genes are generally filtered out if they fail to meet the threshold specified by the user, which is typically either a minimum number of cells expressing the gene or a minimum count number the gene was captured across cells. Notably, empty droplets that do not contain a single cell will still contain free-floating, ambient mRNA from other dissociating cells prior to library preparation, which can confound downstream analysis. To address this issue, several computational methods have been developed, such as *EmptyDrops* [144] or *SoupX* [145], which distinguish between empty droplets and cells, thus estimating and removing cell-free ambient mRNA contamination, which has been shown to effectively reduce technical noise and improve the quality of downstream analysis [130].

2.1.2 Normalization

ScRNA Seq data can show considerable variation in the amount of RNA per cell, both within a cell type and between different cell types. This is due to technical factors resulting from the different steps of the sequencing protocol, such as mRNA capture or reverse transcription during sequencing, as well as biological factors, such as cell size [138, 146]. Therefore, normalization is a crucial step in the preprocessing of scRNA Seq data which aims to account for this technical read variation between cells to enable meaningful results for downstream analysis.

The choice of normalization approach depends heavily on the gene coverage of the sequencing protocol used. For instance, full-length sequencing protocols, such as Smart-Seq2 [147] sequence the entire mRNA transcript and therefore need to account for gene length, as longer genes generate more reads. In contrast, droplet-based methods, such as the one used in this thesis from 10X Genomics, only sequence the 3' or 5' end of the transcript, and thus do not require

normalization to account for differences in gene length.

The most straightforward normalization approach is total count normalization, i.e. scaling the counts of each cell by a fixed factor, thereby accounting for differences in library size effects between cells. Total count normalization approaches like counts per million (CPM) normalization assumes that all cells in the dataset initially contained an equal number of mRNA molecules and count depth differences arise only due to sampling [138]. CPM divides the counts of each cell by the total read count and multiplies the result by a scaling factor (one million in the case of CPM) to ensure that the sum of all counts is constant across all cells. While CPM can be effective, it does not account for other sources of technical or biological variation, such as cell size and differences in overall gene expression between cell types and may not be suitable for datasets with significant technical or biological variability [138].

To address these limitations, the single cell specific *scran* algorithm [148] pools cells with similar total per-cell read count, thus accounting for the inherent sparsity in scRNA Seq data. These summed expression values are then deconvolved to estimate size factors based on a linear regression over genes for subsequent normalization for each cell, thus heavily reducing biases from differentially expressed genes, an advantage over classical total count normalization approaches [149].

2.1.3 Batch correction and data integration

In single cell sequencing experiments, both technical and biological variations contribute to the total variability of the data. Batch effects arise from various sources, such as differences in experimental sample handling, chosen sequencing platform, or sample preparation time. Depending on the experimental design, technical and biological variations can either be distinct or confounded [150]. For instance, in a scenario where two samples, one from a healthy individual and another from a diseased patient, are sequenced in separate runs, technical variation is confounded with biological variation. These effects can introduce unwanted variation in the data, which is unrelated to the biological variability of interest. This variation can be mistaken for true biological signal and must be accounted for to avoid spurious downstream correlations [151].

While batch correction methods aim to remove technical batch effects between samples or cells in the data, I refer to the task of integrating data from multiple experiments as data integration. Both batch corrections and data integration methods aim to remove unwanted sources of variation in the data while preserving true biological variation, but data integration methods usually deal with challenges involving compositional differences between datasets [138, 152].

As the number and scale of single cell datasets continues to increase, it becomes crucial to address batch effects and integrate data from multiple sources. These issues are particularly relevant for large-scale efforts such as the Human Cell Atlas, which aims to integrate data from

diverse tissues, platforms, and individuals [104, 151]. Consequently, numerous methods have been developed and benchmarked against one another to address these challenges [151, 153]. I refer to Lücken et al. [151] for an overview and benchmark of most commonly used methods for batch correction and data integration and only highlight key methods that were used in this thesis.

ComBat [154], originally borrowed from bulk microarray data, is a parametric empirical Bayes method that models batch effects as random effects and uses a linear regression framework to adjust for batch effects by adjusting the mean and variance between the specified batches. *ComBat* assumes that all differences between batches can be corrected by this, relatively naive, linear scaling approach. The increasing complexity of single cell datasets, however, often leads to variations in sample size, sampling time, conditions, and sequencing platforms, demanding computational methods that address these complex, nonlinear batch effects [151].

One of these data integration methods is *Harmony* [155], which first embeds cells in a low dimensional PCA space and iteratively removes batches by clustering similar cells from differently specified batches or datasets. At each iteration, cluster specific centroids from each batch are used to compute and apply linear correction factors for each cell. *Batch balanced k nearest neighbors (BBKNN)* [156], another graph-based approach, uses k-nearest neighbor (k-NN) graphs to align cells from different batches. *BBKNN* first constructs k-NN graphs for each batch independently and subsequently integrates them into a joint graph, thereby applying a correction factor to the weights of edges between cells from different batches to balance the contribution of each batch. Additionally, *single cell Variational Inference (scVI)* [157] is a deep generative model that utilizes a conditional variational autoencoder (VAE) to condition the dimensionality reduction process on a certain batch covariate, so that the covariate does not affect the resulting low dimensional representation. This method provides denoised counts in gene expression feature space and a low dimensional embedding that can be used for downstream tasks like clustering and visualization.

It is worth mentioning that there is no gold standard algorithm for batch correction or data integration, as each algorithm usually involves a trade-off between batch effect removal and preservation of biological variability [138]. Therefore, the selection of a suitable approach should be based upon comparison of the results obtained from multiple different batch correction or data integration algorithms.

2.1.4 Feature selection and visualization in a low dimensional embedding

In scRNA Seq experiments, due to limited amounts of starting material and sequencing depth, a significant number of expressed genes remains undetected in certain cells [128]. Consequently, a large proportion of entries in the resulting gene expression matrix contain zeros. These zero counts may arise from either true low or zero expression levels in cells or from technical artifacts, resulting in significant heterogeneity that poses a challenge in accurately interpreting the data [138, 158].

As every gene represents a dimension in scRNA Seq, even after filtering out these zero-count genes during quality control, the feature space for a scRNA Seq dataset can easily exceed 15,000 genes. This large amount of captured genes make scRNA seq data highly dimensional and noisy.

Therefore, it is crucial to select informative features that describe the biological variability of interest while filtering out noise, technical artifacts, and low quality genes to reduce data dimensionality for subsequent visualization and downstream analyses. Highly variable genes (HVGs) [159] can be identified based on normalized dispersion [136], i.e. by ordering genes along their mean expression in several bins and selecting them according to their highest variance-to-mean ratio. However, most feature selection methods typically consider the overall variability across the entire dataset, thereby masking genes that are differentially expressed in rare cell populations, as these genes only contribute minimally to the total variability [128].

For effective data visualization, it is crucial to capture and describe the underlying data structure as preserving as possible in two dimensions. Principal component analysis (PCA), a linear transformation method, is one of the most commonly used dimensionality reduction methods. PCA identifies the major sources of variability in the data and projects the data onto principal components that capture the largest proportion of the variance. The number of components retained for later analysis will depend on the complexity of the dataset [160]. Although PCA can effectively reduce the dimensionality of the data, it assumes that the variance is normally distributed, which may not be the case with nonlinear single cell data.

Uniform Manifold Approximation and Projection (UMAP) [161] is a popular, nonlinear dimensionality reduction and visualization method that has been widely used in the analysis of scRNA Seq data. By constructing a cell-to-cell nearest-neighbor network, the algorithm approximates the topology of the data, and subsequently estimates a low dimensional embedding that maximally preserves the underlying structure [128]. One of the advantages of UMAP over other dimensionality reduction methods is its ability to preserve the global structure of the data while still retaining the local relationships. However, it is important to note that UMAP visualizations depend on the choice of parameters and the initial conditions. This may lead to overinterpretation of relationships and heterogeneity in the data when relying on UMAP visualization alone [162].

2.1.5 Clustering

Clustering is a fundamental task in the analysis of scRNA Seq data. It groups cells with similar gene expression profiles into discrete clusters or subpopulations. This approach is commonly used to identify cell types and states, investigate data heterogeneity and dynamics and explore relationships between cells. Graph-based clustering methods, such as Louvain [163] and Leiden [164] clustering, rely on constructing a graph or network between cells, where cells are represented as nodes and edges are weighted according to the transcriptomic similarity of the connected

cells. These clustering algorithms then partition the graph into densely connected subgraphs and identify densely connected communities within the graph.

Louvain and Leiden clustering have been widely used in the analysis of scRNA Seq data and have been shown to outperform other clustering methods in terms of accuracy and speed [164–166]. Nonetheless, both algorithms have their own set of parameters that can significantly affect downstream results and interpretations, and their arbitrary nature makes reproducing prior clustering particularly challenging [167, 168]. For instance, the *resolution* parameter for Louvain and Leiden clustering determines the cluster size, and there are no definitive guidelines for identifying an optimal value. Thus, users must make informed decisions based on the particular dataset being analyzed.

2.1.6 Differential gene expression and enrichment analysis

Differential expression analysis is a statistical approach that enables the identification of genes that are highly enriched across groups, samples, or conditions in the data. This analysis step is commonly used in scRNA Seq analysis to annotate and characterize cell clusters and uncover molecular mechanisms or pathways driving differences between cells. Several differential expression analysis methods have been applied to or developed for scRNA Seq data, and their respective strengths and limitations have been thoroughly investigated [169, 170]. Hence, this section solely highlights key approaches that have been used for the analysis in this thesis.

Single cell methods that compare individual cells, such as the widely used Student's t-test compare the mean expression of a gene in two different groups of cells, assuming that the data is normally distributed and that the variances are equal in the groups. While this test provides a fast and simple approach to compute the top differentially expressed genes and usually picks up strong differences between groups [171], scRNA Seq data is often highly variable, which can violate the assumption of normal distributions. The Wilcoxon Rank Sum Test is a non-parametric test that does not make any assumptions about the distribution of the data. Instead, it ranks the expression values and compares the ranks between the two groups of cells, making it more robust to the variability of scRNA Seq data. Pseudobulk methods, such as *limma* [172] or *edgeR* [173] aggregate cells within a biological replicate to so-called pseudobulks before applying a statistical test. These methods have been shown to outperform single cell approaches, accounting for technical variation between replicates when comparing multiple groups of cells, thereby leading to less false positive discoveries [170, 171].

While differential expression analysis is a powerful tool for identifying genes that are differentially expressed between different conditions, the large number of results generated by this analysis can often make it difficult to extract meaningful biological insights. One approach to addressing this challenge is gene enrichment analysis, which can help identify biological pathways, molecular

functions, and cellular processes that are enriched in a set of differentially expressed genes. Tools like *gprofiler* [174] utilize a hypergeometric test to compare a specified number of differentially expressed genes to gene collections from various databases, such as Gene Ontology (GO) [175] or databases about biological pathways, diseases, and drugs such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [176]. These tools calculate the significance and overlap between the specific gene list and the list from the database, providing insights into the potential biological relevance of the differentially expressed genes.

2.1.7 Cell cycle inference

The cell cycle, a coordinated series of events that all cells undergo as they grow and divide, is an important consideration in the analysis of scRNA Seq data. Knowledge of the cell cycle phase is biologically informative at the compositional and gene level, as the cell cycle is known to affect gene expression [177].

Marker-based cell cycle inference relies on the expression of a small set of marker genes that are known to be associated with distinct cell cycle stages [177, 178]. Specifically, a cell cycle score is calculated for each cell by subtracting the average expression of a set of cell cycle genes from the average expression of a randomly sampled background set with expression values within the same range [134]. Once cells have been assigned to a certain cell cycle stage, the effect and variation caused by differences in the cell cycle between single cells can be regressed out using linear models. It should be noted, however, that regressing out cell cycle-caused differences can lead to overcorrection in certain circumstances and should be performed and evaluated with care [138].

2.1.8 Reference mapping and label transfer

Leveraging existing annotated reference datasets to annotate new scRNA Seq data can ease downstream analysis and biological interpretation despite issues such as technical variability and biological noise and can be particularly useful for ongoing initiatives such as the Human Cell Atlas [179]. However, the quality of the transferred annotations depends on the quality of the reference data, the model, and their compatibility with the dataset [130].

A simple, yet effective approach to integrating embeddings and annotations of queried data with a provided reference dataset is through Scanpy's *ingest* approach. This method projects the query data onto a PCA or UMAP space that has been fitted on the reference data, and, using a k-NN classifier, maps and projects the cell labels onto the embedding. It is noteworthy that Scanpy's *ingest* does not learn a joint representation of query and reference data but solely projects the query data onto the reference embedding.

Approaches that do learn a joint representation include *scANVI* [180], an extension to the previ-

ously discussed *scVI* [157] model. This semi-supervised variational autoencoder model initially learns a low dimensional representation of the query and the reference data and subsequently trains a classifier to learn cell type knowledge on the labeled reference data. The classifier is then used to predict and infer labels for cells in the unlabeled query data.

2.1.9 Trajectory inference and pseudotime analysis

Developmental processes, such as stem cell differentiation, occur gradually and involve a progressive shift in gene expression. The study of these differentiation processes on a single cell level has led to the revival of Waddington's landscape [181], a concept which envisions cells during developmental processes as balls rolling down a hill with numerous hills and valleys that represent the complex network underlying cellular fate decisions. Ideally, one would be able to repeatedly profile a single cell over time to investigate the underlying molecular characteristics. However, due to the destructive nature of single cell assays, each cell can only be measured once, leading to static snapshots of cellular states within such a landscape [182].

However, as biological events unfold unsynchronized across multiple cells, inferring a gradual order of gene expression profiles through trajectory inference methods is possible when sampling enough single cells along a developmental process [183]. This allows us to construct a pseudotime, which represents the order of transcriptomic similarity and typically, to some extent, corresponds to the cellular intrinsic temporal dynamics of differentiation. There are several computational approaches for trajectory inference and subsequent investigation of gradual changes in gene expression patterns along pseudotime [184]. However, these methods typically require prior biological knowledge to determine the direction of the trajectory and fail to assign directions to the recovered trajectories [183, 185].

Circumventing these issues, RNA velocity [186] enables an additional layer of information by incorporating the dynamics of mRNA transcription and degradation. Newly transcribed immature mRNA is still unspliced and, compared to mature and spliced mRNA, contains intronic sequences that are detectable in scRNA Seq experiments [186]. Assuming a simple model for each gene that relates the abundance of immature and mature mRNA, the change in mRNA abundance, known as RNA velocity, can be inferred [186]. Higher abundance of unspliced mRNA for a particular gene indicates upregulation in a cell and vice versa. Combining these velocities across genes can then be used to estimate the future state of an individual cell.

The *scVelo* [182] package estimates RNA velocity through a stochastic or dynamical model, allowing RNA velocity to be adapted to various specifications and datasets. Computational tools such as *cellrank* [187] build on top of *scVelo* by combining trajectory inference with directional information obtained from RNA velocity. *Cellrank* detects initial, intermediate and terminal populations and predicts cell fate probabilities towards any of these terminal states. These fate

probabilities then allow uncovering gene lineage drivers and visualizing gene expression trends along each individual lineage.

Waddington Optimal Transport (WOT) [188] builds on the optimal transport theory and can be used to compute distances between cells based on their gene expression profiles and infer developmental trajectories by minimizing the transportation cost between cell states for time-lapsed scRNA Seq data. To incorporate time information, *WOT* considers a sequence of probability distributions that correspond to the gene expression profiles of cells at different time points. By computing a time-dependent transportation plan between these distributions, *WOT* can describe the transition probabilities of cells from one time point to another. *WOT* has been implemented in *cellrank* as a module for the inference of developmental trajectories, computation of fate probabilities and identification of gene lineage drivers in time-lapsed scRNA Seq data.

2.2 Single cell chromatin accessibility

Regulatory mechanisms, such as epigenetics and chromatin accessibility, play a crucial role in controlling gene expression [189, 190]. To gain a deeper understanding of chromatin state dynamics at the single cell level, scATAC Seq can be used to quantify genome-wide chromatin accessibility in individual cells. Although scRNA Seq and scATAC Seq share similarities in the basic outline of their data preprocessing workflow, there are fundamental differences in individual steps [130, 191–193]. Unlike scRNA Seq, which captures mRNA from a fixed set of genes, scATAC Seq samples fragments from the entire accessible genome and thereby lacks a standardized feature set, leading to data that is more sparse, noisier, and usually bigger compared to scRNA Seq [130].

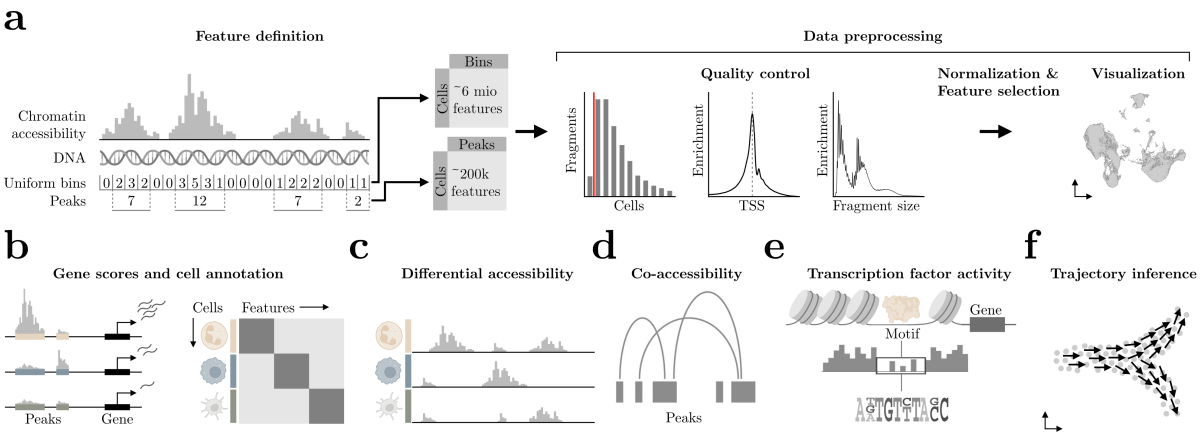

Figure 2.3: Workflow for single cell ATAC sequencing data preprocessing and analysis. **a)** Starting with a fragment file, a feature set first has to be defined which serves as the basis for subsequent data preprocessing steps. Data preprocessing includes quality control, normalization, feature selection and visualization in a low dimensional embedding. **b)** Gene scores can be estimated based on the local accessibility of the gene region, including the promoter and gene body, across all cells in the dataset. **c)** Differential accessibility analysis identifies genomic regions that show differential chromatin accessibility between different cell types or conditions. Pinpointing these regions allows identification of key regulatory elements, transcription factors, genes and pathways that drive cell type-specific functions and transitions. **d)** Co-accessibility methods identify correlations between accessible genomic regions in single cells, assuming that regions involved in similar biological processes have correlated accessibility patterns that may be associated with the expression of nearby genes. **e)** Transcription factor activity can be inferred based on motif enrichment, deviations of per-cell accessibility of a given motif or through footprinting analysis. **f)** Trajectory inference methods model the developmental trajectories of cells by ordering them along a pseudotime axis. This allows the identification of cell fate decisions and the reconstruction of differentiation pathways, as well as the detection of genes and regulatory elements that are driving cell fate transitions. Adapted from Heumos et al. [130] using elements from BioRender.com (2023).

Preprocessing and analyzing of scATAC Seq data begins with a fragment file that contains detailed information about all sequenced DNA fragments generated from Tn5 transposition events,

including their positions on the reference genome (see section 1.4.2 for a detailed explanation of the scATAC Seq workflow). This fragment file can be obtained using ready-made analysis pipelines such as *Cellranger ATAC* from 10X Genomics, which, similar to the *Cellranger* pipeline for scRNA Seq data, automatically processes, aligns and demultiplexes raw single cell data from 10X Genomics ATAC protocols.

Figure 2.3 displays typical preprocessing and selected subsequent analysis steps that I used for this thesis, starting with a fragment file. Below, the individual steps of data preprocessing and analysis along with computational methods employed for this thesis are discussed in more detail. As the rationale behind data preprocessing steps are similar to those of scRNA Seq, the focus will be on the specifics of scATAC Seq. Various analysis platforms exist for the analysis of scATAC Seq data, which combine different ready-to-use tools and pipelines for data preprocessing to high level downstream analysis. A widely used platform for the analysis of scATAC Seq data for this thesis is *ArchR* [194], a package written in R that offers an extensive suite of analysis tools for processing and analysis of scATAC data.

2.2.1 Peak calling or binning for generation of feature matrices

In scATAC Seq, a standardized feature set is not available as it is in scRNA Seq and first has to be identified, which is mostly done using either cell-by-peak or cell-by-bin matrices [130]. Peaks refer to variable regions of open chromatin with an enrichment of Tn5 transposition events over background noise. Bins, on the other hand, are uniformly sized windows across the genome that capture all Tn5 transposition events. Binning algorithms capture Tn5 transposition events in equal-sized genomic regions (bins).

While some analysis pipelines combine peak calling and bin calling algorithms, the choice of peaks or bins as features depends on the research question and the characteristics of the dataset. Peak matrices provide a more direct measure of the chromatin accessibility of each genomic element. Due to the binary nature of scATAC Seq data, where regions are either accessible or not accessible, peak calling is generally performed on groups or clusters of cells rather than on an individual cell basis [130, 194]. Hence, identifying peaks requires a sufficient number of cells and therefore may fail in rare cell types or in experiments with low amounts of captured cells [130, 195]. Bin matrices, on the other hand, can suffer from increased noise or bias when using too few or too large genomic bins. If the bin size is too small, resulting bin matrices can be too sparse and suffer from a high false positive rate. Conversely, if the bin size is too large, resulting bin matrices may not capture subtle differences in chromatin accessibility or cell-to-cell variability within a genomic region.

For this thesis, peaks were called from scATAC Seq data using *Macs2* [196]. Originally designed for ChIP Seq data, *Macs2* estimates the local background noise level using a sliding window

approach. It then models the read count distribution with a Poisson distribution using a dynamic thresholding method to determine the significance of the peaks. The dynamic thresholding method calculates a False Discovery Rate (FDR) to control for multiple testing, and adjusts the threshold to ensure that the FDR is below a user-defined value.

2.2.2 Quality control

To ensure downstream analysis contains biologically meaningful features and not noise, quality control is essential for identifying and removing low quality cells and features. Similar to scRNA Seq quality control, there are several criteria that can be used to filter cells and peaks based on their quality.

Low quality cells can arise from technical issues during the experiment, such as improper cell lysis or amplification. One metric used to evaluate the quality of scATAC Seq cells is the total number of fragments per cell, which is directly proportional to the amount of chromatin accessible to the transposase enzyme used in the scATAC Seq protocol. Cells with very few usable fragments will not provide enough data to make useful interpretations and should therefore be excluded [130]. Upper and lower thresholds for filtering cells according to captured fragments are defined as F_{min} and F_{max}. However, the total number of fragments alone is not sufficient to assess cell quality since this can vary widely depending on the type of cell, cell cycle stage, or the amount of input material.

Transcription start site (TSS) enrichment scores measure the enrichment of chromatin accessibility at gene promoters compared to random genomic regions. ATAC Seq data is universally enriched at gene TSS regions compared to other genomic regions, due to large protein complexes that bind to promoters [126, 130, 191, 194]. The TSS enrichment score represents the ratio between the peak of the enrichment at the TSS and flanking regions for each cell. Cells with low TSS enrichment scores are usually of poor quality, as the signal may be too weak to detect genuine chromatin accessibility [197]. Lower thresholds for filtering cells according to TSS enrichment scores are defined as TSS_{min}.

Nucleosomes, composed of DNA wrapped around histone proteins, restrict the accessibility of DNA and thus hinder transposase cleavage, leading to low accessibility of DNA that is tightly wrapped around a nucleosome (see section 1.4.2 for a detailed explanation of the scATAC Seq workflow). Because of the patterned way that DNA wraps around nucleosomes, there is a nucleosomal periodicity in the distribution of fragment sizes in scATAC Seq data as fragments must span 0, 1, 2, etc. nucleosomes [130, 191]. The nucleosome signal is therefore defined as the ratio of long fragments resulting from one or multiple histones bound between the Tn5 transposition sites and short nucleosome-free fragments. This signal consists of distinct peaks and valleys upon inspection of fragment size and is often used to assess the quality of scATAC Seq data [197].

Several other criteria are used to filter out poor quality peaks from scATAC Seq data. For example, peaks from non-standard chromosomes or chromosome scaffolds not found in the reference genome are typically excluded as they are likely to represent experimental artifacts or low quality data. Additionally, the ENCODE project has provided a list of certain genomic regions which are known to have low chromatin accessibility, termed blacklist regions [198]. Peaks identified in these blacklist regions are often associated with artifactual signals and should be removed from downstream analysis.

2.2.3 Normalization, feature selection and low dimensional visualization

Similarly to scRNA Seq data, subsequent preprocessing of scATAC data generally involves normalization, feature selection and visualization in a low dimensional embedding. However, the computational approaches used for each step in scATAC Seq data analysis differ due to the increased inherent sparsity of scATAC Seq data compared to scRNA Seq data. Since there are only two copies of DNA in each cell in a diploid organism, the maximum number of counts for a given base-pair position is two [130, 191, 193]. Thus, a zero in scATAC Seq could indicate that a certain genomic region is either inaccessible or that it was not sampled. This inherent sparsity causes high inter-cell similarity at all zero positions and therefore makes traditional approaches such as PCA unsuitable for analysis of scATAC Seq data [194].

To address this challenge, latent semantic indexing (LSI), a natural language processing method, originally designed to assess document similarity based on word counts, has been adapted for scATAC Seq data analysis [194, 199, 200]. LSI is a layered dimensionality reduction approach and combines term frequency-inverse document frequency (TF-IDF) normalization and singular value decomposition (SVD). In the case of scATAC Seq, each sample is treated as a document and each peak is treated as a term. First, the term frequency is calculated for each peak in each single cell, which reflects the number of times a particular peak is accessible. Due to the nature of scATAC Seq, a peak can be accessible 0,1, or 2 times. This term frequency is then depth normalized by a constant 10,000. Then, the values are normalized by the inverse sample frequency across all cells to identify the peaks that are more specific rather than commonly accessible. Finally, the resulting matrix is then log-transformed, generating a TF-IDF normalized matrix that reflects the importance of each peak to each sample. Next, SVD is performed to identify the most valuable information across samples and represent them in a lower-dimensional space. Finally, a more conventional dimensionality reduction technique, such as UMAP can be used to visualize the data.

2.2.4 Gene scoring and cluster annotation

Features in scATAC Seq data are relatively hard to interpret compared to a fixed set of genes in scRNA. When corresponding RNA data is available for the same cells, as is the case in this thesis, gene annotations from the RNA signal can be transferred and systematically refined in ATAC

resolution to further elucidate cellular identity, if necessary. In the absence of corresponding RNA data, gene expression for cell type-specific marker genes can be estimated from chromatin accessibility data alone to aid in reliable cell type annotation. This can be accomplished by calculating gene scores, which estimate the level of gene expression based on the local accessibility of the gene region, including the promoter and gene body, across all cells in the dataset. Adjusting for the distance to the gene and large differences in gene size, the resulting gene scores are useful for cell type identification based on marker gene expression and for performing differential gene expression analysis between groups of cells [194].

To further enhance the visualization and simplify the process of cell type annotation, simple models such as *MAGIC* [201] learn the manifold data and use the resulting graph to denoise gene activity scores.

2.2.5 Differential chromatin accessibility

Similarly to scRNA Seq, differential accessibility analysis employs statistical models to identify variations in accessible genomic regions across cells, conditions, or samples, enabling the discovery of underlying molecular mechanisms or pathways among cells. It is worth noting that while scRNA Seq data performs analyses at the gene level, scATAC Seq data may use peaks, genes based on gene scoring methods, or even transcription factor motifs (discussed in section 2.2.7).

For an unbiased identification of marker features in a cell population of interest, first, background cells are selected that closely resemble the cells of interest based on their quality control metrics, such as TSS enrichment score and normalized amount of fragments per cell [194]. The normalization of these values using quantile normalization ensures that the variance of each dimension is distributed uniformly across the same relative scale. Subsequently, the nearest neighbors with the most similar bias in terms of TSS enrichment and amount of fragments per cell are determined in this normalized multidimensional space using the Euclidean distance metric [194]. The generation of a bias-matched background cell group is crucial as it allows for a more robust determination of significance even in smaller cell groups. Enriched features are then identified by comparing cell groups using pairwise statistical tests such as the Wilcoxon Rank Sum Test and ranked based on their adjusted p-value and FDR.

2.2.6 Peak co-accessibility

Co-accessibility methods identify correlation between peaks across many single cells, based on the assumption that accessible genomic regions in the same cells are involved in similar biological processes and their accessibility patterns may be correlated with the expression of nearby genes [202]. The peak correlation-based approach implemented in *ArchR* measures the correlation between the binary accessibility profiles of all pairs of genomic regions, resulting in a matrix that provides co-accessibility scores for each pair of peaks. It is worth mentioning that

although these co-accessibility analyses are essential for understanding regulatory networks, cell type-specific peaks are often linked as they are typically accessible within a single cell type. This leads to a strong correlation between these peaks, but does not necessarily indicate a regulatory relationship between them [194].

2.2.7 Transcription factor activity

After having identified a set of peaks or features that are specific to a certain cell population or condition, it becomes possible to predict which transcription factors (TFs) may be mediating the binding events that create those accessible chromatin sites. As key lineage-defining TFs are often found in cell type specific accessible chromatin regions, the activity of transcription factors can shed further light into the regulatory mechanisms governing gene expression.

In scATAC Seq data, one of the most straightforward ways to identify potential TF activity is to search for enriched DNA binding motifs of TFs within a set of peaks or features. This approach assumes that the presence of a particular DNA motif within an open chromatin region indicates the binding of the corresponding TF. Databases like JASPAR [203] and Cis-BP [204] provide detailed information about binding motifs of TFs from multiple species, which can be used to identify enriched DNA motifs. In addition, the ENCODE consortium [205] has mapped TF binding sites across a wide array of cell types, creating a TF motif database that can help characterize unknown cell types, similar to the enrichment analyses described in section 2.1.6

However, these enrichment analyses are typically not performed on the single cell level and fail to take the insertion sequence bias of the transposase into account. To address these limitations, *chromVar* [206] calculates TF activity by measuring the deviation of per-cell accessibility of a given motif from the expected accessibility based on the average of all cells or samples. This approach calculates a deviation score for each motif for all cells, reflecting the estimated activity of each TF across individual cells.

Motif enrichment analysis has a notable limitation in that the mere presence of a DNA binding motif may not necessarily imply actual binding by the corresponding TF [207]. The DNA motif may instead be bound by a different TF, be part of a larger regulatory complex, or remain unbound altogether. Moreover, some TFs can bind to regions without any known motif, posing challenges in identifying their activity through motif enrichment analysis alone [206, 208].

To circumvent this limitation, TF footprinting can be used to identify TFs that are bound to a particular genomic region, based on the pattern of chromatin accessibility in that specific region. TF footprinting is based on the notion that the binding of a TF to the DNA protects the underlying DNA sequence from being cut by the transposase and sequenced. Consequently, genomic regions that are bound by TFs exhibit lower levels of chromatin accessibility compared to

flaking regions that lack TF binding. To this end, peak regions are initially scanned for any DNA sequence that matches a motif, and the genomic locations of the relevant motifs are obtained. To account for Tn5 insertion biases, *ArchR* generates a matrix of all possible hexamer position frequencies and k-mer frequencies at Tn5 insertion sites [194]. Since conducting TF footprinting at a single site level would typically require a substantial sequencing depth, transposase insertion sites for a specific motif are typically merged across cells of a specific group to establish an aggregate TF footprint and accurately predict TF binding events [194].

2.2.8 Trajectory inference and pseudotemporal ordering

Inferring a pseudotemporal ordering of single cells in a low dimensional space can be used to investigate gradual changes in chromatin accessibility and TF activity over time. This involves defining a trajectory backbone in *ArchR*, which provides a rough ordering of cell across timepoints or cell clusters for time-lapsed or snapshot data [194]. For time-lapsed differentiation data, the first timepoint was used as the backbone. This backbone enables the calculation of the distance for each cell from the first cluster to the mean coordinates of the next cluster, resulting in a pseudotime value for all cells along the previously defined trajectory. Aligning cells to the trajectory based on their Euclidean distance to the nearest point along the manifold facilitates the analysis of changes in chromatin accessibility peaks, gene scores, or TF deviations across the inferred pseudotime. TFs where gene expression is positively correlated to changes in the accessibility of their corresponding motif can then be identified by integrating of gene scores along with TF motif accessibility across pseudotime. This can potentially reveal drivers of differentiation or regulatory elements that are dynamic throughout the cellular trajectory [194].

3 Single cell transcriptomics for CAR target identification in AML

CAR T cell therapy has shown promise in treating B cell malignancies, but its effectiveness in other malignancies is hindered a crucial absence of safe targets (see section 1.1.2). Previous target identification strategies using transcriptomics typically rely on bulk data, which lack the granularity necessary for comprehensive analysis.

In this chapter, I first describe a new computational approach for de novo target identification for chimeric antigen receptor (CAR) T cell therapy using single cell gene expression data. Applying this approach to acute myeloid leukemia (AML), I identified two previously unrecognized targets, CSF1R and CD86. We subsequently demonstrated their efficacy in vitro and in vivo.

The goal of this project was to utilize single cell gene expression data to identify CAR targets with an optimal expression profile regarding their therapeutic efficacy and safety. To achieve this goal, I developed a stepwise computational framework specifically designed to identify potential CAR targets based on their single cell gene expression profiles. Applying my approach to AML, I leveraged gene expression data from over 500,000 single cells collected from 15 AML patients and 9 healthy tissue samples. Aided by this high resolution, single cell expression approach, I computationally identified colony-stimulating factor 1 receptor (CSF1R) and cluster of differentiation 86 (CD86) as promising targets for CAR T cell therapy in AML. Functional validation of these established CAR T cells shows robust in vitro and in vivo efficacy in cell line- and human-derived AML models with minimal off-target toxicity toward relevant healthy human tissues, providing a strong rationale for further clinical development.

This chapter is based on and partly identical to the following publication in Nature Biotechnology:

1. Gottschlich, A.*, **Thomas, M.***, Grünmeier, R.*, ... Marr, C. and Kobold, S., 2023. Single-cell transcriptomic atlas-guided development of CAR-T cells for the treatment of acute myeloid leukemia. *Nature Biotechnology, pp.1-15. Shared first authorship.*

Note that "*" marks an equal contribution. See section 1.5 of this thesis for a detailed summary of individual contributions.

3.1 A computational framework for CAR target identification

CAR T cells targeting B cell lineage antigens such as cluster of differentiation 19 (CD19) or B cell maturation antigen (BCMA) have demonstrated clinical efficacy in heavily pretreated individuals suffering from different B cell malignancies, such as B cell lymphoma, B cell acute lymphoblastic leukemia and multiple myeloma [37–39]. However, CAR T cells targeting non-B cell-associated epitopes have yet to show similar response rates [209] (see section 1.1.2). For instance, in myeloid malignancies, such as acute myeloid leukemia (AML), common target structures are often coexpressed on vital tissues, such as endothelial cells or hematopoietic stem and progenitor cells (HSPCs), increasing the risk for on-target off-tumor toxicity [54, 210]. Identifying safe target structures is thus pivotal to translate the vast potential of CAR T cell therapy to myeloid neoplasms and other malignancies.

Newly developed CAR T cells are often directed to target structures that have already been used for antibody therapy. By contrast, unbiased de novo target screenings for CAR T cell therapy have rarely been conducted [50]. In addition, until recently, off-tumor antigen projections could only leverage bulk sequencing data, missing detailed information about cell-type-specific target antigen expression patterns [50]. Conveniently, recent advancements in single cell technologies have provided extensive expression data, offering precise insights into healthy and malignant cells [211] (see section 2.1). This data, still a mostly untapped resource for therapeutic development, at least in the context of de novo antigen predictions and CAR T cell development, enables thorough on- and off-tumor antigen prediction [212], enhancing CAR T cell development with unprecedented resolution.

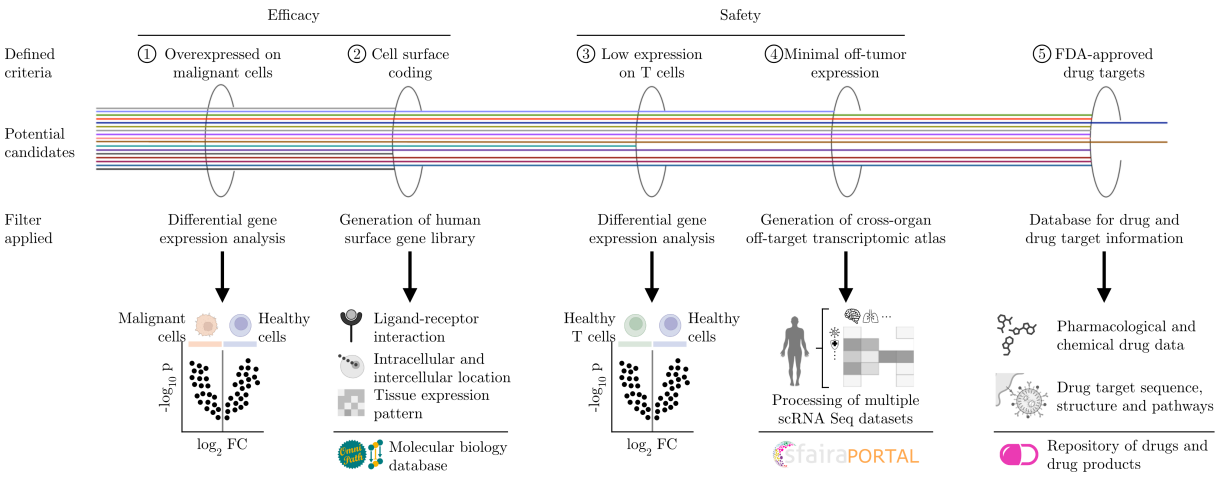

Figure 3.1: Computational scRNA Seq-based CAR target antigen identification approach. To ensure CAR efficacy, a suitable candidate is overexpressed in malignant cells and located on the cell surface. In terms of CAR safety, the candidate should not be expressed on T cells and show minimal expression across vital, healthy tissues. To further optimize the safety profile of CAR targets, I focused on targets with known pharmacological action of FDA-approved drugs.

I thus developed a computational approach, leveraging single cell RNA sequencing (scRNA Seq) data for CAR target identification (Figure 3.1).

To ensure CAR efficacy, a suitable candidate is (1) overexpressed in malignant cells and (2) located on the cell surface. Overexpressed targets were identified by performing differential expression analysis between malignant and healthy cell populations using a using a t-test with overestimated variance (see section 2.1 of this thesis for a detailed explanation of computational methods). To identify candidates accessible for CAR T cells on the target cell surface, I used OmniPath [213], a large-scale molecular database, to integrate data from ligand-receptor interactions, intracellular and intercellular location and tissue expression patterns [214–217] into a comprehensive human surface gene library. This newly generated library served as a resource for identifying candidates that are accessible for CAR T cells on the target cell surface (Figure 3.1).

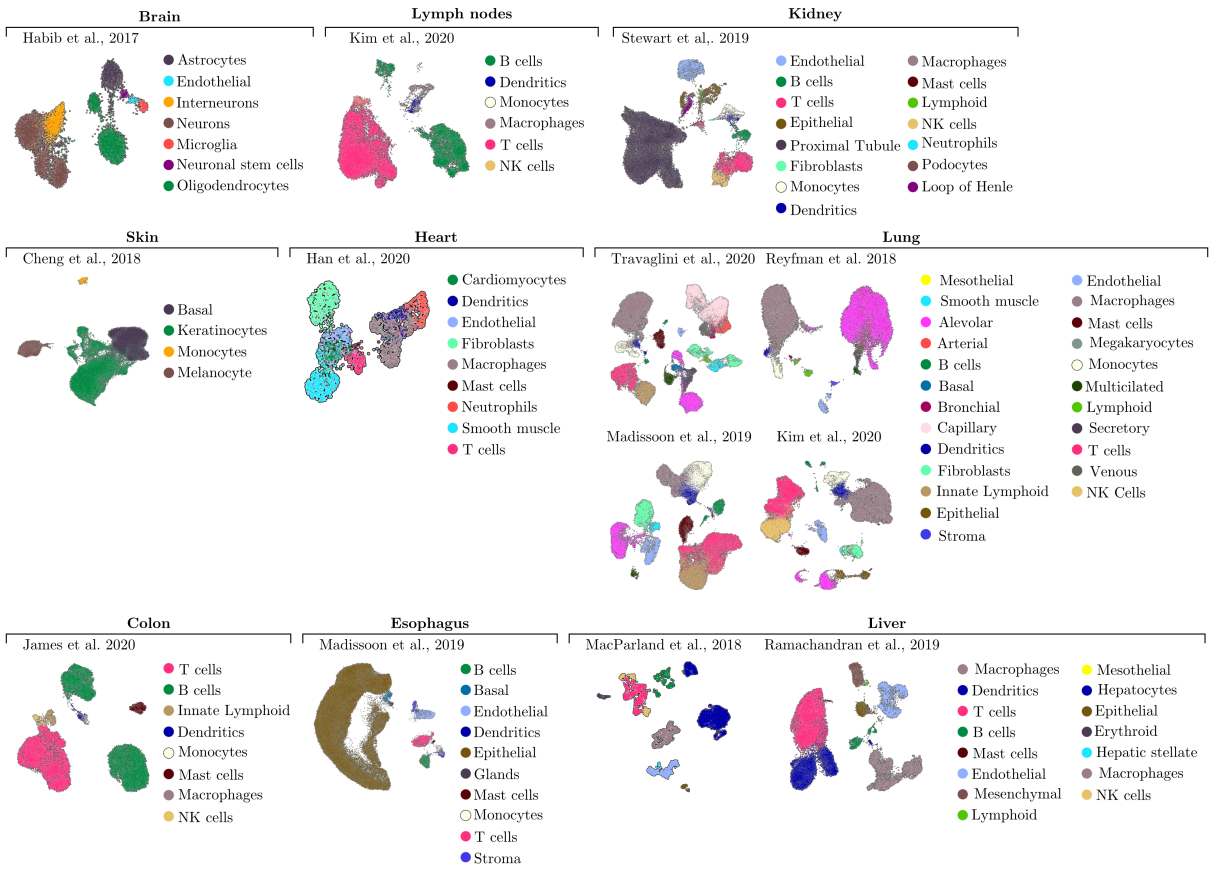

Figure 3.2: Generation of a cross-organ off-target transcriptomic atlas (COOTA). UMAP plots of 11 scRNA Seq datasets from various healthy tissues with colors highlighting clustering of respective cell types. Cell annotations were provided by the authors of the respective studies. Figure adapted from Gottschlich et al. [127].

In terms of CAR safety, the candidate should (3) not be expressed on T cells and (4) show minimal expression across vital, healthy tissues. I identified enriched marker genes in T cells by comparing the mean expression of healthy T cells to the mean expression remaining healthy

cell types using a t-test with overestimated variance (see section 2.1 of this thesis for a detailed explanation of computational methods). To minimize on-target off-tumor effects, I therefore processed 11 scRNA Seq datasets from nine healthy human tissues (brain, lung, lymph nodes, heart, skin, liver, kidney, colon and esophagus) [218–228] (Figure 3.1, Figure 3.2, Table 1). These datasets, consisting of over 500,000 single healthy cells in total, contributed to the comprehensive evaluation of target expression levels across various healthy tissues within the cross-organ off-target transcriptomic atlas (COOTA). A detailed summary of all scRNA Seq datasets used to construct COOTA is provided in Table 1. Cell annotation labels provided by the respective study authors were thoroughly inspected and, if necessary, relabeled to facilitate global comparisons of cell populations. I excluded targets highly expressed in vital non-immune cell lineages or on cell types of tissues in direct proximity to infused T cells (that is, endothelium, arteries, veins, bronchial vessels, capillary and smooth muscle cells). To mitigate batch-effect induced changes in gene expression across tissues in COOTA, I defined thresholds based on absolute percentage values rather than raw expression counts.

Table 1: Overview of COOTA scRNA Seq Datasets.

Organ	**Method**	**Sequenced cells**	**Cell types**	**Abbreviation**	**Publication**
Brain	DroNc Seq	13,067	7	Hb2017	Habib et al., 2017 [218]
Lymph nodes	10X	37,446	6	Km2020	Kim et al., 2020 [219]
Kidney	10X	40,268	15	Sd2019	Stewart et al., 2019 [220]
Skin	10X	68,036	4	Cg2018	Cheng et al., 2018 [226]
Heart	Mircowell Seq	2,753	9	Hn2020	Han et al., 2020 [228]
Lung	10X	60,633	24	Ti2020	Travaglini et al., 2020 [221]
		41,503		Rn2018	Reyfman et al., 2018 [223]
		56,304		Mn2019	Madissoon et al., 2019 [222]
		42,995		Km2020	Kim et al., 2020 [219]
Colon	10X	41,322	8	Js2020	James et al., 2020 [227]
Esophagus	10X	87,947	10	Mn2019	Madissoon et al., 2019 [222]
Liver	10X	6,271	15	Md2018	MacParland et al., 2018 [224]
		33,037		Ra2019	Ramachandran et al., 2019 [225]

To further optimize the safety profile of newly developed CAR T cells (5), I reasoned that, if targeted therapies for any of the so far identified candidates have already been approved by the Food and Drug Administration (FDA), the risk for unexpected, severe on-target off-tumor toxicities will be minimized. In addition, this could shorten time and decrease regulatory hurdles for translation of newly developed CAR T cells into clinical routines, as safety of target-directed therapies was previously demonstrated. Thus, I utilized a database that contains information on the interactions, pharmacology and chemical structures of all monitored FDA-approved drugs and drug targets [229] and defined druggable targets as targets with known pharmacological action of FDA-approved drugs.

3.2 Identification of CSF1R and CD86 as CAR targets in AML

AML is the most common acute leukemia in adults, and its molecular heterogeneity has complicated the successful development of new therapeutic agents [230]. Despite upfront curative intent in most individuals with combinatorial chemotherapy, disease relapse is frequent, occurring in over 50% of treated individuals [231]. After relapse, allogeneic hematopoietic stem cell transplantation (allo-HSCT) remains the only curative approach; but even then, long-term survival probabilities are below 20%. Therefore, innovative treatment options represent a high unmet medical need. Currently, CAR T cells targeting AML-associated target antigens CD33 and interleukin-3 receptor-α (IL3RA, CD123) are undergoing clinical investigation. Due to preclinical evidence of off-tumor toxicity toward HSPCs, most clinical trials are evaluating the potential of anti-CD123 or anti-CD33 CAR T cells as a bridge-to-transplant regimen before allo-HSCT. Early reports of these trials have shown only limited therapeutic efficacy [232–234]. Yet, more complete results of these clinical studies in AML are eagerly awaited. Meanwhile, other targets, such as CD70, C-type lectin-like molecule-1, FMS-like tyrosine kinase-3 (FLT3), CD44 variant 6 (CD44v6), sialic acid-binding Ig-like lectin-6 (Siglec-6) or CD117, have been tested in preclinical studies as alternative CAR targets [235–239]. However, clinical validation is pending, and expression profiles of most of the targets raise at least some uncertainties regarding their clinical safety and efficacy.

3.2.1 Analysis of AML and healthy cells for CAR target identification

To apply my computational approach for de novo CAR target identification to AML, I used publicly available scRNA Seq data from 15 individuals with AML [240]. Here, I excluded individual AML916, as it had a mixed AML phenotype expressing markers of stem cells, myeloid, T and B lineages. Barcodes were filtered for each sample for high quality cells based on the total distributions of unique molecular identifier counts and genes. I excluded cells with captured counts C_{max} >12,000, genes G_{max} >4,000, or fraction of mitochondria-derived genes MT_{max} >20% from further analysis (see section 2.1.1 for a definition and detailed explanation about applied filtering thresholds). I excluded barcodes that could not be confidently assigned to either healthy or tumor, as well as genes detected in less than 2% of cells from further analyses. I normalized unique molecular identifier counts of each cell using the *scran* algorithm [148]. I identified the top 4,000 highly variable genes based on normalized dispersion [136]. On these, I performed dimension reduction by computing 50 principal components. To account for technical batches, I used *Harmony* [155] to integrate data from the respective individuals. Next, I computed a neighborhood graph on all 50 harmony-adjusted principal components with 15 neighbors. For two-dimensional visualization, I embedding the neighborhood graph via uniform manifold approximation and projection (UMAP) [161] with an effective minimum distance between embedded points of 0.5 (see methods section 2.1 of this thesis for a detailed explanation of applied preprocessing steps).

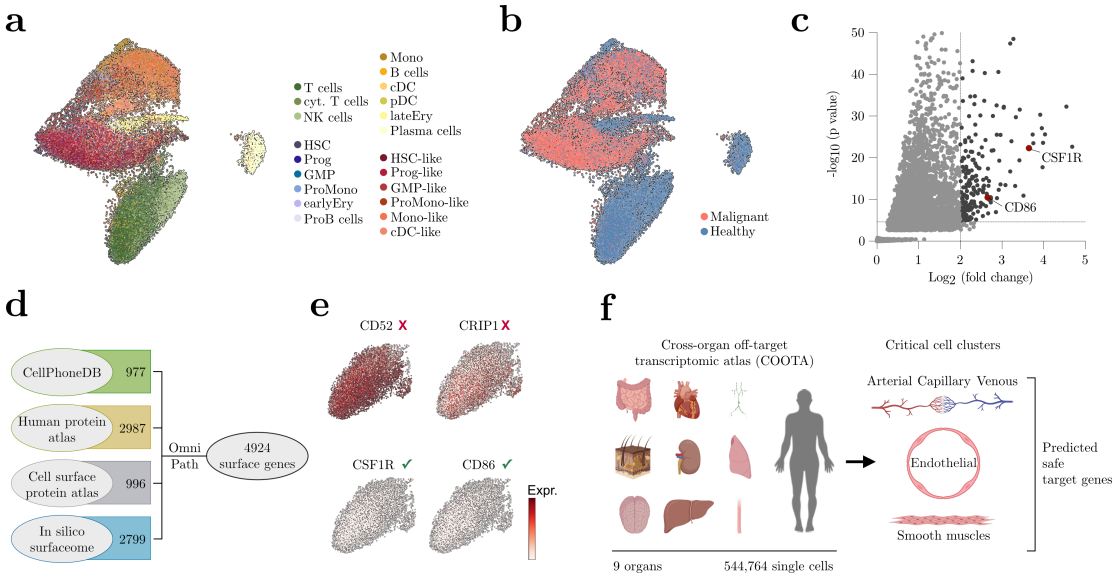

Figure 3.3: ScRNA Seq-based screening approach identifies CSF1R and CD86 as potential CAR targets in AML. a-b) UMAP showing 28,404 healthy and malignant cells from data of 15 previously published individuals with AML harboring 15 different mutations [240]. Normalized gene expression values were log transformed. Colors highlight the different cell types **(a)** and condition **(b)**. Cell annotations are provided; NK cells, natural killer cells; GMP, granulocyte–monocyte progenitors; ProMono, promonocytes; EarlyEry, early erythrocytes; ProB cells, pro-B cells; Mono, monocytes; cDC, conventional dendritic cells; pDC, plasmacytoid dendritic cells; LateEry, late erythrocytes. **c)** Volcano plot showing the remaining two target antigens with their respective FDR-adjusted log10 p value and log fold change values from differential expression analysis between malignant HSPC-like cells and healthy HSPCs using a t-test with overestimated variance. Dashed lines indicate applied thresholds at a log fold change >2 and p value <0.01. **d)** Summary of databases used to identify genes coding for proteins on the cell surface. **e)** Quantification of T cell expression of newly identified targets. Red crosses indicate targets with high expression on T cells, which were excluded from further analyses. Green check marks indicate no significant expression on T cells. **f)** Harmonization of 11 scRNA Seq datasets from nine healthy human tissues into a COOTA consisting of 544,764 cells. A detailed summary of all used datasets is provided in Table 1. Targets highly expressed in non-immune cell lineages or on cell types in direct proximity to infused T cells (critical cell clusters: arterial, capillary, venous, endothelial and smooth muscle cells) were excluded from further analysis. Figure adapted from Gottschlich et al. [127].

A total of 28,404 sequenced healthy and malignant bone marrow cells passed quality control (Figure 3.3a,b). Next, I sought to identify candidates with higher expression on malignant HSPC-like cells (herein termed hematopoietic stem cell (HSC)-like and progenitor (Prog)-like) than on healthy cells. I identified characteristic gene signatures of AML HSPCs by performing separate differential expression analysis of AML HSC-like and Prog-like cells against their healthy equivalents, respectively. I defined enriched marker genes as genes with log fold change >2 and FDR-adjusted p values ≤ 0.01. Differential gene expression analyses between malignant and healthy HSPCs revealed 96 genes that were strongly overexpressed in HSPC-like cells and were

used for further downstream analyses (Figure 3.3c).

To obtain genes encoding proteins on the cell surface, I used OmniPath [213], a large-scale molecular database, to access data from (1) the mass spectrometry-based cell surface protein atlas (CSPA) [214], (2) CellPhoneDB [216], a repository of curated receptors, ligands and their interactions, (3) the machine learning-based in silico human surfaceome [215] and (4) the human protein atlas (HPA) [217] v20.1 (https://www.proteinatlas.org/). Permissive integration of all datasets was critical, as cell surface expression showed strong variability between databases. Consequently, I used the union of these databases for all subsequent analyses (Figure 3.3d). Of the 96 genes overexpressed in HSPC-like cells, 36 genes were present in the generated library of human genes coding for proteins on the cell and were used for further downstream analyses.

Targets should not be expressed on T cells to avoid unintended CAR T fratricide that can result from shared target expression between malignant cells and CAR T cells [50, 241]. I identified enriched genes in T cells by comparing the mean expression of healthy T cells to the mean expression of all other healthy cell types using a t-test with overestimated variance. Testing was performed on the log-transformed normalized data to account for differences in sequencing depth between samples. For further target antigen filtering, I considered upregulated genes with false-discovery rate (FDR)-adjusted p values ≤ 0.01 and a log fold change >0. Genes that therefore passed all previous filters but showed high expression on T cells (for example, CD52 and CRIP1) were excluded from further analysis (Figure 3.3e).

To quantify off-target effects, I analyzed and combined a total of 11 scRNA Seq datasets across 9 healthy tissues [218–228]. I obtained raw annotated scRNA Seq data from the respective studies [218, 220, 222, 224–228] using the Python-based data repository *sfaira* [242]. To quantitatively analyze the expression of possible CAR T cell therapy targets across healthy tissues, I performed comparable preprocessing steps for each dataset separately, which involved removing low quality cells and lowly expressed genes, normalizing cell counts using *scran*, selecting highly variable genes based on normalized dispersion and visualizing the cells in a two-dimensional UMAP embedding as described above. For the lung datasets of Travaglini, Madissoon and Reyfman [221–223], I used publicly available data with cell annotations derived from a study integrating multiple scRNA Seq datasets [243]. To account for technical batches along the respective samples, I calculated batch-balanced k-nearest neighbors [156] for the datasets of Travaglini, Madissoon, Reyfman, Ramachandran, Cheng, James and Han [221–223, 225–228]. Finally, I concatenated processed and annotated count matrices on union variables and used the resulting matrix for target antigen filtering. Genes were excluded if they were expressed in over 2% of cells of a critical cell cluster (endothelial, arterial, bronchial, capillary, venous and smooth muscle cells) (Figure 3.3f).

To identify genes that encode druggable proteins, I used DrugBank [229], a database containing information on the interactions, pharmacology and chemical structures of drugs and drug targets.

I defined druggable genes as targets with known pharmacological action of FDA-approved drugs. Subsequently, I further restricted the gene list to include only those meeting these criteria.

Using this approach, two potential candidates for CAR development remained: colony-stimulating factor 1 receptor (CSF1R) and cluster of differentiation 86 (CD86). Interestingly, most of the described CAR targets for AML (n = 20) failed the thresholds of my stringent analyses at different levels Table 2. For example, prototypic AML antigens CD33 and CD123 did not fulfill my strict criteria of overexpression in malignant HSPCs, most likely due to expression of both antigens on healthy HSPCs. In addition, CD123 had high expression levels across endothelial and various lung cell types (see section 3.2 for detailed analysis).

Table 2: Current CAR targets in AML were cross-referenced to filters used for the single cell-based target screening approach. Log2FC HSC-/Prog-like >2: overexpressed on HSC-/Prog-like cells with log fold change >2 and FDR-adjusted p value ≤0.01, using a t-test with overestimated variance. Hyphen: Antigen did not fulfill respective threshold or criteria. Check mark: Thresholds or criteria were passed.

AML antigen	Log2FC HSC-like >2	Log2FC Prog-like >2	Expressed on cell surface	Low expression on T cells	Low expression on critical cell types	Target of FDA-approved drugs
CD13	✓	-	✓	✓	-	-
CD33	-	-	✓	✓	✓	✓
CD34	-	-	✓	✓	-	-
CD38	-	-	✓	✓	✓	-
CD44	-	-	✓	✓	✓	✓
CD70	-	-	✓	✓	✓	-
CD117	-	-	✓	✓	✓	✓
CD123	-	✓	✓	✓	-	✓
CLEC12A	-	-	✓	✓	✓	-
CSF2RA	-	-	✓	✓	✓	✓
CSF2RB	-	-	✓	✓	✓	✓
FLT3	-	-	✓	✓	✓	✓
FOLR1	-	-	✓	✓	-	-
IL1RAP	-	-	✓	✓	✓	-
NCAM1	-	-	✓	✓	✓	-
NKG2D	-	-	✓	✓	✓	-
LILRB4	-	✓	✓	✓	✓	-
LY6E	-	-	✓	-	-	-
PROM1	-	-	✓	✓	✓	-
SIGLEC6	-	-	✓	✓	✓	-

To the best of our knowledge, neither anti-CD86 nor anti-CSF1R CAR T cells have been implicated for CAR T cell therapy in AML. I thus decided to further investigate their potential.

3.2.2 Computational assessment of identified targets in AML and healthy cells

Next, I benchmarked the two identified target antigens CSF1R and CD86 to the reference genes CD123 and CD33 to ease interpretation of receptor expression on a transcriptomic level (Figure 3.4). CSF1R was expressed in all six malignant cell clusters, but was expressed the highest on monocyte-like or conventional dendritic cell-like clusters. CD86 was most strongly expressed in

monocyte-like, promonocyte-like and conventional dendritic cell-like clusters (Figure 3.4a,b). In terms of expression in malignant HSPC clusters, CSF1R expression was higher than CD86, albeit lower than CD123 and CD33 reference genes (Figure 3.4a,b). In contrast, CD123 or CD33 were detected in healthy HSCs and progenitors, while both CSF1R and CD86 were only minimally expressed among these cells (Figure 3.4c).

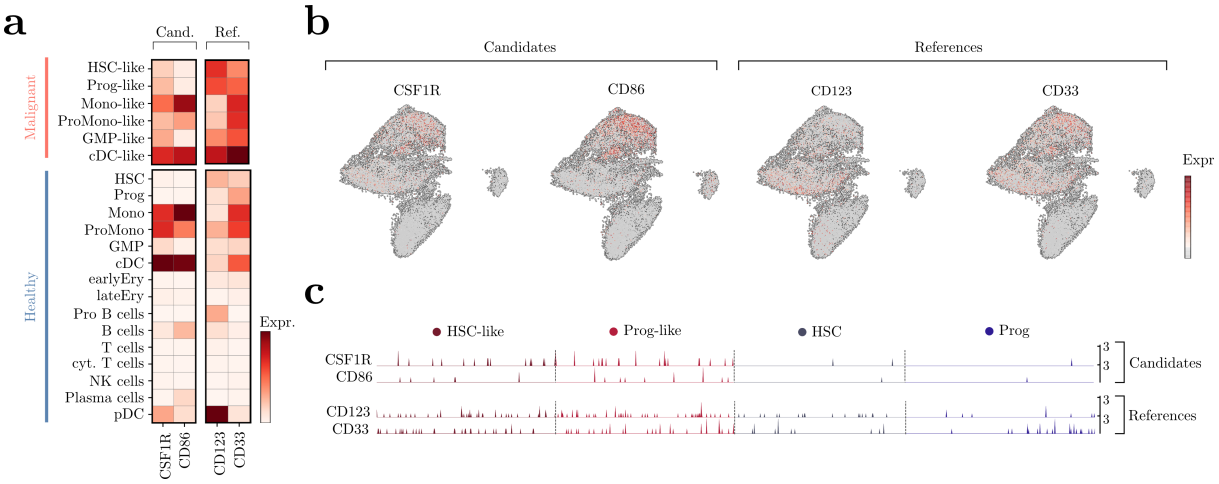

Figure 3.4: CSF1R and CD86 are preferentially expressed on malignant HSPC-like cells compared to healthy HSPCs. a) Expression of target and reference genes (CD123 and CD33) in single healthy and malignant cell types. Normalized expression values were log transformed and scaled to unit variance; Cand.: candidates; Ref.: references. **b)** Expression of CSF1R and CD86 target genes in healthy and malignant cells from 15 individuals with AML. Normalized gene expression values were log transformed and visualized in a UMAP embedding. **c)** Expression of CSF1R and CD86 target genes in malignant (HSC-like and Prog-like; left) and healthy (HSC and Prog; right) stem cells. For visualization purposes, normalized expression values of healthy HSPCs and a random subsample of malignant HSPCs were log transformed and scaled to unit variance. Each peak corresponds to a cell, and peak height indicates expression intensity. Figure adapted from Gottschlich et al. [127].

COOTA analysis revealed target antigen expression mainly in immune cells of myeloid origin (monocytes, macrophages and dendritic cells), similar to the peripheral expression profile of CD33 (Figure 3.5a). CSF1R and CD86 were not highly expressed on epithelial or stromal cells (Figure 3.5a,top). In organ-specific cell clusters (Figure 14a, bottom), expression was restricted to microglia cells in the brain, as described in the literature [244]. Additionally, in line with my COOTA prediction, CSF1R is known to be expressed on microglia [245], raising additional safety concerns. ScRNA Seq analysis of single human and murine brain tissue [218] confirmed expression of CSF1R in microglia and similar expression patterns in tissue-resident myeloid cells (Figure 3.5b,c).

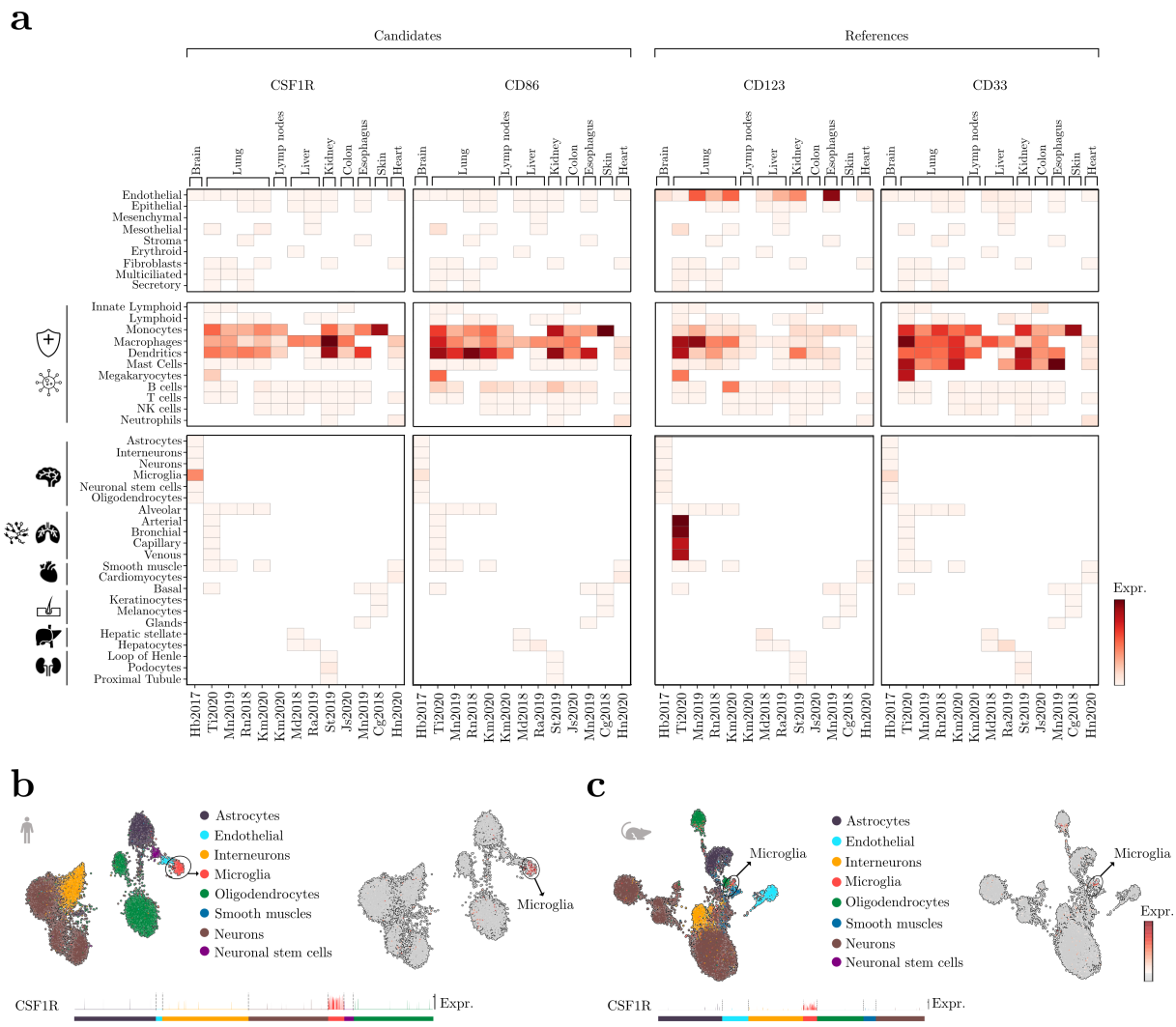

Figure 3.5: Off-tumor expression of CSF1R and CD86 is restricted to infiltrating or tissue-resident immune cells. a) Single-cell COOTA screening for target candidates (CSF1R and CD86) and reference (CD123 and CD33) genes. The single cell transcriptomic atlas consists of a total of 544,764 sequenced cells from nine different organs. Each field represents the mean expression value per cluster. Blank fields indicate cell types not present in a study. A detailed summary of all used datasets is provided in Table 1. **b)** Quantification of log-transformed normalized target expression in single human brain cells is shown on the left. Each peak corresponds to a cell, peak height indicates expression intensity. A UMAP plot illustrating the expression pattern of CSF1R in human brain cells is shown on the right. **c)** Csf1r expression in single mouse brain cells. A UMAP embedding of sequenced brain cells is shown on the left. Each peak corresponds to a cell, peak height indicates expression intensity. Normalized, log-transformed Csf1r expression per cell type is shown on the right. Figure adapted from Gottschlich et al. [127].

To ensure the validity of my analyses and to better reflect the cytogenetic diversity of AML as a disease, I next sought to further increase the size the patient cohort. Thus, I obtained a second publicly available scRNA Seq dataset of five additional individuals with AML [246] (Figure 3.6). For the cross-validation of my computational target identification approach, I used *scArches* [179], leveraging *scANVI*, a semi-supervised variational autoencoder [180], to map the unlabeled data

from Petti et al. [246] onto a newly generated reference map of van Galen et al. [240] (Figure 3.6a,b).

Briefly, I trained the *scVI* model on the raw, annotated reference data from van Galen et al. [240] with 2 hidden layers and a dropout rate of 0.2. Next, I initialized the *scANVI* model from the pretrained *scVI* model, and trained it for 20 epochs with 100 samples per label. Subsequently, I created a new query model instance before training the query data. I obtained a latent representation and label predictions before computing a neighborhood graph with 15 neighbors using the *scANVI* representation, and embedded the graph via UMAP as mentioned before (see methods section 2.1.8 of this thesis for a detailed explanation).

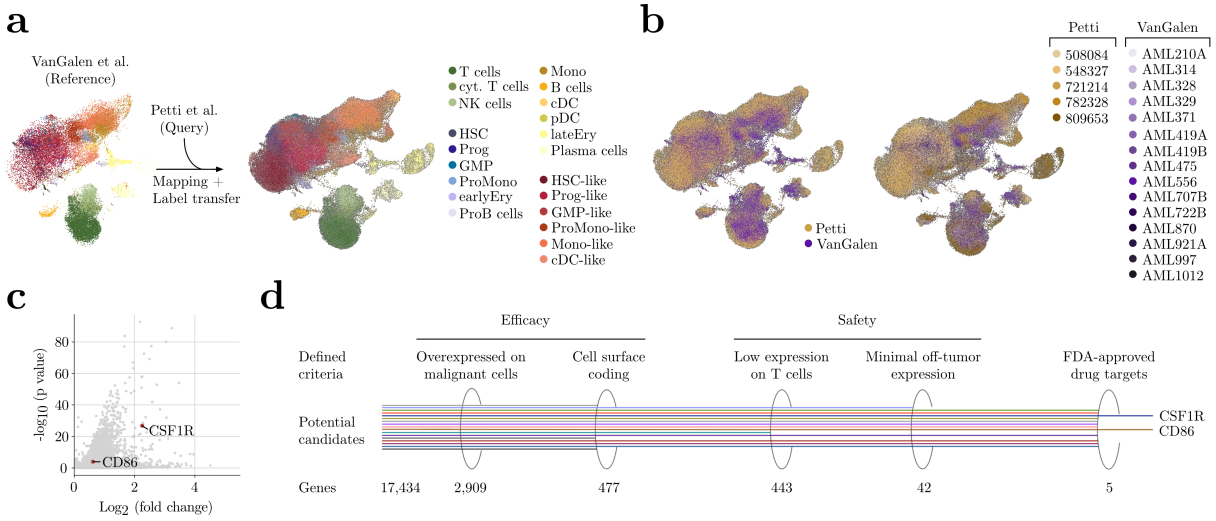

Figure 3.6: Analysis of an additional cohort of five AML patients identifies CSF1R and CD86 as suitable targets. a) Data from van Galen et al. [240] was used as a reference (left) to map cells from Petti et al. [246] (right) using *scANVI* [180]. **b)** UMAP representation showing the mapped query and reference data together (left) and the individual AML patients from the two datasets (right). **c)** Volcano plot showing CD86 and CSF1R target genes with their respective FDR-adjusted log10 p value and log fold changes from differential expression analysis between malignant HSPC-like and healthy HSPC using a t-test with overestimated variance. **d)** Computational CAR target antigen identification using the mapped dataset of Petti et al. [246] by stepwise evaluation against a set of criteria for an ideal and effective CAR target antigen. The decreasing number of screened AML target genes are shown on the bottom. Figure adapted from Gottschlich et al. [127].

In line with the results above, CSF1R and CD86 were preferentially expressed in malignant cells compared to healthy hematopoietic cells (Figure 3.6c). I applied my target identification approach to these five additional AML patients and again identified both CSF1R and CD86 as suitable target antigens for CAR therapy in this AML cohort (Figure 3.6d).

Both antigens were highly expressed across malignant cells in 100% of the individuals with AML with captured malignant blasts (Figure 3.7a,b), despite the heterogeneous molecular profile of the

participant collective [240, 246].

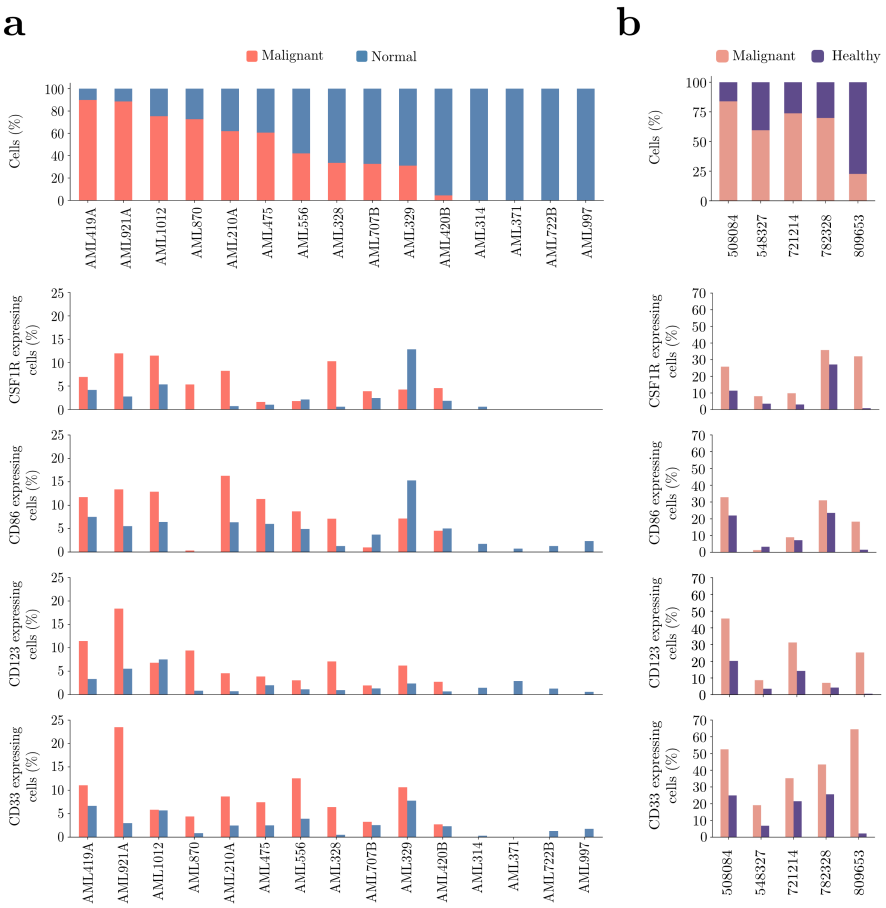

Figure 3.7: CSF1R and CD86 are consistently expressed across multiple patients with differing molecular subtypes. a-b) Top: Amount of malignant and normal cells per AML patient of van Galen et al. [240] **(a)** or Petti et al. [246] **(b)**. Bottom: Percentage of malignant and normal cells expressing target genes CSF1R, CD86 and reference genes CD123, CD33 for each sequenced AML patient of van Galen et al. [240] **(a)** or Petti et al. [246]**(b)**. Figure adapted from Gottschlich et al. [127].

In summary, using two independent scRNA Seq AML cohorts consisting of a total of 20 individuals and over 500,000 single cells from vital healthy human tissues, I identified CSF1R and CD86 as potential CAR targets for AML therapy.

3.2.3 Functional validation of CAR targets

Following the identification of CSF1R and CD86 as promising candidates for CAR T cell therapy in AML, functional validation of these CAR T cells was carried out in vivo and in vitro. Note that the experimental work that ensued was primarily conducted by Adrian Gottschlich and Ruth Grünmeier under the supervision of Prof. Sebastian Kobold (see section 1.5 of this thesis for a detailed summary of individual contributions). As the focus of this thesis lies on the

computational approach, I refer to the corresponding manuscript for a detailed description of the subsequent functional validation, including the full list of authors involved and their respective contributions.

Briefly, after having verified the expression of CSF1R and CD86 on a protein level, CAR T cells against both targets were developed. Efficacy of CSF1R and CD86-directed CAR T was extensively tested in vitro and in vivo using syngeneic mouse models, human-derived models, human cell lines and primary AML blasts. Assessment of advanced in vitro primary cell cultures for target-expressing cell types demonstrated a better discriminatory capacity than established anti-CD33 CAR T cells. In addition, utilizing several in vivo and in vitro models, we were able to mitigate safety concerns regarding potential off-target toxicity.

In summary, functional validation of these established CAR T cells shows robust in vitro and in vivo efficacy with minimal off-target toxicity toward relevant healthy human tissues, providing a strong rationale for further clinical development.

3.3 Discussion

Despite the remarkable progress made in the treatment of B cell lymphomas, a broad applicability of CAR T cells beyond B cell malignancies is still hindered by a lack of safe target structures. A major concern in target identification strategies followed so far is their à posteriori nature, where safety is considered after efficacy and primarily not for target selection or therapeutic design. Shifting this paradigm to a criteria-based target identification from first principles requires the integration of gene expression of all cells in all relevant tissues. Until now, such an endeavor was technically not feasible. Conveniently, the revolution in single cell technologies in the last decade generated massive single cell expression data that provide detailed information about the transcriptomic anatomy of healthy and malignant cells [104, 211], a yet mostly untapped resource for therapeutic development. In the context of CAR T cell target screening, these advancements allow in-depth on- and off-tumor antigen prediction [104, 212], offering unique insights into patients and various cell types at an unmatched resolution.

I developed the first scRNA Seq approach for de novo target identification and in-depth, high resolution off-tumor mapping across multiple tissues that is specifically tailored to predict potential candidates for CAR T cell therapy. Applying the approach to AML, I identified two previously unrecognized target antigens: CSF1R and CD86. Extensive in vitro and in vivo validation revealed broad expression on AML blasts, strong and durable treatment responses of newly developed CAR T cells in vitro and in vivo and minimal toxicities toward relevant healthy cells and tissue.

For target screening, I leveraged single cell sequencing data from 15 primary AML specimens with differing cytogenetic and mutational profiles [240], providing a direct representation of patient heterogeneity and disease complexity. These primary samples served as invaluable resources for identifying potential therapeutic targets with clinical significance. In addition, I validated the obtained results in an independent cohort of five additional AML patients [246]. This cohort was primarily employed for validation due to its small sample size and lack of crucial clinical information regarding AML subtype and patient mutational profiles. Yet, given the highly complex molecular landscape of AML, rare AML subtypes might still not be fully represented in my analyses. In addition, it is important to acknowledge the inherent selection bias in data inclusion from both malignant and healthy tissues used for target antigen filtering, as data selection for COOTA was based on data quality and comparability, as well as the relevance of specific tissues for assessing safety issues with CAR targets. Despite these limitations, this study clearly demonstrates the translational potential of scRNA Seq-based screening approaches and provides proof of principle of the whole spectrum of scRNA Seq-guided drug development spanning from computational target identification to preclinical investigation of newly developed CAR T cells.

CSF1R has been previously implicated as a target for small-molecule inhibition in AML [247]. However, its expression was thought to be restricted to a small subset of AML-supportive cells in certain individuals, while the majority of human blasts are regarded as antigen negative [248]. Using single cell gene expression data along with protein expression measurements, I was able to confirm high CSF1R expression on AML blasts. These reported ambiguities of CSF1R expression on malignant AML blasts encourage the use of RNA-based screening algorithms for target identification and prioritization, as methodological or biological confounders can easily mask protein expression analysis. Nevertheless, it is crucial to bear in mind that scRNA Seq-centered strategies come with their own limitations, such as the zero or dropout problem [249] and in any case require protein validation.

Unsurprisingly, CSF1R was expressed on microglia, which share a common monocytic precursor, as also known for CD33 [250]. Clinical investigation of the so far only CSF1R-directed monoclonal antibody did not reveal neurotoxicity as a concern when depleting CSF1R-positive cells from the periphery [251]. However, given the different mode of action of cellular- versus antibody-based therapies, these results might not be directly transferable to anti-CSF1R CAR T cell therapy. In addition, CAR T cells are known to be able to cross the blood-brain barrier [252, 253], and peak levels of proinflammatory cytokines further increase permeability of this tightly regulated barrier [254, 255]. Because of these considerations, we rigorously tested the possibility for neurotoxicity in fully syngeneic mouse models, in which we implanted large quantities of CAR T cells directly into mouse brains. Yet, we did not observe any signs of neurotoxicity. Nevertheless, future clinical validations will need to include well-designed protocols to vigilantly detect any signs of neurotoxicity.

CD86 is expressed on malignant AML blasts, and high receptor expression is associated with shortened overall survival of individuals with AML [256, 257], but, to the best of our knowledge, CD86 has never been explored as a target for (immuno)therapy of cancer. The expression of CD86 is not limited to AML and has also been reported in numerous B cell malignancies [258]. As such, the use of CD86 CAR T cells promises not only treatment options for AML but also applications for a variety of other hematological malignancies, such as multiple myeloma [259] and acute lymphoblastic leukemia [260]. Nevertheless, CD86 is also expressed on healthy macrophages and dendritic cells [261–263] and might increase the risk of immunosuppression and ensuing severe infection. However, CTLA-4 fusion proteins, such as abatacept (targeting both CD80 and CD86), have received approval by the FDA and are clinically used for the treatment of autoinflammatory disorders [264]. In clinical studies, Abatacept was generally well tolerated [264].

For both CSF1R and CD86, the measured antigen densities were rather low, especially compared to the high CD33 expression. Yet, despite our extensive functional validation, we did not observe marked differences between CSF1R and CD86 CAR T cells compared to established CD33 CAR T cells. Along these lines, we did not observe a correlation between lysis capacity of CAR T cells

and site density of the respective target antigen. To a certain extent, these findings are in line with recent reports observed for anti-mesothelin CAR T cells in solid tumors [265]. Several factors, such as affinity and binding properties of the used single-chain variable fragment (scFv) and conformation of the target antigen, can positively or negatively influence CAR–tumor cell interactions. Ultimately, while high target antigen expression undoubtedly increases killing efficacy, our data suggest that in some cases, functional cross-comparison might help to identify promising target antigens, despite on first glance rather low antigen expression.

Similar to previous results in AML [50], I was not able to identify target antigens with expression limited to a single immune cell lineage, as is the case for CD19 or BCMA in B cell malignancies. However, expression of prime candidates CSF1R and CD86 is limited to immune cells of myeloid origin (monocytes, tissue-resident macrophages and dendritic cells), with minimal detection on stem or progenitor cells. Thus, these candidates could bear the advantage of clinical application without the risk for severe bone marrow toxicity, which is a current concern of AML-targeted treatments [232]. It should be noted, however, that to date, clinical outcome of off-tumor gene expression on HSPCs remains elusive. Along these lines, precise projection of off-tumor antigen expression is one of the central objectives of my single cell approach, because unwanted toxicity may be inferred from high transcriptomic off-target antigen expression [52, 104, 212]. Yet, as outlined above, the risk of severe adverse effects caused by off-tumor activity of CAR T cells is not fully understood, and different outcomes have been reported [266]. As such, the latest trials evaluating the safety of CD123 CAR T cells did not show sustained cytopenia [266].

However, in most anti-CD123 CAR T cell trials currently being conducted, participants eventually received allogeneic hematopoietic stem cell transplantation, which presumably eradicated CAR T cells. Of note, the development of fatal cytokine release syndrome and capillary leak syndrome following CD123 CAR T cell infusion, potentially due to off-target expression of CD123 on small vessels, has been reported [210]. Altogether, current clinical evidence does not support a clear definition of the critical cell types and expression thresholds that would preclude the development of CAR T cells against a certain target to avoid unmanageable toxicities. In the long run, detailed knowledge of off-tumor expression will allow vigilant monitoring of 'high risk off-tumor organs' in clinical trials and might enable rapid side effect-mitigating treatments. Similarly, clinical lessons from anti-CD19 or anti-BCMA CAR T cell therapy deem lineage-restricted expression patterns as highly desirable, providing further strong evidence for the use of single cell technologies for de novo target identification, as this might aid the search for unrecognized target antigens with minimal off-tumor expression in healthy tissues.

I implemented stringent criteria during target identification, such as overexpressed genes on malignant cell populations with log fold change >2 and COOTA expression filtering thresholds of 2% in critical cells, which inherently limited the number of identified targets. This stringency was deliberate to ensure the selection of high-confidence targets with minimal off-target

effects. Also, while this approach is robust in identifying potential CAR targets, it's important to acknowledgethat parameter choices played a crucial role in shaping the final selection of targets, with thresholds for overexpression and COOTA inclusion impacting the final outcome. I iteratively adjusted parameters based on domain knowledge and preliminary validations across diverse scRNA Seq datasets to ensure the robustness of results across diverse datasets. Future studies could benefit from comprehensive robustness analyses to further validate the stability and generalizability of this approach.

Many of the currently investigated CAR targets in AML failed the thresholds of overexpression on malignant HSPCs compared to their healthy counterparts in my analyses. Herein, to a certain extent, this data contradict data from publications of colleagues [235, 267]. Sauer et al., for example, illustrated higher expression of CD70 in bone marrow biopsies of individuals with AML than in bone marrow samples of healthy donors by using immunohistochemistry [235]. This discrepancy is most likely due to my restrictive analyses, in which I have chosen rather high cutoff criteria to ensure maximal safety of identified target antigens. Adjustment of these thresholds will yield different results, and many of the previously identified target antigens (for example, CD123, CD33, CD70, FLT3, C-type lectin-like molecule-1 and CD44v6) will most likely be of aid to improve clinical care of individuals with refractory or relapsed AML. Precision medicine in AML at the individual patient level remains challenging due to the considerable variation observed among patients [268]. Nonetheless, my approach provides a foundation for future endeavors aimed at refining and customizing CAR T cell therapies. The data clearly demonstrates the value of CSF1R and CD86 as targets for CAR T cell therapy in AML, and, considering the complex molecular landscape of AML and its highly diverse subsets, these targets are expected to be valuable additions to the immunotherapeutic repertoire in AML.

In summary, my results highlight the potential of using high resolution, single cell transcriptomic data for CAR T cell target selection and development. Leveraging these data and the appropriate high dimensional analyses as standard operating procedures enables identification of potential new target structures for targeted immunotherapy in a wide range of malignant disorders.

4 Estimating on-tumor efficacy and off-tumor toxicity of CAR targets by screening single cell gene expression

While CAR T cell therapy has made significant progress in recent years, its ability to achieve the desired therapeutic effect depends on targeting tumor-specific antigens, as shared target expression on healthy tissues can lead to severe on-target, off-tumor toxicities (see section 1.1.2). Administration of approved CAR T cells have shown a wide range of side effects, such as neurotoxicity, hematotoxicity, and cytokine release syndrome. Importantly, these were retrospective findings, discovered after introduction of these treatments into clinical practice. Flow cytometry has traditionally served as the primary method for immunophenotypic measurements, but its low throughput and reliance on predetermined hypotheses limit exploratory molecular profiling of CAR T cells and their targets [269].

In this chapter, I therefore examined single cell gene expression profiles of CAR targets derived from clinical trials involving malignancies treated with CAR T cell therapy approved by the FDA. The primary objective of this chapter was to identify potential differences that could account for the observed toxicities in clinical settings. Furthermore, I sought to evaluate the expression profiles of CAR targets in malignant and healthy tissues and identify novel targets that demonstrate a favorable expression pattern across both malignant and healthy cells.

To that end, I leveraged single cell gene expression data from over 300,000 cells obtained from patients suffering from follicular lymphoma, multiple myeloma, and acute lymphoblastic leukemia. Additionally, I conducted an extensive screening of CAR target expression in over 3 million cells derived from various healthy tissues, including bone marrow, brain, lung, lymph nodes, spleen, heart, liver, and kidney. Through my analysis, I identified notable disparities in the expression levels of CAR targets and interpreted novel clinical outcome data with regard to the expression of CAR targets across both healthy and malignant cell populations. Finally, I identified novel CAR targets for the treatment of follicular lymphoma, multiple myeloma, and acute lymphoblastic leukemia that exhibit a favorable gene expression pattern across both healthy and malignant cell populations using a computational screening approach based on single cell gene expression data.

My findings provide guidance in identifying potential off-tumor toxicities and aid in the selection of CAR targets in a wide range of hematological malignancies prior to their translation into clinical applications. Consequently, this study serves as a resource for both retrospective and prospective evaluation of therapeutic targets, offering significant support for enhancing the safety and efficacy of future CAR T cell therapies.

This chapter is based on and partly identical to the following publication:

1. **Thomas, M.***, Brabenec, R.*, ... Kobold, S. and Marr, C., 2023. Single-cell gene expression screening for efficacy and safety prediction of CAR T cell therapy targets. *In preparation.*

Note that "*" marks an equal contribution. See section 1.5 of this thesis for a detailed summary of individual contributions.

4.1 Single cell transcriptomics for CAR target tumor efficacy estimation

With the development and approval of CD19- or BCMA-directed CAR T cells, the field of cellular immunotherapy has seen tremendous progress in recent years and has evolved as a part of the standard for patients suffering from various relapsed or refractory B cell malignancies [37–39]. Enormous effort has been invested in moving these advanced therapeutic products into other hematological malignancies and, importantly, solid tumors, with limited reported success [56, 270]. A major determinant of CAR T cell safety but also efficacy is the choice of antigen. This should be ideally ubiquitously expressed on the cancer cell and only there, be of key importance to the disease and of no or very limited expression in healthy tissues [43, 50]. In reality, however, these characteristics rarely apply and antigen expression is shared with other cells and tissues. Such "on-target, off-tumor" effects [43] can result in a range of often severe, long-term and sometimes life threatening consequences, including cytokine release syndrome (CRS), neurotoxicity, and serious organ damage [22, 43].

For instance, treatment with anti-CD19 CAR T cells has been associated with severe neurotoxicity, hematotoxicity, as well as CRS in varying frequencies [271, 272]. These side effects have been partially linked to target expression and target availability, for example in the case of neurotoxicity or cytokine release [43, 270]. Importantly, these findings were retrospective and discovered only after the introduction of these treatments into clinical practice, fortunately proving to be manageable to a large extent [43, 273]. In other cases, however, CAR-induced toxicity might be less manageable or even fatal. Examples include CAIX (CA9)-targeting CAR T cells used for treatment of renal cell carcinoma (RCC) that caused hepatobiliary toxicity, most likely due to shared CAIX expression on the bile duct epithelium [54, 55]. Cardiorespiratory failure was fatal in a patient suffering from colorectal cancer upon treatment with CAR T cells targeting HER2, due to reactivity against pulmonary and cardiac tissue [56, 57]. Along these lines, it appears evident that a better a priori understanding of both the product and the target might help to mitigate risks and prioritize concepts with higher potential for efficacy with controllable safety.

The rapid growth, resolution and availability of scRNA Seq data has revolutionized our understanding of complex cellular biology [105, 274, 275]. Emerging large-scale atlases of healthy human tissues and tumor mass enable therapeutic target screening across various cell populations on a transcriptomic level, a significant aid in predicting potential risks of on-target, off-tumor toxicities and for assessing the target specificity on malignant cells for CAR T cell therapy [104, 276].

Parker et al. demonstrated the efficacy of scRNA Seq atlases in evaluating the safety properties of CAR targets by associating neurotoxicity of CD19-targeting CAR T cells with CD19 expression in a small subset of brain mural cells responsible for maintaining blood-brain barrier permeability [52]. Similarly, neurotoxicity in a patient suffering from multiple myeloma treated with

BCMA-targeting CAR T cells was linked to BCMA expression in the caudate nucleus of healthy human brain tissues [53]. Previous studies have predominantly focused on exploring a specific subset of cells expressing the target of interest, rather than considering the broader context of the global expression profile [52, 53]. However, with the rapid increase in clinical trials and the consequent expansion of CAR targets, an extensive analysis of the global expression profiles of CAR targets on both healthy and malignant tissues is lacking [212, 277].

To investigate whether the global gene expression profile of a CAR target could serve as an indicator for toxicity effects, I focused my analyses on malignancies for which CAR T cell therapy has FDA approval. For these malignancies, we then screened ongoing clinical trials utilizing FDA-approved or experimental CAR target antigens for overall response rates, as well as reported rates of neurotoxicity, hematotoxicity and cytokine release syndrome among patients who underwent CAR T cell therapy. Finally, I analyzed transcriptomic profiles of these CAR targets in malignant and healthy cell populations and interpreted global expression patterns of these targets in conjunction with clinical patient outcome data (Figure 4.1).

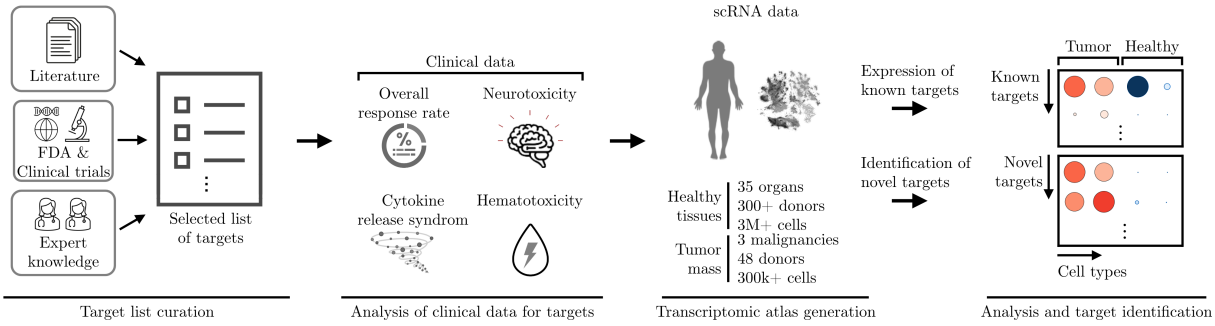

Figure 4.1: Computational assessment of efficacy and safety of CAR targets from clinical trials. CAR targets were acquired through screening the literature and clinical trials. Clinical data provided insights about the efficacy and safety of CAR targets. Target gene expression was evaluated computationally on malignant and healthy tissues for CAR targets from clinical trials. Using a single cell gene expression-based approach, novel CAR targets were identified.

The subsequent sections detail the different steps, including curating the CAR target list from clinical trials, analyzing clinical outcome data derived from patients that underwent CAR T cell therapy, generating a transcriptomic atlas encompassing malignant and healthy cell populations, and conducting an in-depth analysis and interpretation of target expression profiles in relation to the available clinical outcome data. Finally, using a scRNA Seq-based screening approach, I identified novel candidates for CAR T cell therapy in these malignancies, which exhibit promising expression patterns across both healthy and malignant cell populations.

4.1.1 Approved and investigational CAR targets in clinical trials

To determine the malignancies of interest, I initially aimed to identify all malignancies for which the FDA has approved CAR T cell therapy. As of February 2023, there are a total of six FDA approved CAR T cell therapies, all utilizing CD19 or BCMA (TNFRSF17) as target antigens to treat patients suffering from follicular lymphoma (FL), multiple myeloma (MM), acute lymphoblastic leukemia (B-ALL), large B cell lymphoma (LBCL), diffuse large B cell lymphoma (DLBCL), primary mediastinal B cell lymphoma (PMBCL), high-grade B cell lymphoma (HGBL), and mantle cell lymphoma (MCL) (Table 3).

Table 3: FDA-approved CAR T therapeutics and their respective target antigens. Targets were obtained from the website http://www.fda.gov as of February 2023.

Therapeutic	Approved for	Targeting antigen
ABECMA	Refractory MM	BCMA (TNFRSF17)
BREYANZI	Relapsed or refractory LBCL	CD19
CARVYKTI	Relapsed or refractory MM	BCMA (TNFRSF17)
KYMRIAH	Relapsed or refractory DLBCL	CD19
	Relapsed or refractory B-ALL	CD19
TECARTUS	Relapsed or refractory MCL	CD19
	Relapsed or refractory B-ALL	CD19
YESCARTA	DLBCL	CD19
	PMBCL	CD19
	HGBL	CD19
	DLBCL resulting from FL	CD19
	FL	CD19

To identify additional targets for CAR T cell therapies in clinical trials, we performed a comprehensive screening of www.clinicalTrials.gov. The applied search criteria included the intervention/treatment terms "CAR T Cell Therapy," "CAR T," and "CAR" applied to the above mentioned malignancies. For FL treatment, CD19 is currently the only FDA-approved product targeting a single antigen (Table 3). However, two additional targets, CD20 and CD22, are being investigated in clinical trials as of February 2023 (Table 4).

Table 4: Targets from clinical trials in follicular lymphoma (FL). Targets were obtained from the website http://www.clinicaltrials.gov as of February 2023.

Target	Gene	Status	Malignancy	ClinicalTrials Identifier
CD19	CD19	Recruiting	FL	NCT05326243
CD20	MS4A1	Recruiting	FL	NCT03277729
CD22	CD22	Unknown status	FL	NCT03999697

Regarding the treatment of MM, BCMA serves as the only target antigen for the two FDA-approved CAR T cell products (Table 3). Furthermore, we identified eight additional antigens under investigation for multiple myeloma in clinical trials (Table 5).

In the case of B-ALL treatment, CD19 is the sole target for FDA-approved CAR T cell therapies (Table 3). We found five additional antigens being explored in clinical trials (Table 6).

Table 5: Targets from clinical trials in multiple myeloma (MM). Targets were obtained from the website http://www.clinicaltrials.gov as of February 2023.

Target	Gene	Status	Malignancy	ClinicalTrials Identifier
BCMA	TNFRSF17	Not yet recruiting	MM	NCT05181501
CD19	CD19	Unknown status	MM	NCT04182581
CS1	SLAMF7	Completed	MM	NCT03958656
GPRC5D	GPRC5D	Active, not recruiting	MM	NCT04555551
NKG2D	KLRK1	Completed	MM	NCT02203825
CD138	SDC1	Unknown status	MM	NCT01886976
CD38	CD38	Terminated	MM	NCT03473496
CD44v6	CD44	Terminated	MM	NCT04097301
CD56	NCAM1	Terminated	MM	NCT03473496

Table 6: Targets from clinical trials in acute lymphoblastic leukemia (B-ALL). Targets were obtained from the website http://www.clinicaltrials.gov as of February 2023.

Target	Gene	Status	Malignancy	ClinicalTrials Identifier
CD19	CD19	Withdrawn	B-ALL	NCT04094766
BAFFR	TNFRSF13C	Recruiting	B-ALL	NCT04690595
CD22	CD22	Withdrawn	B-ALL	NCT04094766
CD20	MS4A1	Recruiting	B-ALL	NCT05418088
CD123	IL3RA	Active, not recruiting	B-ALL	NCT02159495
ROR1	ROR1	Recruiting	B-ALL	NCT05694364

B cell malignancies, such as large B cell lymphoma (LBCL), diffuse large B cell lymphoma (DLBCL), primary mediastinal B cell lymphoma (PMBCL), high-grade B cell lymphoma (HGBL), and mantle cell lymphoma (MCL), are of great clinical significance due to their prevalence and high rates of morbidity and mortality. They represent a diverse set of malignancies, ranging from aggressive to indolent forms of lymphoma. Currently, CD19 is the only FDA-approved target antigen for all these B cell malignancies (Table 3). We identified seven additional antigens being investigated in various clinical trials (Table 7).

Table 7: Targets from clinical trials across diverse B cell malignancies. Targets were obtained from the website http://www.clinicaltrials.gov as of February 2023.

Target	Gene	Status	Malignancy	ClinicalTrials Identifier
CD19	CD19	Recruiting	LBCL, DLBCL, PMBCL	NCT04812691
CD19	CD19	Recruiting	HGBL	NCT05418088
CD19	CD19	Recruiting	MCL	NCT04484012
CD20	MS4A1	Recruiting	LBCL, DLBCL, PMBCL	NCT03277729
CD20	MS4A1	Recruiting	HGBL	NCT05418088
CD20	MS4A1	Enrolling by invitation	MCL	NCT04911478
CD22	CD22	Unknown status	LBCL, DLBCL, PMBCL	NCT03999697
CD22	CD22	Recruiting	HGBL	NCT05418088
CD22	CD22	Unknown status	MCL	NCT02721407
BAFFR	TNFRSF13C	Recruiting	MCL	NCT05370430
ROR1	ROR1	Recruiting	LBCL, DLBCL, PMBCL	NCT05694364
ROR1	ROR1	Terminated	MCL	NCT02706392
CD79a	CD79A	Recruiting	LBCL, DLBCL, PMBCL	NCT05169489
CD30	TNFRSF8	Completed	LBCL, DLBCL, PMBCL	NCT03049449
CD137	TNFRSF9	Unknown status	MCL	NCT02685670

4.1.2 Analysis of clinical outcomes in CAR T cell therapy patients

We subsequently screened clinical trials involving CAR T cell therapy that reported patient outcome data. The screening process was primarily carried out by Lisa Gregor under supervision of Prof. Sebastian Kobold (see section 1.5 of this thesis for a detailed summary of individual contributions). We obtained a list of ongoing clinical trials utilizing CAR T cell therapy as of March 2023 by examining the websites http://www.clinicaltrials.gov (search criteria: "car-t" or "chimeric antigen receptor", filtering by completed and terminated trials) and https://pubmed.ncbi.nlm.nih.gov/ (search criteria: "car-t" or "chimeric antigen receptor", filtering by clinical trials). Papers of respective clinical studies were obtained from https://pubmed.ncbi.nlm.nih.gov/ and were screened to extract the clinical trial identifier, which was then used to eliminate duplicates and exclude any trials evaluating non-CAR T products. All trials were screened by two independent individuals to avoid human error and were carefully assessed to acquire specific information about the CAR target employed for treating each malignancy.

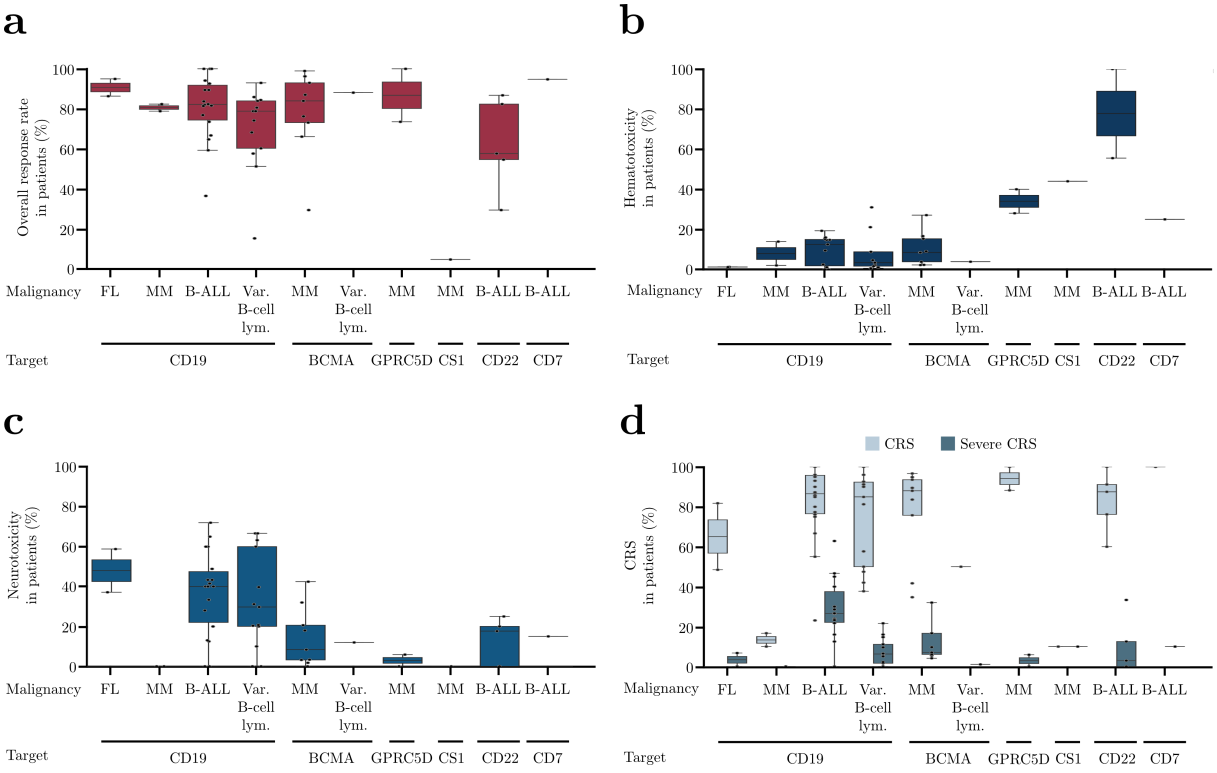

Figure 4.2: Clinical patient outcome data following administered CAR T cell therapies. **a)** Overall response rate (ORR) showing the percentage of treatment-responsive patients following CAR T cell therapy for each target and treated malignancy. ORR was calculated by dividing the number of patients exhibiting partial or complete response by the total number of patients involved in the study. **b-d)** Percentage of patients showing signs of hematotoxicity **(b)**, neurotoxicity **(c)**, as well as CRS (defined as stages 1 and 2) and severe CRS (defined as stages 3 and 4) **(d)** for each target and treated malignancy. Each dot represents one clinical study.

For each study, we further analyzed the number of patients exhibiting signs of progressive disease, stable disease, partial response, and complete response, which we used to calculate the overall response rate (ORR) by dividing the number of patients exhibiting partial or complete response by the total number of patients involved in the study. In addition, we collected relevant clinical information pertaining to CAR safety, such as the number of patients displaying signs of hematotoxicity (incorporating anemia, lymphopenia, thrombocytopenia, and leukopenia), neurotoxicity, or cytokine release syndrome (CRS).

The data obtained presents a highly heterogeneous profile with respect to response rates and observed adverse effects across targets and malignancies (Figure 4.2). Analyzing the ORR (Figure 4.2a), no discernible trend is evident among CAR targets with an average ORR between 60% and 90%. However, compared to other targets, the only available clinical study administering CAR T cells targeting CS1 resulted in a low ORR of 10%. Patients receiving CD19-directed CAR T cells displayed the highest incidence of neurotoxicity (Figure 4.2b), while showing relatively low rates of hematotoxicity (Figure 4.2c). Hematotoxicity was most commonly observed in patients treated with CARs targeting CD22, CS1, GPRC5D, and CD7 (Figure 4.2c). CRS rates (stages 1/2) and severe CRS rates (stages 3/4) showed similar patterns across targets, with no discernible trend (Figure 4.2d).

4.1.3 Harmonization of single cell gene expression data

To investigate the potential for adverse on-target off-tumor effects resulting from high expression of CAR target antigens in healthy cell populations, I conducted an extensive screening of 32 scRNA Seq datasets based on data quality and comparability, as well as the relevance of specific tissues for assessing safety issues with CAR targets (Figure 4.3, Table 8). These datasets encompassed a total of 3,039,758 cells and covered 35 healthy tissues throughout the human body [218–228, 240, 278–296]. Additionally, I examined CAR target expression patterns in over 300,000 single cells obtained from patients suffering from FL, MM, and B-ALL [297–299]. Given the limited availability of scRNA Seq data for other B cell malignancies, such as LBCL, DLBCL, PMBCL HGBL, and MCL, I analyzed the expression profiles of CAR targets used for treating these malignancies in healthy cell populations.

I acquired raw and processed count data for cells from healthy and malignant tissues through data repositories *sfaira* [242] and *cellxgene* (https://cellxgene.cziscience.com). Similar to section 3.2.1 of this thesis, I performed comparable preprocessing steps for each dataset. I converted count data to the anndata format when necessary [300]. For raw and processed count data, I filtered barcodes to retain only high quality cells based on the overall distributions of UMI counts and genes and the fraction of mitochondria-encoded genes per cell after visual examination of each sample, if necessary. See Table 8 for an overview of the utilized scRNA Seq datasets and applied filtering thresholds. For each dataset, I excluded genes detected in fewer than 20 cells from

subsequent analyses. I normalized UMI counts of each cell using the *scran* algorithm [148].

Next, I identified top 4000 variable genes based on normalized dispersion, as previously described [136]. To capture the underlying data structure in two dimensions, I carried out principal component analysis (PCA) dimension reduction by computing 15 principal components on highly variable genes. Next, I constructed a neighborhood graph on the first 50 principal components before embedding the neighborhood graph via UMAP [161]. Cell annotation labels provided by the respective study authors were thoroughly inspected and, if necessary, relabeled to facilitate global comparisons of cell populations. To account for technical batches, such as sequencing depth or library preparation between datasets, I used *scVI* [157] to integrate datasets of respective healthy tissues in an organ-wise manner (see methods section 2.1 of this thesis for a detailed explanation of applied preprocessing steps).

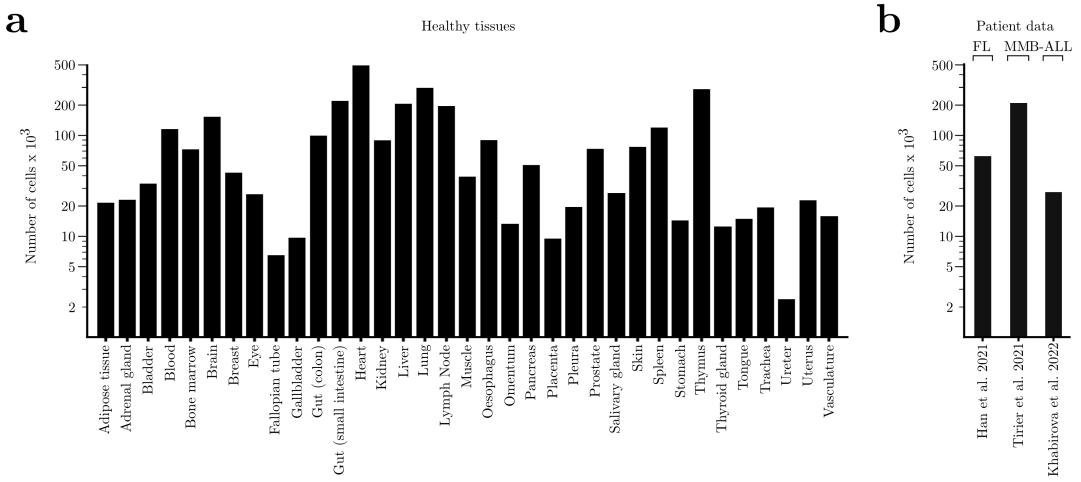

Figure 4.3: Computational assessment of CAR target expression on healthy and malignant cell populations. a) Number of single cells after quality control in scRNA Seq datasets from 35 healthy human tissues used to quantify CAR target off-tumor expression on healthy cell populations. **b)** Number of single cells in scRNA Seq datasets from follicular lymphoma (FL), multiple myeloma (MM) and acute lymphoblastic leukemia (B-ALL) tumors used to quantify CAR target on-tumor expression.

To obtain reliable cell annotations for B-ALL cells, I mapped raw count data of unlabeled B-ALL cells [299] to an annotated reference of fetal bone marrow cells [301] using the semi-supervised variational auto-encoder model *scANVI* [180]. Briefly, I trained the *scVI* model on the raw reference data with 2 hidden layers and a dropout rate of 0.2. I initialized the *scANVI* model from the pretrained *scVI* model, before training it for 20 epochs with 100 samples per label. Next, I created a new query model instance before training the query data. I obtained a latent representation and label predictions before computing a neighborhood graph with 15 neighbors using the *scANVI* representation, and embedded the graph via UMAP as mentioned before.

Table 8: ScRNA Seq datasets and cell filtering thresholds used to assess CAR target expression. Hyphen: Threshold was not applied. C_{max}/G_{max} = upper thresholds for filtering cells according to captured UMI counts (C_{max}) or genes (G_{max}). MT_{max} = upper thresholds for filtering cells according to their fraction of mitochondria-derived genes. See methods section 2.1.1 for a definition and explanation about filtering thresholds.

Condition	Entity or Tissue	C_{max}	G_{max}	MT_{max}	Reference
Tumor	FL	20,000	6,000	20%	Han et al., 2021 [297]
Tumor	MM	80,000	7,500	20%	Tirier et al., 2021 [298]
Tumor	B-ALL	-	-	-	Khabirova et al., 2021 [299]
Healthy	Multiple tissues	4,000	2,000	20%	Han et al., 2020 [228]
Healthy	Multiple tissues	16,000	9,000	20%	The Tabula Sapiens Consortium, 2022 [278]
Healthy	Multiple tissues	8,000	4,3000	20%	Conde et al., 2022 [279]
Healthy	Multiple tissues	4,000	1,700	20%	Szabo et al., 2019 [280]
Healthy	Brain	14,000	4,000	25%	Habib et al., 2017 [218]
Healthy	Brain	-	-	25%	Berg et al., 2020 [281]
Healthy	Brain	-	-	20%	Bakken et al., 2021 [282]
Healthy	Eye	45,000	10,000	20%	Lukowski et al., 2019 [283]
Healthy	Kidney	-	7,500	20%	Stewart et al., 2019 [220]
Healthy	Kidney	4,000	2,500	20%	Muto et al., 2021 [284]
Healthy	Thymus	-	-	25%	Park et al., 2020 [285]
Healthy	Lymph nodes	-	-	25%	Xiang et al., 2020 [286]
Healthy	Lymph nodes Lung	50,000	6,000	20%	Kim et al., 2020 [219]
Healthy	Lung	30,000	5,500	20%	Reyfman et al., 2018 [223]
Healthy	Lung	50,000	6,000	20%	Travaglini et al., 2020 [221]
Healthy	Lymph nodes Oesophagus	30,000	5,500	20%	Madissoon et al., 2019 [222]
Healthy	Colon	60,000	5,000	20%	James et al., 2020 [227]
Healthy	Colon	9,000	6,000	20%	Fawkner-Corbett et al., 2021 [287]
Healthy	Colon Rectum	38,000	5,500	25%	Wang et al., 2019 [288]
Healthy	Liver	25,000	4,000	20%	MacParland et al., 2018 [224]
Healthy	Liver	42,000	9,000	20%	Andrews et al., 2022 [289]
Healthy	Liver	-	-	20%	Ramachandran et al., 2019 [225]
Healthy	Pancreas	-	-	20%	Baron et al., 2016 [290]
Healthy	Pancreas	-	-	20%	Peng et al., 2019 [291]
Healthy	Pancreas	-	7,500	20%	Enge et al., 2017 [292]
Healthy	Pancreas	-	-	20%	Muraro et al., 2016 [293]
Healthy	Heart	4,500	1,600	10%	Litviňuková et al., 2020 [294]
Healthy	Bone Marrow	17,000	4,200	20%	van Galen et al., 2019 [240]
Healthy	Skin	-	-	20%	Cheng et al., 2018 [226]
Healthy	Prostate	7,000	3,000	20%	Henry et al., 2018 [295]
Healthy	Breast	40,000	7,000	20%	Bhat-Nakshatri et al., 2021 [296]

In summary, I leveraged publicly available scRNA Seq data derived from diverse healthy tissues (Figure 4.3a) and FL, MM and B-ALL tumors (Figure 4.3b). I employed this gene expression data to investigate the gene expression patterns of CAR targets that have undergone clinical trials for the treatment of malignancies for which the FDA has approved CAR T cell therapy.

4.2 Analysis of CAR targets across malignant and healthy cells

The success of CAR T cell therapy relies heavily on the precise targeting of tumor-specific antigens, enabling specific recognition and subsequent elimination of cancer cells while minimizing damage to healthy tissues. Achieving target selectivity is crucial for reducing side effects and enhancing the overall safety profile of the therapy. If the target antigen is expressed at very low levels or is absent on cancer cells, CAR T cells may fail to recognize and eliminate malignant cells. Therefore, it is essential to thoroughly evaluate the expression levels and density of target antigens on malignant cells prior to administering CAR T cell therapy.

4.2.1 Tumor expression of approved and investigational CAR targets

I therefore analyzed expression levels of approved and currently investigated targets for the treatment of FL, MM and B-ALL (Figure 4.4). For FL, CD19 is presently the sole target antigen approved for a single-target CAR T cell therapy (Table 3). Clinical trials are currently investigating two additional targets: CD20 and CD22 (Table 4). The expression of these antigens is restricted to malignant B-cell lymphoma cells, plasma cells, and non-malignant B cells (Figure 4.4a-c), with variations in overall and mean expression levels (Figure 4.4b-c). Notably, CD20 also demonstrated antigen expression on erythroid cell types (Figure 4.4c).

Regarding MM, BCMA currently serves as the only target antigen for the two FDA-approved CAR T cell products (Table 3). We identified eight additional antigens being explored in clinical trials for MM (Table 5). CAR targets for the treatment of MM exhibited high expression levels on tumor cells overall (Figure 4.4d-f). NKG2D demonstrated the lowest overall expression across tumor and healthy cell populations. BCMA, CS1, and SDC1 were found to be coexpressed on plasma or B cells, while also displaying high expression levels on tumor cells. CD44 was found to be strongly coexpressed across a wide range of healthy cell populations (Figure 4.4e-f).

In the case of B-ALL, CD19 is currently the sole target for FDA-approved CAR T cells (Table 3). We identified five additional antigens for B-ALL that are under investigation in clinical trials (Table 6). All CAR targets employed in the treatment of B-ALL exhibited moderate expression levels within malignant cells. Specifically, both CD19 and CD22 were expressed in malignant and healthy cells belonging to the B cell lineage (Figure 4.4g-i). Conversely, CD20 exhibited elevated expression within healthy cells of the B cell lineage but relatively lower expression levels within malignant cells of the same lineage (Figure 4.4h-i).

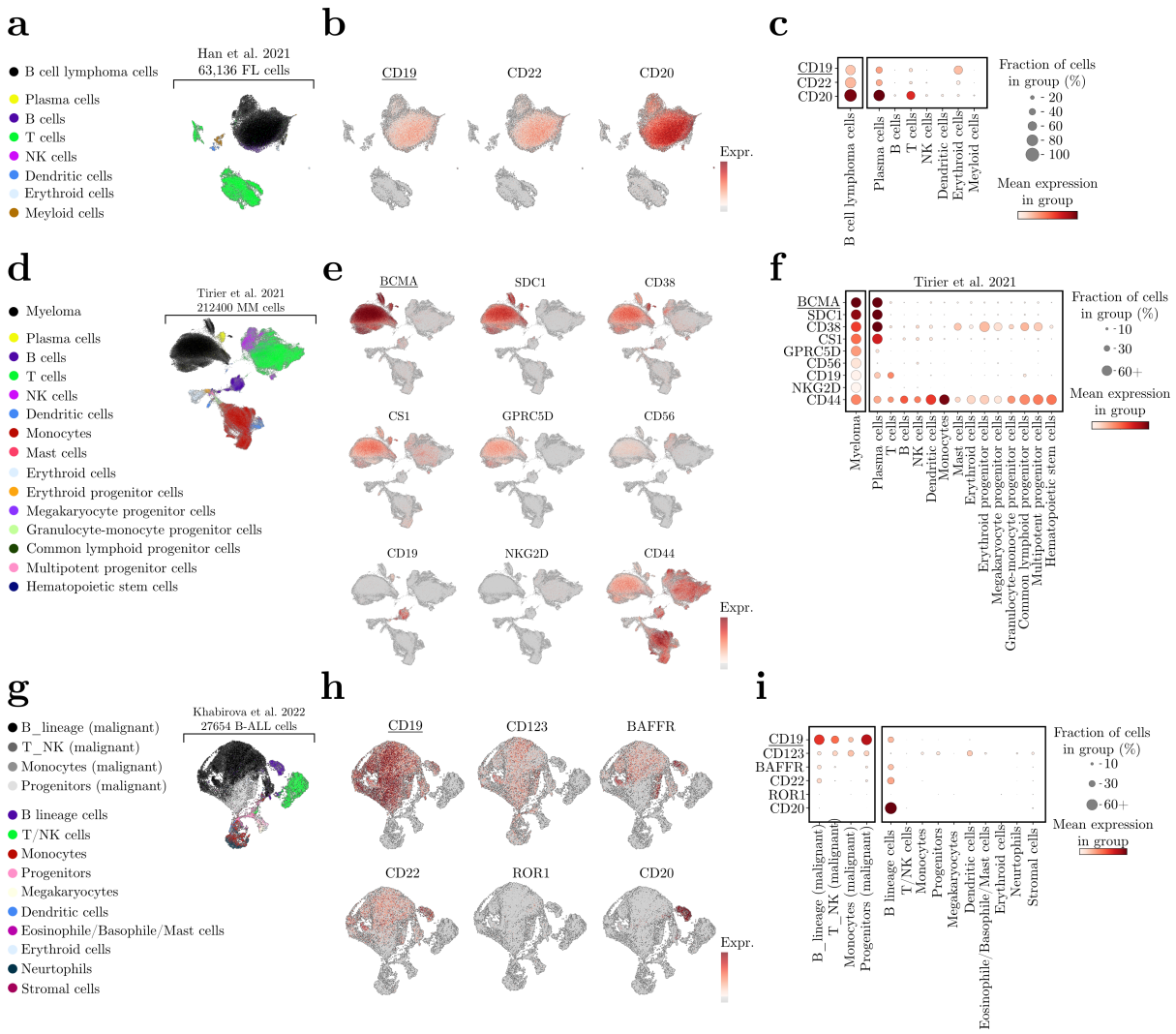

Figure 4.4: Assessment of tumor expression of CAR targets for the treatment of follicular lymphoma (FL), multiple myeloma (MM) and acute lymphoblastic leukemia (B-ALL). a,d,g) UMAP showing 63,236 healthy and malignant cells from FL **(a)**, 212,400 healthy and malignant cells from MM **(d)** and 27,654 healthy and malignant cells from B-ALL **(g)**. Normalized gene expression values were logarithmized. Colors highlight the different cell types. Annotations of FL and MM cells were provided by the authors. Annotations of B-ALL cell types were obtained by matching cells to a reference of fetal bone marrow cells [301]. **b,e,h)** Expression of CAR targets from clinical trials across FL **(b)**, MM **(e)** and B-ALL **(h)** datasets. Normalized gene expression values were logarithmized and visualized in a UMAP embedding. FDA-approved targets have been underlined. Targets were ordered according to their smallest euclidean distance in gene expression space to the approved targets for each malignancy. **c,f,i)** Expression of CAR targets on tumor and healthy cells for FL **(c)**, MM **(f)** and B-ALL **(i)** datasets. Dot size indicates the fraction of cells per cell type expressing a target, color intensity shows mean normalized gene expression per cell type. FDA-approved targets have been underlined. Targets were ordered according to their smallest euclidean distance to the approved targets.

Cancer is a highly heterogeneous disease, leading to variations in the expression of target antigens not only across different cancer types but also among patients with the same cancer type [302]. The presence of antigen heterogeneity poses challenges in predicting and identifying suitable candidates for CAR T cell therapy. Patient eligibility for CAR T cell therapy is often determined based on the expression of the target antigen. Therefore, when the antigen exhibits variability in expression among patients, it becomes essential to accurately evaluate antigen expression in individual patients to ensure their suitability for the therapy. I therefore evaluated the expression of target antigens in individual patients suffering from FL, MM and B-ALL (Figure 4.5).

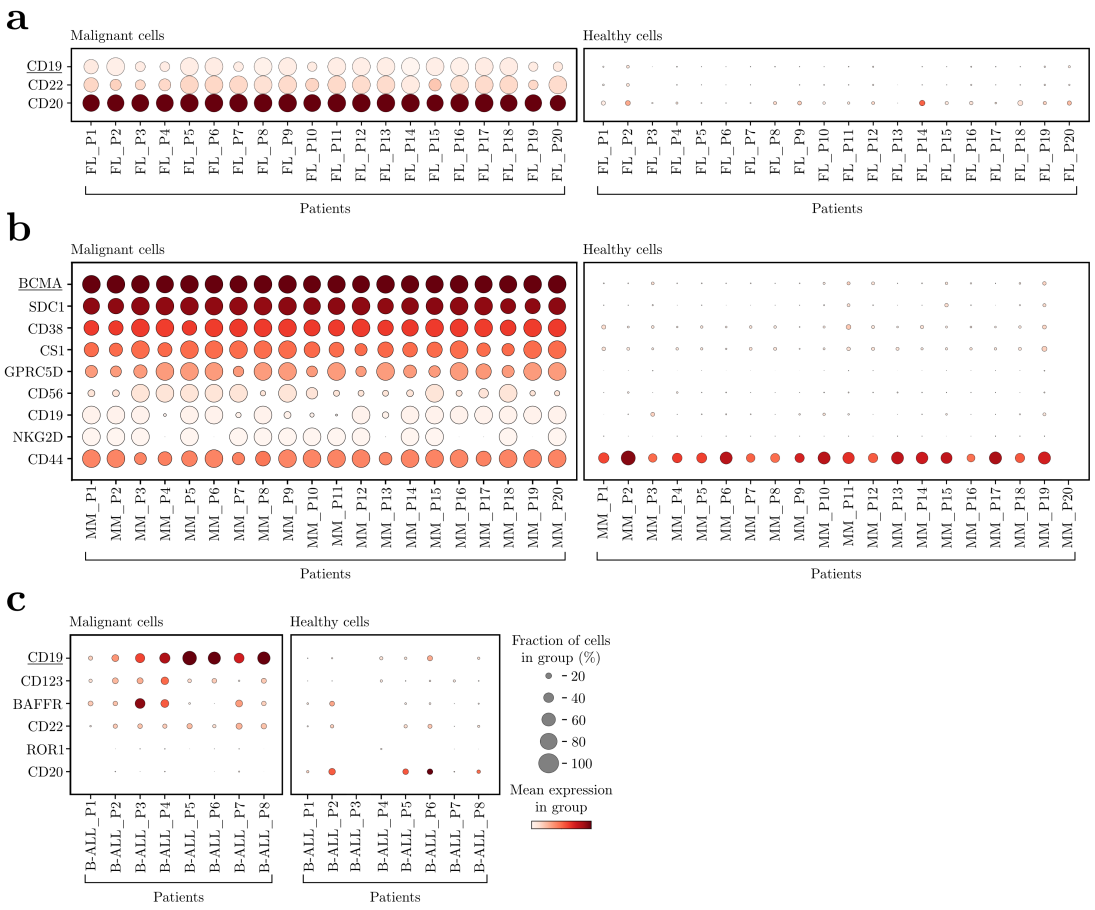

Figure 4.5: CAR target expression on healthy and malignant cells per patient. a-c) Expression of CAR targets on healthy and malignant cells per patient for FL **(a)**, MM **(b)**, and B-ALL **(c)** datasets . Dot size indicates the fraction of cells per cell type expressing a target, color intensity shows mean normalized gene expression per cell type.Targets were ordered according to their smallest euclidean distance in gene expression space to the approved targets for each malignancy.

All targets employed in FL treatment consistently exhibited higher expression levels in malignant cells compared to healthy cell populations across all 20 patients in the FL dataset (Figure 4.5a). Similarly, targets for MM treatment demonstrated preferentially higher expression in malignant cells across all 20 patients compared to healthy cell populations (Figure 4.5b). Notably, in MM,

CD19, NKG2D, and CD56 did not exhibit strong expression across malignant cells from all patients. CD44 showed strong expression across healthy cell populations in the majority of patients (Figure 4.5b). For targets related to B-ALL treatment, their expression densities were generally low, resulting in an even distribution between malignant and healthy cell populations (Figure 4.5c). However, CD19 and BAFFR exhibited the greatest difference in expression between malignant and healthy cells per patient, with predominant expression observed across malignant cell populations for each patient (Figure 4.5c).

In summary, I meticulously examined the expression of CAR targets in malignant cells using three scRNA Seq datasets comprising over 300,000 cells from FL, MM, and B-ALL [297–299] (Figure 4.3b). It should be noted that scRNA Seq data is currently unavailable for other B-cell lymphomas mentioned in Table 3.

4.2.2 Healthy expression profiles of approved and investigational CAR targets

CAR T cells are designed to recognize and attack cells that express the target antigen upon infusion into patients. However, there is a risk of unintended damage to healthy tissues if the target antigen is also expressed on healthy cells. This can result in adverse toxicity effects, particularly when the target antigen is found on vital organs or tissues. Therefore, it is crucial to carefully select target antigens for CAR T cell therapy that exhibit high expression in cancer cells while showing limited or no expression in healthy tissues. Achieving this selectivity is challenging but essential to maximize the therapeutic index of CAR T cell therapy.

To investigate the potential association between high expression of CAR target antigens in healthy cell populations and the observed adverse on-target off-tumor effects, I conducted a comprehensive screening of CAR targets used for the treatment of FL, MM, B-ALL and various other B cell malignanciesacross over 3 million single cells derived from multiple healthy tissues throughout the human body (Table 8, Figure 4.6).

Regarding the targets with patient outcome data availability, I observed that CD19 and BCMA expression was predominantly restricted to B cells or cells of the B cell lineage, such as plasma cells (Figure 4.6a-h). GPRC5D also exhibited a similar pattern, although with lower overall expression compared to CD19 and BCMA. Additionally, GPRC5D showed minimal expression in brain cell populations, including astrocytes, microglial cells, and neurons (Figure 4.6b). CD22 expression was mostly confined to B cells, but was also expressed in brain oligodendrocytes (Figure 4.6b) and myeloid cells, such as basophils or mast cells (Figure 4.6c-d). CS1 demonstrated a broader expression pattern across immune cells in multiple tissues, including T cells, NK cells, and myeloid cells. CD7 exhibited a broader expression across immune cells, present in a wider range of cell populations, including T cells (Figure 4.6a-h).

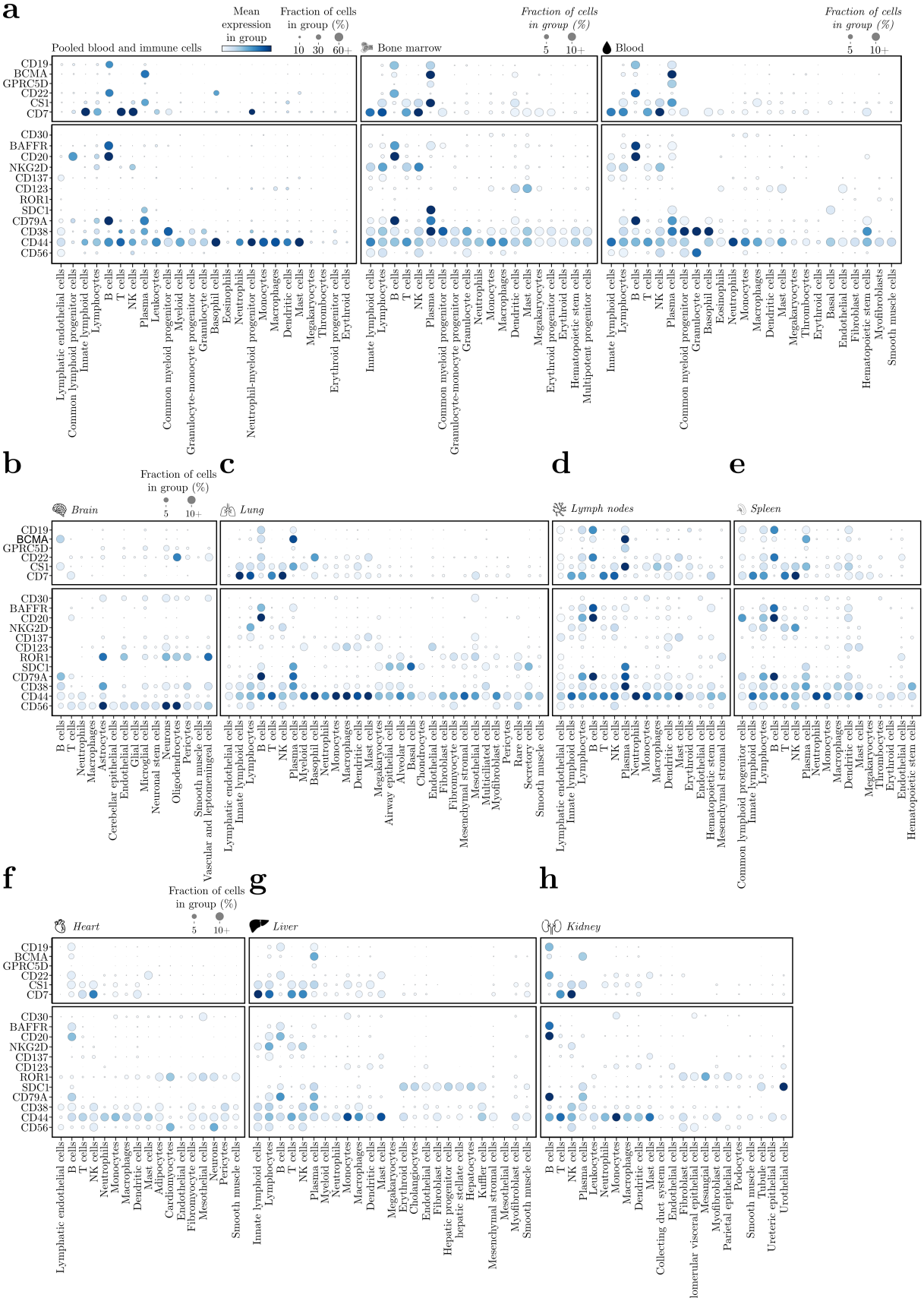
a
Mean expression in group
Fraction of cells in group (%)
10 30 60+
Pooled blood and immune cells
Bone marrow
Blood
5 10+
CD19
BCMA
GPRC5D
CD22
CS1
CD7
CD30
BAFFR
CD20
NKG2D
CD137
CD123
ROR1
SDC1
CD79A
CD38
CD44
CD56
Lymphatic endothelial cells
Common lymphoid progenitor cells
Innate lymphoid cells
Lymphocytes
B cells
T cells
NK cells
Plasma cells
Leukocytes
Common myeloid progenitor cells
Myeloid cells
Granulocyte-monocyte progenitor cells
Granulocyte cells
Basophil cells
Eosinophils
Neutrophils
Neutrophil-myeloid progenitor cells
Monocytes
Macrophages
Dendritic cells
Mast cells
Megakaryocytes
Thrombocytes
Erythroid progenitor cells
Erythroid cells
Innate lymphoid cells
Lymphocytes
B cells
T cells
NK cells
Plasma cells
Common myeloid progenitor cells
Granulocyte-monocyte progenitor cells
Granulocyte cells
Neutrophils
Monocytes
Macrophages
Dendritic cells
Mast cells
Megakaryocytes
Erythroid progenitor cells
Erythroid cells
Hematopoietic stem cells
Multipotent progenitor cells
Innate lymphoid cells
Lymphocytes
B cells
T cells
NK cells
Plasma cells
Common myeloid progenitor cells
Granulocyte cells
Basophil cells
Eosinophils
Neutrophils
Monocytes
Macrophages
Dendritic cells
Mast cells
Megakaryocytes
Thrombocytes
Erythroid cells
Basal cells
Endothelial cells
Fibroblast cells
Hematopoietic stem cells
Myofibroblasts
Smooth muscle cells
b
Fraction of cells in group (%)
5 10+
Brain
B cells
T cells
Neutrophils
Macrophages
Astrocytes
Cerebellar epithelial cells
Endothelial cells
Glial cells
Microglial cells
Neuronal stem cells
Neurons
Oligodendrocytes
Pericytes
Smooth muscle cells
Vascular and leptomeningeal cells
c
Lung
Lymphatic endothelial cells
Innate lymphoid cells
Lymphocytes
B cells
T cells
NK cells
Plasma cells
Myeloid cells
Basophil cells
Neutrophils
Monocytes
Macrophages
Dendritic cells
Mast cells
Megakaryocytes
Airway epithelial cells
Alveolar cells
Basal cells
Chondrocytes
Endothelial cells
Fibroblast cells
Fibromyocyte cells
Mesenchymal stromal cells
Mesothelial cells
Multiciliated cells
Myofibroblast cells
Pericytes
Rare cells
Secretory cells
Smooth muscle cells
d
Lymph nodes
Lymphatic endothelial cells
Innate lymphoid cells
Lymphocytes
B cells
T cells
NK cells
Plasma cells
Neutrophils
Monocytes
Macrophages
Dendritic cells
Mast cells
Erythroid cells
Endothelial cells
Hematopoietic stem cells
Mesenchymal stromal cells
e
Spleen
Common lymphoid progenitor cells
Innate lymphoid cells
Lymphocytes
B cells
T cells
NK cells
Plasma cells
Neutrophils
Monocytes
Macrophages
Dendritic cells
Mast cells
Megakaryocytes
Thrombocytes
Erythroid cells
Endothelial cells
Hematopoietic stem cells
f
Fraction of cells in group (%)
5 10+
Heart
Lymphatic endothelial cells
B cells
T cells
NK cells
Neutrophils
Monocytes
Macrophages
Dendritic cells
Mast cells
Adipocytes
Cardiomyocytes
Endothelial cells
Fibromyocyte cells
Mesothelial cells
Neurons
Pericytes
Smooth muscle cells
g
Liver
Innate lymphoid cells
Lymphocytes
B cells
T cells
NK cells
Plasma cells
Myeloid cells
Neutrophils
Monocytes
Macrophages
Dendritic cells
Mast cells
Megakaryocytes
Erythroid cells
Cholangiocytes
Endothelial cells
Fibroblast cells
Hepatic progenitor cells
hepatic stellate cells
Hepatocytes
Kuffler cells
Mesenchymal stromal cells
Mesothelial cells
Myofibroblast cells
Smooth muscle cells
h
Kidney
B cells
T cells
NK cells
Plasma cells
Leukocytes
Neutrophils
Monocytes
Macrophages
Dendritic cells
Mast cells
Collecting duct system cells
Endothelial cells
Fibroblast cells
Glomerular visceral epithelial cells
Mesangial cells
Myofibroblast cells
Parietal epithelial cells
Podocytes
Smooth muscle cells
Tubule cells
Ureteric epithelial cells
Urothelial cells

Figure 4.6: Gene expression screening of CAR targets on healthy cell populations. **a)** Expression of CAR targets on pooled immune cells from all healthy tissues (left), bone marrow (middle) and blood (right) tissues. **b-h)** Expression of CAR targets across brain **(b)**, lung **(c)**, lymph nodes **(d)**, spleen **(e)**, heart **(f)**, liver **(g)** and kidney **(h)** tissues. Dot size indicates the fraction of cells per cell type expressing a target, color intensity shows mean normalized gene expression per cell type. Targets with and without clinical data were respectively ordered according to their smallest euclidean distance in gene expression space to the approved targets.

For target antigens without patient outcome data from clinical trials, CD30 exhibited to most similar expression pattern compared to approved targets and was expressed in a narrow range of immune cells, particularly monocytes, but also in various brain cell types, such as astrocytes, microglial cells, and neurons (Figure 4.6b). BAFFR expression was mainly confined to B cells and plasma cells (Figure 4.6a-h). CD20 expression was mostly confined to cells of the B cell lineage, such as B cells and plasma cells (Figure 4.6a-h). NKG2D expression was limited to a narrow range of immune cells, including T cells and NK cells (Figure 4.6a). CD137 showed low expression levels across immune cell types from multiple tissues, primarily lymphocytes and lymphatic endothelial cells (Figure 4.6a-h). CD123 was predominantly expressed in monocytes, macrophages, dendritic cells, and mast cells but also exhibited expression in neurons (Figure 4.6b). ROR1 exhibited a narrow expression pattern across immune cells but was highly expressed in vital cell types of the brain, including astrocytes, neurons, oligodendrocytes, microglia, and was also expressed in cardiomyocytes of the heart (Figure 4.6b,f). SDC1 expression was predominantly restricted to plasma cells, but could also be detected in specific cell populations of the liver (Figure 4.6g) and kidney (Figure 4.6h). CD79A was predominantly restricted to B cells and plasma cells (Figure 4.6a-h). CD38 was observed across a broad spectrum of immune cell types, including lymphatic endothelial cells, innate lymphoid cells, T cells, and NK cells, but also in astrocytes, neurons, and pericytes (Figure 4.6b,f). CD44 displayed the broadest expression across screened tissues, with elevated expression levels in almost all cell types (Figure 4.6a-h). CD56 was expressed in endothelial cells (Figure 4.6a) and showed high expression in various brain cell types, including astrocytes, cerebellar epithelial cells, microglial cells, neurons, and oligodendrocytes (Figure 4.6b).

In summary, I have generated an extensive transcriptional atlas comprising over 3 million cells derived from healthy tissues. I examined CAR target expression globally across blood and immune cells from various healthy human tissues, and assessed target expression in healthy tissues such as the bone marrow, blood, lymph nodes, brain, lung, spleen, heart, liver, and kidney, which are considered crucial in the pathogenesis and dissemination of B cell malignancies [303].

4.2.3 Interpreting CAR target expression profiles with clinical outcome data

With the increasing availability of CAR T clinical data through the accumulation of numerous clinical trials, a collection of patient outcome data has started to emerge. This valuable resource provides significant insights into the efficacy, safety, and long-term consequences of CAR T cell therapy in real-world clinical settings. Leveraging this wealth of patient outcome data, facilitates a better understanding of the intricate relationship between molecular characteristics, treatment response, and toxicity, thereby potentially identifying key factors that influence the success of the therapy.

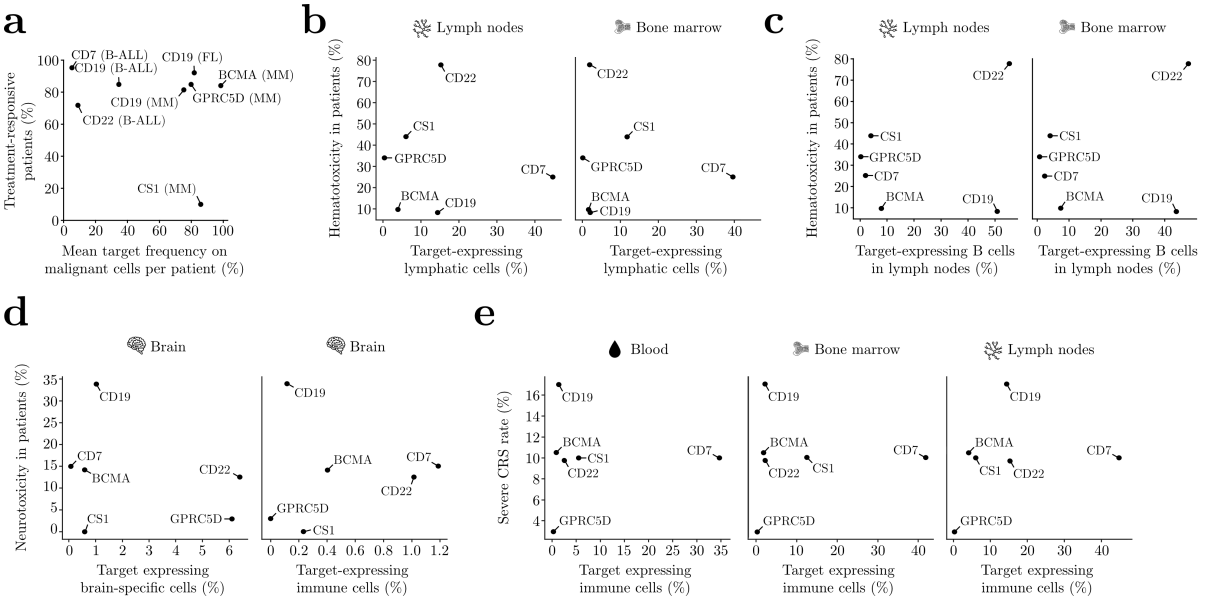

Figure 4.7: Reported CAR efficacies and toxicities do not arise from a direct causal relationship with global CAR target gene expression profiles. a) Correlation between treatment-responsive patients (calculated by dividing the number of patients exhibiting partial or complete response by the total number of patients involved in the study) and mean CAR target frequency on malignant cell populations per patient. **b)** Correlation between clinically reported hematotoxicity in patients and percentage of CAR target-expressing lymphatic cells in lymph nodes (left) and bone marrow (right) tissues. **c)** Correlation between clinically reported hematotoxicity in patients and percentage of CAR target-expressing B cells in lymph nodes (left) and bone marrow (right) tissues. **d)** Correlation between clinically reported neurotoxicity in patients and percentage of CAR target-expressing brain-specific cells (left) and CAR target-expressing immune cells (right) in brain tissues. **e)** Correlation between clinically reported severe CRS (defined as stages 3 and 4) in patients and percentage of CAR target-expressing immune cells in blood (left), bone marrow (center), and lymph node tissues (left).

To explore the potential associations between global single cell gene expression profiles and reported patient outcomes, I examined response rates and survival outcomes in patients who had undergone CAR T cell therapy targeting specific antigens for various malignancies (Figure 4.7).

Intriguingly, my analysis did not reveal any significant correlation between clinical patient outcomes and global single cell gene expression profiles. However, I did observe a marginal increase in the overall response rate (ORR) that corresponded to the mean frequency of target gene expression on malignant cells per patient (Figure 4.7a). Remarkably, when assessing hematotoxicity (Figure 4.7b-c), neurotoxicity (Figure 4.7d), and CRS (Figure 4.7e), I observed a highly heterogeneous pattern that lacked a clear association between global scRNA Seq profiles of targets and the occurrence of toxicity effects. These findings suggest that these toxicities may not arise from a direct causal relationship with the global expression of target genes across healthy cell populations.

4.3 Computational identification of potential CAR target candidates

Target specificity is a critical determinant of successful CAR T cell therapy. Considering the substantial heterogeneity exhibited by certain targets across tumor and healthy cell populations, I sought to leverage single cell gene expression data to identify additional candidate targets for CAR T cell therapy in these malignancies. In chapter 3 of this thesis, I presented a computational approach that enables the identification of suitable targets for CAR T cell therapy based on single cell gene expression data. Consequently, I employed this target identification approach to analyze FL, MM, and B-ALL, resulting in the identification of multiple targets suitable for CAR T cell therapy against these malignancies (Figure 4.8). To the best of our knowledge, none of these targets have been utilized for CAR T cell therapy in the context of these malignancies thus far.

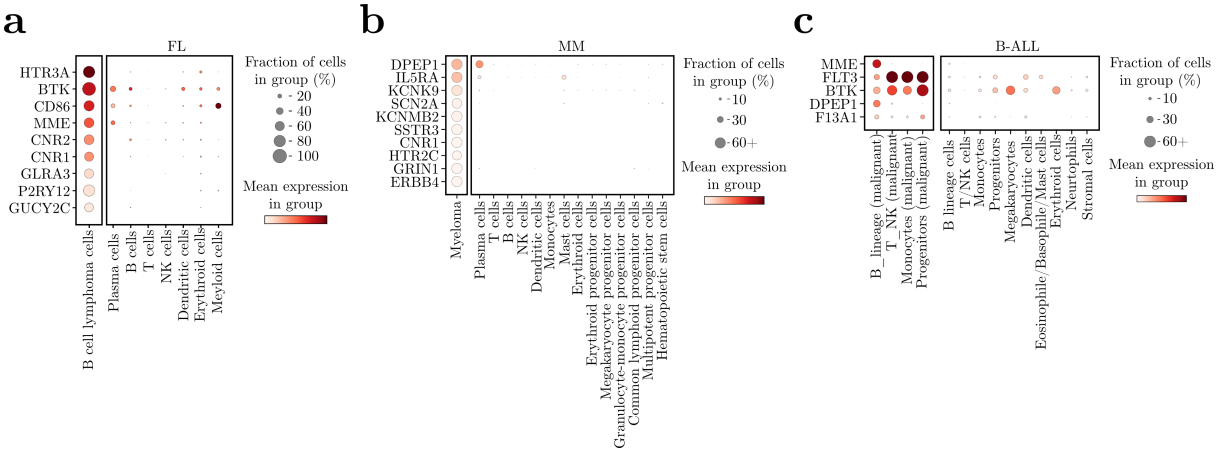

Figure 4.8: Computationally identified CAR target candidates are highly expressed across malignant cell populations. a-c) Expression of identified CAR target candidates on tumor and healthy cells for FL **(a)**, MM **(b)**, and B-ALL **(c)** datasets. Dot size indicates the fraction of cells per cell type expressing a target candidate, color intensity shows mean normalized gene expression per cell type.

Interestingly, for FL, my computational approach revealed CD20, currently under investigation in clinical trials, and FDA-approved CD19 as the top two hits. Additionally, I identified 9 novel candidates as potential targets for CAR T cell therapy in FL. These candidates exhibited remarkably high and highly restricted expression on B cell lymphoma cells (Figure 4.8a). Noteworthy among them is BKT (Bruton's tyrosine kinase), a crucial mediator in B cell receptor signaling, which has demonstrated remarkable effectiveness in mantle cell lymphoma treatment [304, 305]. Furthermore, CD86, which was previously identified for AML in chapter 3 of this thesis, has displayed promising therapeutic efficacy in AML treatment. Moreover, GUCY2C, a CAR target utilized for colorectal cancer treatment [306, 307], emerged as another candidate.

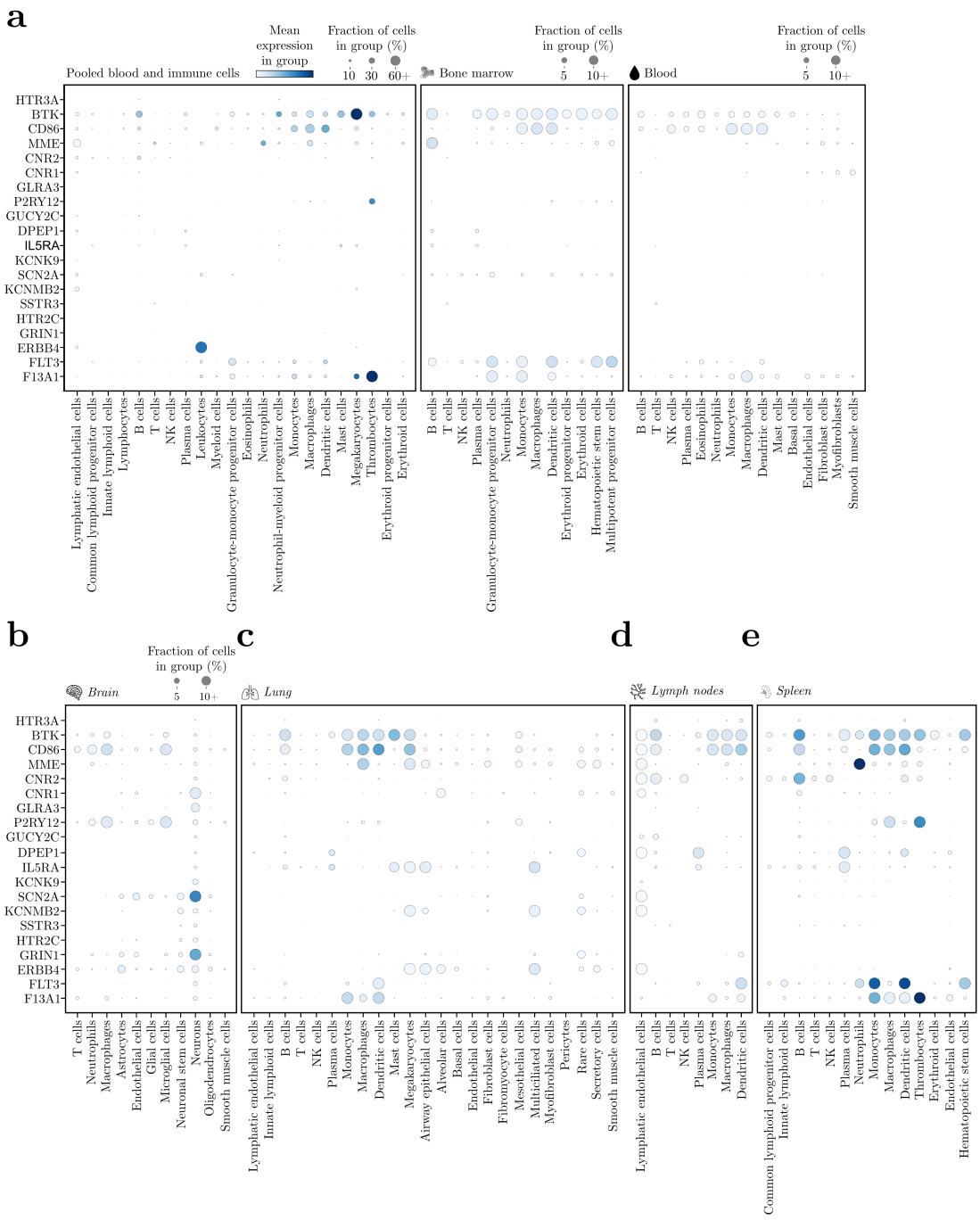

Figure 4.9: Gene expression screening of CAR target candidates on healthy cell populations. a) Expression of CAR target candidates on pooled immune cells from all healthy tissues (left), bone marrow (middle) and blood (right) tissues. **b-e)** Expression of CAR target candidates across brain **(b)**, lung **(c)**, lymph nodes **(d)** and spleen **(e)** tissues. Dot size indicates the fraction of cells per cell type expressing a target candidate, color intensity shows mean normalized gene expression per cell type.

In the context of MM, my computational approach identified BCMA as the top hit, which currently stands as the only approved CAR product for MM treatment. Furthermore, my analysis identified CS1, which is already undergoing investigation in clinical trials for MM treatment

(Table 5). Additionally, I identified ten potential targets that exhibited elevated expression levels on malignant cells (Figure 4.8b).

Regarding B-ALL, my analysis identified CD19, the sole approved CAR T cell therapy product against B-ALL, as the leading candidate. Furthermore, I identified five additional candidates, all displaying heightened expression across malignant cell populations (Figure 4.8c). Notably, MME, BTK, and DPEP1 emerged once again as potential candidates. Among the other candidates was FLT3, which has been previously utilized in AML treatment with CAR T cells [267, 308].

Subsequent to evaluating the expression profiles within malignant cell populations, I next assessed the expression of target candidates across various healthy cell populations. For this, I employed the transcriptomic atlas to examine the cross-tissue expression patterns of these targets (Figure 4.9). Remarkably, all identified targets exhibited minimal global expression across blood and bone marrow cells (Figure 25a), as well as brain, lung, lymph node, and spleen tissues (Figure 4.9b-e). BTK, CD86, MME, and FLT3 displayed the broadest expression patterns. Furthermore, P2RY12, SCN2A, and GRIN1 exhibited slight expression in brain cell populations, primarily neurons and microglial cells (Figure 4.9b).

In summary, I analyzed gene expression patterns of approved CAR targets and targets under investigation in clinical trials across both healthy and malignant cell populations. The observed differences in target gene expression serve as a valuable resource for retrospective and prospective target evaluation. Additionally, I explored the variations in gene expression among targets in conjunction with clinical data from patients who underwent CAR T cell therapy, revealing limited correlation between global target gene expression patterns and the observed toxicity effects in clinical settings. To further inform the selection of CAR targets prior to clinical translation, I put forth several novel target candidates for CAR T cell therapy against these malignancies, leveraging a single cell gene expression-based approach.

4.4 Discussion

With a rapid increase in clinical trials, CAR T cell therapy has gained increasing interest as a promising approach for the treatment of relapsed or refractory B cell malignancies [37–39]. Despite the remarkable clinical success, a wider application is hindered by often life-threatening adverse effects, such as on-target, off-tumor toxicities [22], which have been shown for a number of CARs [52, 309–313], posing significant challenges to broad clinical application. Therefore, careful consideration of the safety and potential risks associated with target antigen selection is essential before initiating clinical testing.

Transcriptomic reference maps of healthy human tissues and tumor mass can be used for screening therapeutic target expression across cell populations, predicting the potential risk for on-target, off-tumor toxicities and the target specificity on tumor cells for CAR T cell therapy [52, 53, 104, 314–316]. However, with the rapid increase in clinical trials and the consequent expansion of CAR targets, an extensive analysis of the global expression profiles of CAR targets on both healthy and tumor tissues is lacking [212, 277].

To address this issue, I used single cell transcriptomics to scrutinize gene expression patterns of CAR target antigens across both tumor and healthy tissues. We provide a comprehensive overview of current target antigens from clinical trials in malignancies with FDA-approval for CAR T cell therapy. Additionally, we screened clinical trials to generate patient outcome data following administered CAR T cell therapies. I interpreted observed toxicity effects in conjunction with global target expression profiles by generating a transcriptional landscape of more than 300k cells across follicular lymphoma (FL), multiple myeloma (MM), and acute lymphoblastic leukemia (B-ALL), as well over 3 million cells across 35 healthy tissues. Finally, using a single cell gene expression-based approach, I propose novel CAR targets for the treatment of FL, MM and B-ALL that exhibit a favorable gene expression pattern across both healthy and tumor cell populations. The findings serve as a resource for both retrospective and prospective evaluations of therapeutic targets and offer significant support for enhancing the safety and efficacy of CAR T cell therapies.

I am not the first to utilize scRNA Seq for assessing the potential of on-target off-tumor toxicities in CAR T cell therapy. However, previous approaches either did not take antigen expression levels into account, lacked gene expression analysis on tumor data, or were limited to a small number of healthy tissues [212, 277]. This study seeks to build on these previous efforts by providing a more extensive analysis of CAR target antigens in tumor and healthy tissues. However, it is important to acknowledge that this study has certain limitations. Currently, available scRNA Seq data for other malignancies with FDA approval for CAR T cell therapy is lacking, most likely due to induced distortions during the dissociation step in the sampling process [317, 318]. Therefore, it is important to acknowledge the inherent selection bias in data inclusion from

both malignant and healthy tissues used for target antigen screening. Additionally, the 3' bias of the sequencing from the chosen public data may lead to an incomplete characterization of isoforms [136, 319]. For example, I was only able to detect the gene CD44, which shows a high expression pattern across a multitude of cell types, rather than specific CD44 isoforms, such as CD44v6, which was shown to be correlated with tumor subtypes [320]. Recent computational approaches such as *drug2cell* integrate drug target interactions from the ChEMBL database (https://www.ebi.ac.uk/chembl/) to evaluate drug target expression in single cells [321]. Similar approaches could be used to predict interactions between CAR T cells and various cell types within the tumor microenvironment, which could help identify potential off-target effects or unintended interactions that may lead to toxicity. Finally, while scRNA Seq undoubtedly has great potential to benefit CAR T research [127], it is possible that gene expression results may not always correlate with protein expression and in vivo or in vitro results. To overcome these limitations, large-scale protein expression screening data may be necessary, which I expect to be accessible in the upcoming years.

In chapter 3 of this thesis, I introduced an unbiased scRNA Seq-based approach for de novo target identification of potential candidates for CAR T cell therapy [127]. Applying this approach to FL, MM, and B-ALL, I identified several target antigens with strong overexpression on tumor cell populations and minimal expression across vital healthy cell populations. Similarly to previous results [50, 127], my computational approach consistently identified FDA-approved targets as the top hits as well as additional genes that have not been utilized for CAR T cell therapy in the context of these malignancies thus far. The observed gene expression and density levels were relatively low for some of these targets. While a higher level of target expression is known to enhance killing efficacy, additional factors, such as binding properties of the utilized single-chain variable fragment (scFv) and target antigen conformation can influence the interaction between CAR T cells and tumor cells. Nevertheless, it is crucial to bear in mind that scRNA Seq-centered strategies come with their own limitations and in any case require protein validation.

For tumor data, rather than relying solely on specific single cell datasets, creating tumor atlases [322, 323] using comprehensive data sources such as The Cancer Genome Atlas (TCGA; https://www.cancer.gov/tcga) could offer a broader and more detailed tumor landscape across a larger patient population. Similarly, atlases of healthy tissues, such as those from initiatives like the Human Cell Atlas (HCA), can also serve as a valuable reference for mapping and contextualization of tumor datasets. These extensive atlases of both healthy and tumor tissues would enable more data-driven approaches to patient stratification, enhancing our understanding of patient heterogeneity in disease context. Identifying distinct patient subgroups and identifying commonalities between them could help link specific patient group characteristics to reported clinical patient outcome data. One potential method for this is multiple instance learning (MIL), where patient samples are considered as "bags" containing multiple instances (individual cells) along with their associated clinical features (e.g., "Responder" or "Non-responder") [324]. This

approach could uncover more meaningful patterns between single cell data from specific patients subgroups and reported clinical toxicities, enhancing our understanding of treatment responses and patient outcomes.

Still, the overall sparsity of data obtained from clinical trials certainly made it difficult to perform correlation approaches between reported clinical patient data and independent single cell data. Obtaining reliable data from clinical trials proved challenging, as discontinued studies often lacked clear explanations for their termination, and not all studies provided detailed and easily interpretable data regarding adverse effects. For example, it was not always clear whether reported toxicities occurred in the same patients or different ones, making data interpretation challenging. Additionally, accessing actual single cell sequencing data from patients in clinical trials for CAR T cell therapy is currently unfeasible, hindering our ability to understand treatment responses in patients who actually underwent CAR T cell therapy at the single cell level. The lack of a centralized approach for transparently depositing such data further complicates efforts to correlate clinical responses with single cell profiles. Given the anticipated increase in CAR T cell therapy clinical trials, establishing a standardized platform for data deposition would greatly facilitate meaningful correlations between clinical response and single cell profiles.

The lack of a clear pattern between global scRNA Seq profiles of targets and toxicity effects observed in this study is therefore not entirely unexpected, considering also the complex nature of these toxicities and their current limited understanding. Hematotoxicity, neurotoxicity, and cytokine release syndrome associated with CAR T cell therapy are multifactorial in nature, influenced by various factors such as immune system activation, cytokine dysregulation, and interactions with the tumor microenvironment [43, 273]. Additionally, the mere expression of a target antigen in scRNA Seq data does not always indicate the presence of toxicity, as evident from CSF1R, which was strongly expressed in microglia cells but did not result in observed toxicity effects upon CAR development [127]. While not all target expression patterns from my global scRNA Seq analysis might be directly relevant to the toxicity effects observed in clinical studies, the data still remains valuable for clinical research and can aid in refining target selection and understanding the underlying biological processes associated with CAR T cell therapy. Broadly expressed targets may pose a higher risk of off-tumor toxicities due to their potential involvement in vital biological processes across various tissues and should therefore be closely monitored to observe potential side effects during clinical studies. Still, to allow for a more stringent analysis of the correlation between single cell expression patterns and target efficacy and potential for toxicity, a detailed data collection of response patterns for CAR T cells during clinical trials is urgently needed.

Finally, it is important to note that currently, the full extent of potential risks associated with off-tumor activity of CAR T cells is still not well understood [266]. Therefore, this study does

not aim to evaluate the relative merits or shortcomings of any given CAR target antigen, it merely provides a way to identify potential off-tumor activity and guide the selection of CAR antigens before translation into clinics. Looking ahead, continued accumulation of scRNA Seq data from patients, along with increasing availability of clinical toxicity data, will play a crucial role in enhancing our understanding of toxicities associated with CAR T cell therapy. Early identification of potential safety concerns associated with CAR T cell therapy facilitates the careful assessment and monitoring of specific cell populations or therapies to ultimately gain a better understanding of their potential risks and benefits.

5 Timelapsed single cell multiomic cross-species characterization of pluripotent stem cells during neural progenitor differentiation

Pluripotent stem cells (PSCs) can give rise to almost all cell types in the body, including specific cell types that are required to repair damaged or destroyed tissues [58] and have garnered significant attention as a potential remedy for the global shortage of transplantable tissues and organs [64]. Steady advancements in cell culture and differentiation techniques have expanded the scope of PSCs in cellular therapies and their potential impact on human health [25, 65]. However, clinical use of PSCs for cell differentiation or tissue repair is still limited to investigational regenerative medicine, as their translational potential is hindered by certain limitations, such as the slow and inefficient differentiation process, as well as the immature characteristics of derived cell types [59, 63, 65, 71, 74, 75]. Consequently, there is a growing interest in accelerating the differentiation of PSCs to produce functional cell types more rapidly and efficiently [14, 75].

The complex process of PSC differentiation can be better understood by studying mammalian developmental timescales underlying cellular differentiation. Exploring beyond human timescales is essential, as comparative studies across species can provide insights into evolutionarily conserved and species-specific aspects of PSC differentiation. As such, studying developmental timescales across species is critical for realizing the full potential of PSCs in cellular therapies [78]. PSCs during differentiation represent an invaluable experimental tool for elucidating the mechanisms underlying species-specific differentiation timelines.

This chapter therefore describes a comprehensive characterization of PSCs derived from three mammalian species as they underwent neural progenitor cell (NPC) differentiation. Combined timelapsed single cell gene expression and chromatin accessibility profiling facilitated a detailed investigation of species-specific timescales of cellular differentiation.

The goal of this chapter was to investigate species-specific developmental speed underlying NPC differentiation. To ensure accurate cross-species comparisons of mammalian PSCs, we first established uniform culture conditions and differentiation protocols. Subsequently, I combined single cell gene expression and single cell chromatin accessibility to profile over 75,000 differentiating PSCs from human, mouse, and macaca over a ten-day NPC differentiation course. By harnessing the gene expression and chromatin accessibility profiles obtained from the same individual cells, I identified both conserved and species-specific transcriptomic and epigenomic signatures associated with NPC differentiation. Differences in the distribution of cells across various cell cycle phases during NPC differentiation and variations in transcript abundance and cell size among species suggest that species-specific cell physiology may set the speed at which differentiation occurs. An integrative analysis of gene scores and transcription factor motif accessibility data unraveled correlated regulatory dynamics and key drivers of NPC differentiation. Furthermore, by mapping cells from different species onto a shared embedding space, I could

discern distinct patterns of progression over the course of differentiation. These findings shed light on the complex molecular machinery governing PSC differentiation into NPC lineages and provide insights into species-specific characteristics that influence the dynamics of cellular differentiation.

This chapter is based on and partly identical to the following publication:

1. **Thomas, M.***, de la Porte, A.*, Schröder, J.*, ... Drukker, M., Schröter, C. and Marr, C., 2023. Single-cell multiomic comparison of differentiating pluripotent stem cells from three mammalian species. *In preparation.*

Note that "*" marks an equal contribution. See section 1.5 of this thesis for a detailed summary of individual contributions.

5.1 Differentiating pluripotent stem cells from three mammalian species under identical conditions

Mammalian embryonic development follows a coordinated sequence of events to ensure proper tissue and organ formation. The duration of these events, however, varies widely among different species. For example, while it takes 13 days for human embryos to progress from oocyte fertilization to gastrulation, it only takes six days for mice to reach this stage [79]. Pluripotent stem cells (PSCs), retain the ability to differentiate into specialized cell lineages representing different stages of development [14]. Consequently, they can serve as an invaluable tool for investigating the regulation of developmental timing across species.

In recent years, the advent of single cell multiomic sequencing technologies has revolutionized our ability to unravel the complexities of cellular diversity and dynamics [116]. Multiomic sequencing goes beyond traditional single cell transcriptomic methods by combining gene expression and chromatin accessibility profiling, thus providing a more holistic understanding of the intricate molecular mechanisms governing PSC differentiation. Capturing gene expression and chromatin accessibility profiles of individual cells at different stages during development enables the dissection of cellular heterogeneity and the reconstruction of lineage trajectories that follow cellular development, thereby enhancing our understanding of developmental processes across multiple species [183].

The utilization of differentiating PSCs from multiple species is crucial for transitioning from basic research to potential clinical applications. By incorporating mouse and non-model organisms such as macaca, which closely resemble human biology, we can effectively bridge the gap between fundamental discoveries and therapeutic interventions. However, conducting comparative studies across species can be challenging due to the inherent differences in culture conditions and growth requirements of PSCs from various species. Typically, cells are cultured in species-specific media formulations tailored to meet their specific needs. This divergence in culture conditions can introduce confounding factors and hinder direct comparisons between differentiating PSCs of multiple species. To mitigate potential biases associated with distinct culture media and focus solely on the inherent characteristics of PSCs and their differentiation capabilities, our initial aim was to establish standardized cell culture conditions for epiblast-like primed-state human embryonic stem cells (ESC), mouse epiblast stem cells (EpiSC), and macaca induced PSCs (iPSCs).

The experimental aspects of harmonizing culture conditions, quantitative reverse transcription PCR (RT-PCR), immunostainings, and neural progenitor cell (NPC) differentiation were performed by Alexandra de la Porte and Julia Schröder under supervision of Prof. Micha Drukker and Dr. Christian Schröter (see section 1.5 of this thesis for a detailed summary of individual contributions), and will be briefly outlined (Figure 5.1).

Following the evaluation of various cell culture media, we identified Universal Primate PSC Media (UPPS) [325] as the most promising medium. Cells were then gradually adapted to the new conditions (Figure 5.1a). Additionally, we established a common protocol for NPC differentiation through dual SMAD inhibition [326] to explore the distinct developmental timing of each species (Figure 5.1b). Using Matrigel coating, clump passaging with EDTA and UPPS media, we were able to successfully cultivate all cell lines and enhance colony morphology (Figure 5.1c). Daily RT-qPCR samples over a ten-day course of NPC differentiation confirmed a prompt downregulation of pluripotency factor POU5F1 (also known as OCT4) upon NPC differentiation, while neural markers SOX1 and PAX6 were gradually upregulated (Figure 5.1d). Similarly, via immunofluorescence, SOX1 and PAX6 proteins were detectable in human and mouse cells following dual SMAD inhibition, whereas POU5F1 was strongly downregulated early into the differentiation protocol (Figure 5.1e).

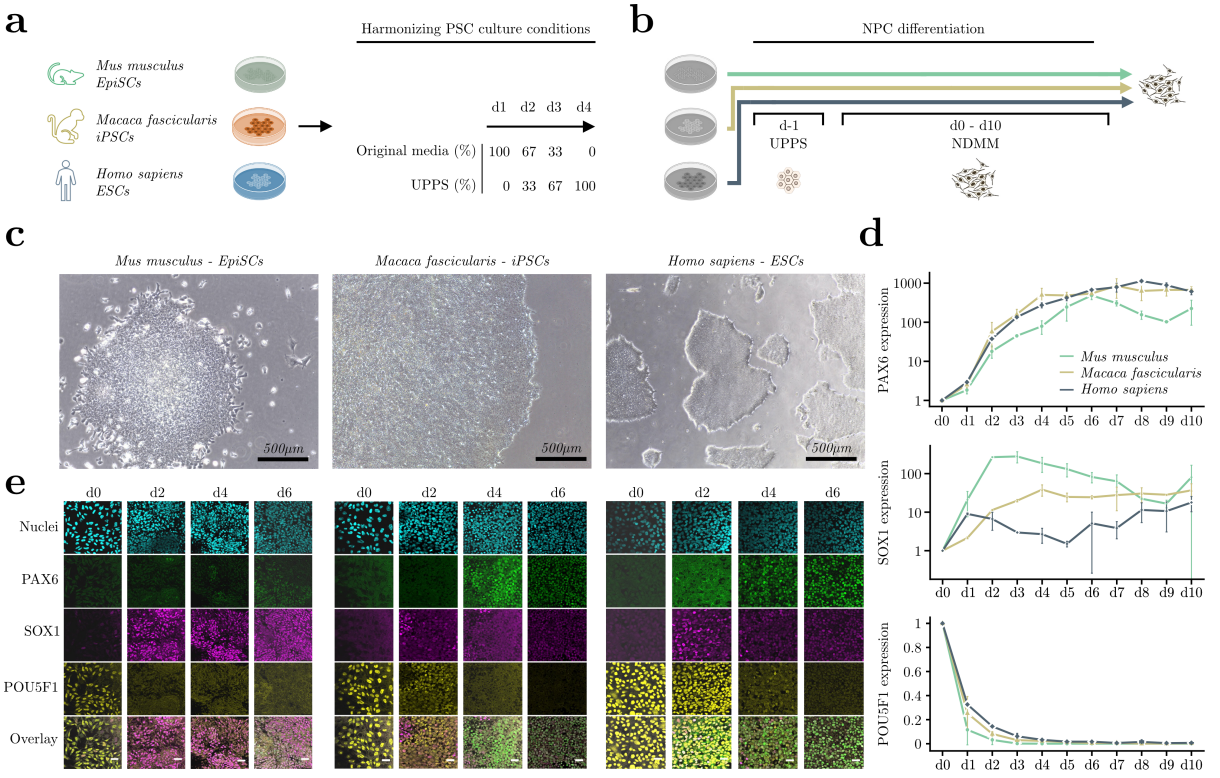

Figure 5.1: Primed PSCs of three mammalian species are maintained and differentiated under identical conditions. a) Establishing standardized cell culture conditions for mouse, macaca, and human PSCs through gradual adaptation to Universal Primate PSC Media (UPPS) [325]. **b)** Mouse, macaca and human PSCs were subject to neural progenitor cell (NPC) differentiation. **c)** Colony morphology from cultivated mouse, macaca and human cell lines. **d)** Daily RT-qPCR samples for PAX6, SOX1 and POU5F1 (OCT4) over a ten-day course of NPC differentiation for mouse, macaca and human cells. **e)** Immunofluorescence measurements of single nuclei, PAX6, SOX1 and POU5F1 (OCT4) expression for mouse, macaca and human cells.

In summary, by employing a common culture medium, we established a platform that allows for direct and unbiased comparisons of differentiating PSCs across three mammalian species. Successful NPC differentiation, as evident from RT-qPCR and immunofluorescence analysis, demonstrates the feasibility of utilizing a standardized differentiation protocol.

5.2 Gene expression and chromatin accessibility dynamics follow species-specific timescales

The subsequent sections detail my computational analysis of differentiating mouse, macaca and human PSCs utilizing combined timelapsed gene expression (scRNA Seq) and chromatin accessibility (scATAC) sequencing at eight distinct timepoints.

5.2.1 Gene expression reflects species-wide differences during differentiation

To mitigate technical cross-species batch effects, we pooled an equal number of cells from the three species for each of the eight timepoints (Figure 5.2). Consequently, I aligned resulting sequencing data against each of the three respective reference genomes separately. To assign each barcode to a species of origin, I employed a two-step approach: initially, I prematurely assigned cells to a species based on which reference genome yielded the highest counts per cell. Subsequently, I used *souporcell* [131] to identify and remove doublet cells, and cluster cells on their genotype and their respective species of origin based on single nucleotide polymorphisms. For each sample, I filtered barcodes to retain high quality cells based on the total distributions of unique molecular identifier counts and genes and the fraction of mitochondria-encoded genes. See Table 9 for applied filtering thresholds. I excluded genes detected in fewer than 20 cells from further analyses.

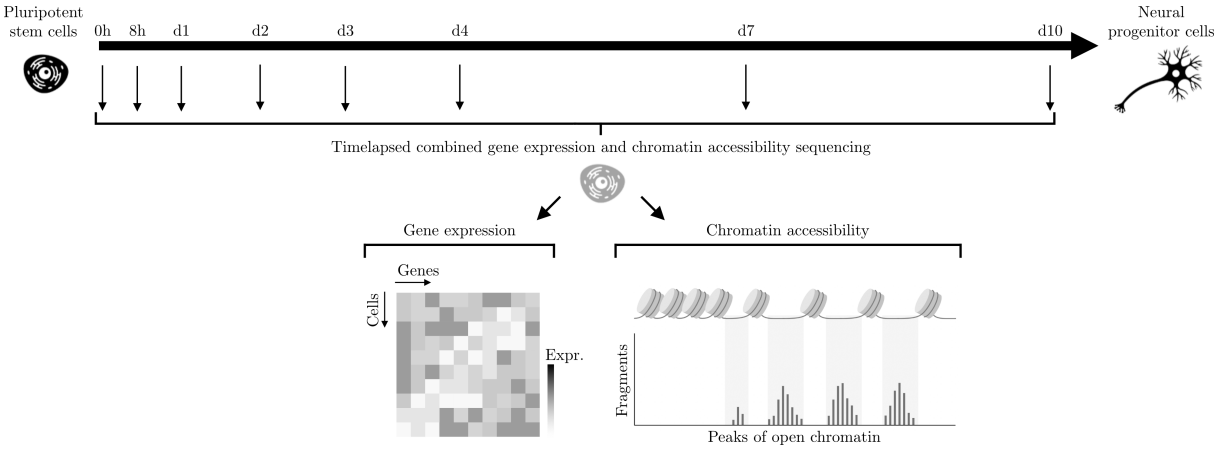

Figure 5.2: Timelapsed profiling of differentiating cells. PSCs of mouse, macaca and human were sampled at 0h, 8h, d1, d2, d3, d4, d7 and d10 during NPC differentiation and profiled using combined gene expression and chromatin accessibility sequencing.

I normalized unique molecular identifier counts of each cell using the *scran* algorithm [148]. I identified the top 4,000 variable genes based on normalized dispersion [136]. I performed principal-component analysis dimension reduction by computing 15 principal components on highly variable genes. Next, I computed a neighborhood graph on the first 50 principal components with 15 neighbors. To identify genes involved in the linear NPC differentiation process, I used *Waddington Optimal Transport (WOT)* [188] to identify potential driver

genes correlating with fate probabilities towards the terminal macrostate as implemented in *cellrank* [187]. *WOT*, primarily used to infer developmental trajectories by minimizing transportation costs between cell states for timelapsed scRNA Seq data (see section 2.1.9), was employed here to identify genes that follow the linear trajectory towards endpoint cells at the final timepoint. For two-dimensional visualization, I embedded the neighborhood graph via uniform manifold approximation and projection (UMAP) [161] on these lineage driver genes identified from *WOT* with an effective minimum distance between embedded points of 0.5 (see methods section 2.1 of this thesis for a detailed explanation of applied preprocessing steps).

Table 9: Cell filtering thresholds during scRNA Seq data preprocessing of differentiating mouse, macaca and human cells. Hyphen: Threshold was not applied. C_{min}/C_{max}= lower and upper thresholds for filtering cells according to captured UMI counts. C_{min}/C_{max}, G_{min}/G_{max} = upper thresholds for filtering cells according to captured UMI counts (C_{min}/C_{max}) or genes (G_{min}/G_{max}). MT_{max} = upper thresholds for filtering cells according to their fraction of mitochondria-derived genes. See methods section 2.1.1 for a definition and detailed explanation about filtering thresholds.

Species	**Timepoint**	$\mathbf{C_{min}}$	$\mathbf{C_{max}}$	$\mathbf{G_{min}}$	$\mathbf{G_{max}}$	$\mathbf{MT_{max}}$	$\mathbf{D_{max}}$
Mouse	0h	1,000	10,000	700	5,000	30%	20%
Mouse	8h	1,250	15,000	1,000	5,000	30%	20%
Mouse	d1	1,000	8,000	500	3,500	30%	20%
Mouse	d2	2,000	10,000	1,000	5,000	30%	30%
Mouse	d3	1,500	10,000	1,000	4,000	30%	30%
Mouse	d4	1,500	10,000	1,000	3,600	30%	30%
Mouse	d7	1,000	10,000	1,000	4,000	30%	30%
Mouse	d10	800	10,000	300	4,000	30%	30%
Macaca	0h	1,500	20,000	1,000	6,000	-	20%
Macaca	8h	1,500	20,000	1,000	6,000	-	20%
Macaca	d1	1,000	30,000	1,000	7,000	-	20%
Macaca	d2	2,500	40,000	1,500	8,000	-	20%
Macaca	d3	1,500	30,000	1,000	7,000	-	20%
Macaca	d4	1,000	30,000	1,000	7,000	-	20%
Macaca	d7	1,000	40,000	1,000	8,000	-	20%
Macaca	d10	1,000	20,000	800	6,000	-	20%
Human	0h	2,000	30,000	1,000	7,000	30%	20%
Human	8h	2,000	40,000	1,500	7,000	30%	20%
Human	d1	2,000	30,000	1,500	7,500	30%	20%
Human	d2	3,000	75,000	2,000	9,000	30%	20%
Human	d3	2,500	40,000	2,000	7,500	30%	20%
Human	d4	3,000	50,000	2,000	8,000	30%	20%
Human	d7	3,500	40,000	1,500	7,000	30%	20%
Human	d10	2,000	30,000	1,000	7,000	30%	20%

UMAP plots of gene expression profiles obtained from differentiating cells accurately reflected the linear trajectory of NPC differentiation for each species (Figure 5.3a) and marker gene expression followed the expected dynamics (Figure 5.3b). For instance, the pluripotency marker POU5F1 exhibited a gradual downregulation during NPC differentiation, while neural markers such as SOX2, PAX6 and MAP2 displayed a progressive upregulation.

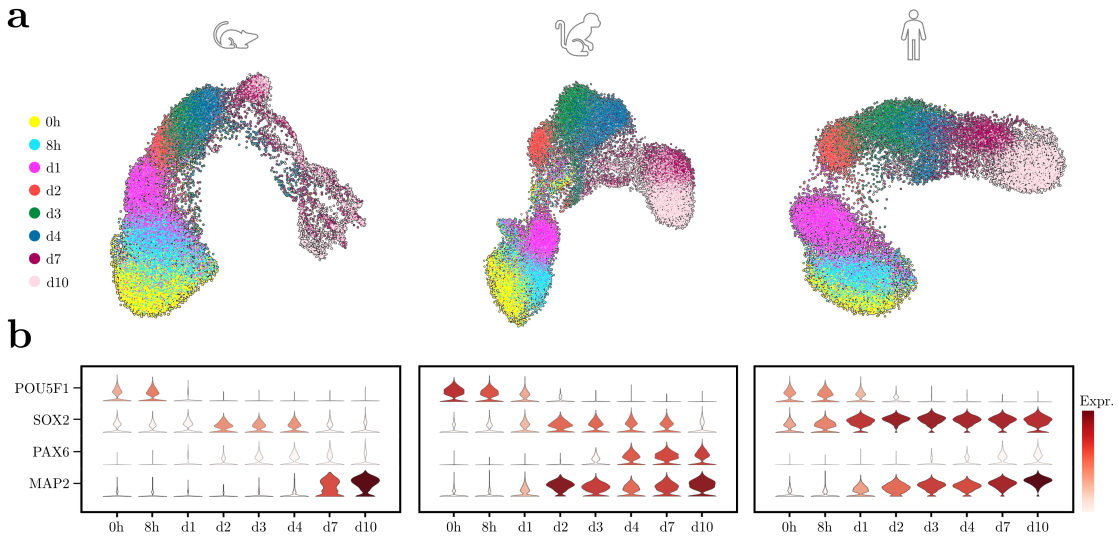

Figure 5.3: Single cell gene expression profiling captures NPC differentiation processes for three mammalian species. a) Gene expression UMAPs showing PSCs of mouse (left), macaca (center) and human (right) sampled at eight time points during NPC differentiation. Colors indicate sampling timepoints. **b)** Expression of selected pluripotency and neuronal marker genes for mouse (left), macaca (center) and human (right) cells. Normalized expression values were logarithmized and scaled to unit variance.

To gain further insights into the cellular states during NPC differentiation, I employed unsupervised clustering and marker genes to analyze the composition of cells throughout the differentiation process (Figure 5.4).

For mouse cells, I observed broadest spectrum of differentiation states, ranging from pluripotent cells to neural cells to mature neuronal cells (Figure 5.4a-d). Macaca (Figure 5.4e-g) and human (Figure 5.4h-j) cells only reached the neural stage within the 10-day differentiation timeframe. Notably, mouse cells exhibited a relatively faster progression, reaching the neural stage around day 2 (Figure 5.4d), while macaca and human cells required three and four days, respectively, to reach the neural stage (Figure 5.4g,j).

In summary, gene expression profiles of single cells from the three mammalian species effectively captured the process of NPC differentiation and revealed differences in the overall speed of differentiation among the species.

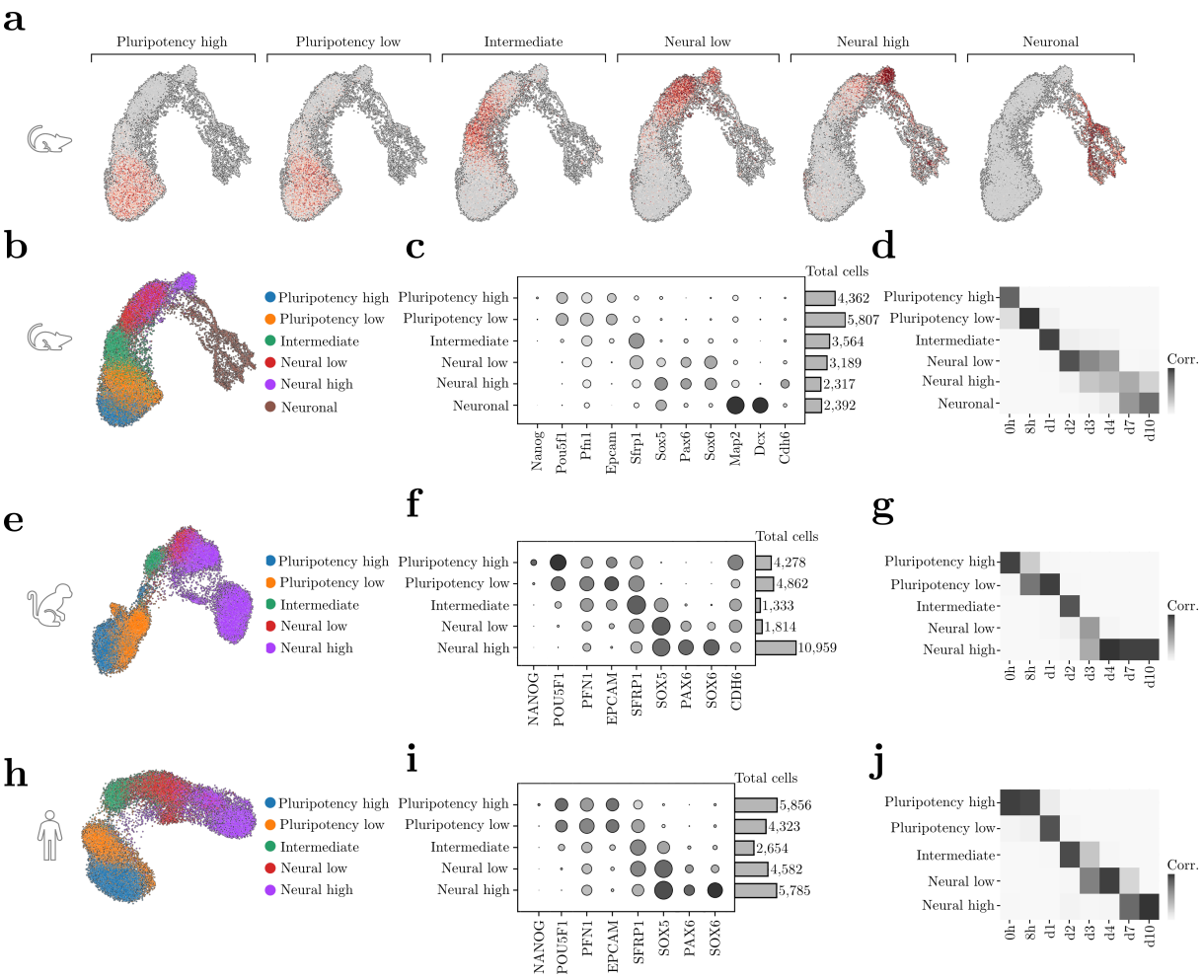

Figure 5.4: Species-specific variations in cellular stages during NPC differentiation. **a)** UMAP showing mouse gene scores for markers used to characterize six cellular stages during differentiation. Scores were calculated by averaging the expression of a set of marker genes subtracted with the average expression of a randomly sampled reference set of genes. **b,e,h)** UMAP projection of differentiating mouse **(b)**, macaca **(e)**, and human **(h)** cells along with their respective annotation of cellular stages. **c,f,i)** Expression of marker genes used to characterize cellular stages during differentiation for mouse **(c)**, macaca **(f)**, and human **(i)** cells. Dot size indicates the fraction of cells per cell cluster expressing a gene, color intensity shows mean normalized gene expression per cell cluster. Bars indicate total cells per stage. **d,g,j)** Correlation between annotated cell clusters and sampling time for mouse **(d)**, macaca **(g)**, and human **(j)** cells.

5.2.2 Species-specific chromatin accessibility profiles during differentiation

After confirming that gene expression profiles accurately capture global cellular changes in cellular states during differentiation, I sought to determine whether the chromatin accessibility of individual differentiating cells exhibits similar patterns.

First, I assigned barcodes to the respective species using the labeled cell barcodes from the scRNA Seq analysis. I then inspected the resulting peak and tile matrices based on the number

of fragments generated from Tn5 enzyme transposition events, transcription start site (TSS) enrichment score and nucleosome signal per cell to obtain high quality cells. I excluded cells with captured fragments F_{min} <1,000 or TSS enrichment scores TSS_{min} <1 from further analysis (see methods section 2.2.2 for a definition and detailed explanation about applied filtering thresholds). Additionally, I excluded peaks located on non-standard chromosomes or chromosome scaffolds, as well as peaks within genomic blacklist regions from further analysis. I then performed layered dimensionality reduction using latent semantic indexing (LSI), consisting of normalization via term frequency-inverse document frequency (TF-IDF) and dimension reduction via singular value decomposition (SVD). Finally, I calculated a UMAP [161] embedding based on the LSI reduced dimensions with 30 neighbors and an effective minimum distance between embedded points of 0.5 [194]. Finally, I transferred cell annotations from scRNA Seq cells (see methods section 2.2 of this thesis for a detailed explanation of applied preprocessing steps).

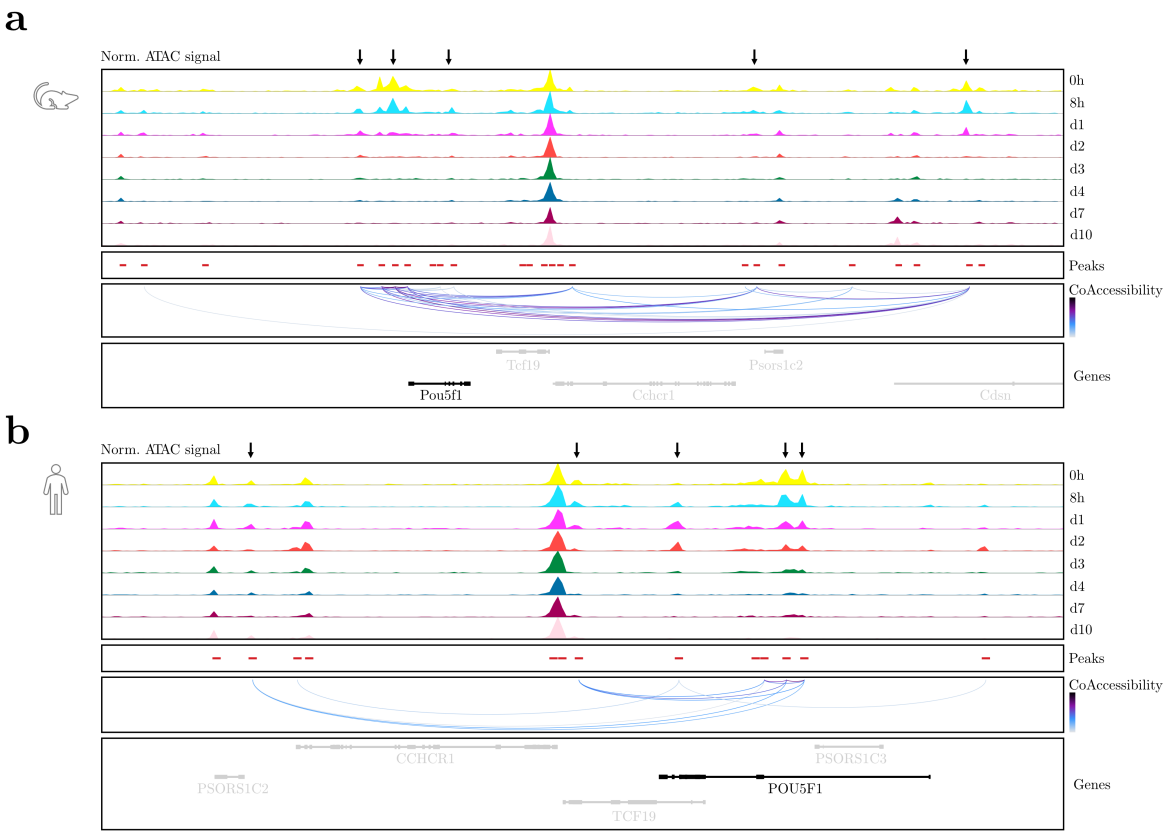

Figure 5.5: Species-specific differences in Tn5 insertions across sampling timepoints. **a-b)** Local chromatin accessibility around the representative region of the POU5F1 gene for mouse **(a)** and human **(b)** cells. Frequency of Tn5 integration is averaged for cells from each timepoint and displayed as pseudo-bulk accessibility tracks to visualize the DNA accessibility in a region. Identified peaks, co-accessible peaks that are potentially correlated to the expression of nearby genes and the location of the genes are visualized alongside these accessibility tracks.

The Tn5 transposase enzyme cuts accessible regions of chromatin DNA, resulting in fragments of open chromatin DNA. I illustrated the frequency of Tn5 insertions across genomic regions (Figure 5.5). These genomic peaks represent pseudo-bulk genomic accessibility tracks, where the Tn5 signal from all cells at each timepoint was averaged to visualize the local DNA accessibility.

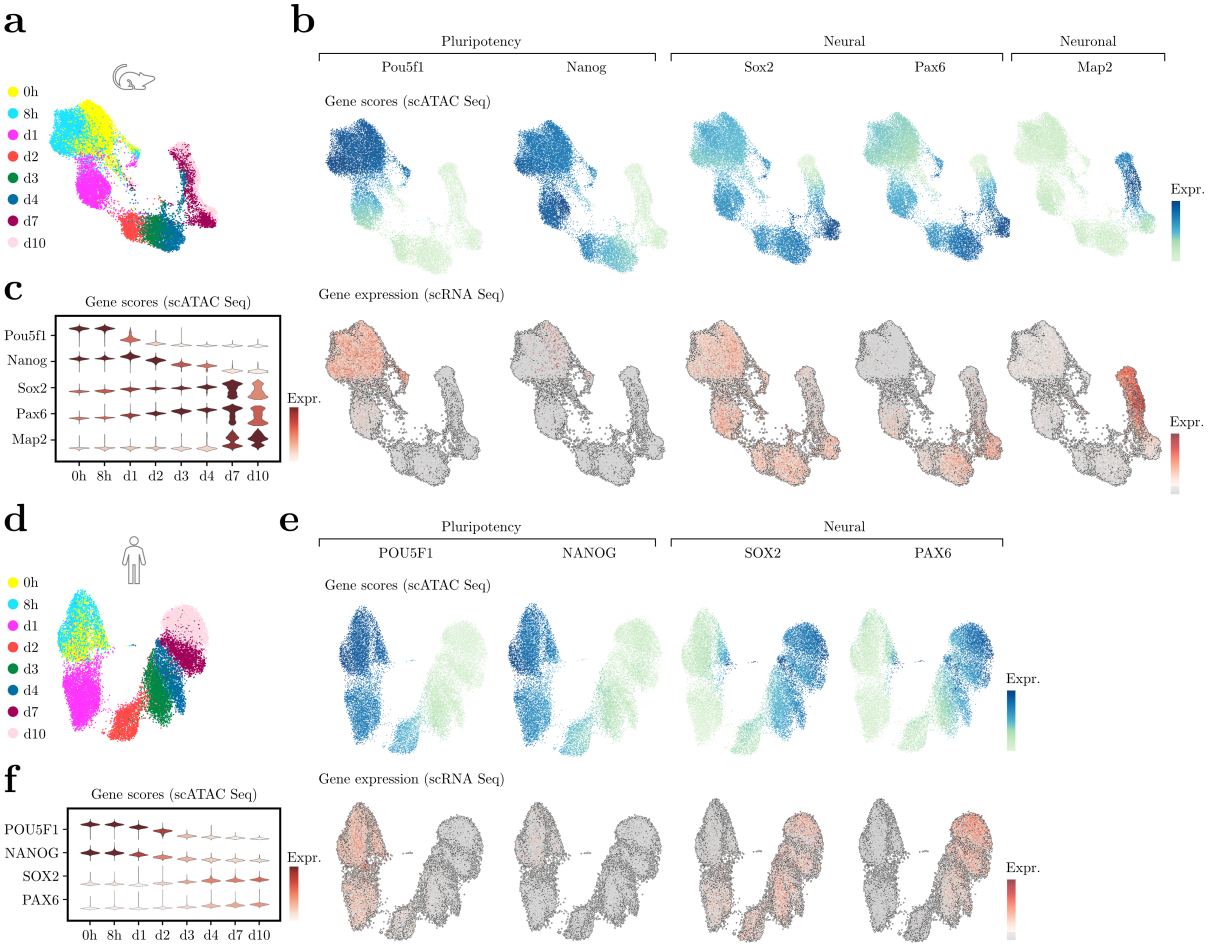

Figure 5.6: Estimated gene expression from chromatin accessibility aligns with scRNA Seq profiles. a,d) Chromatin accessibility UMAPs showing PSCs of mouse **(a)** and human **(d)** sampled at eight timepoints during NPC differentiation. Colors indicate sampling timepoints. **b,e)** Estimated gene expression scores from scATAC Seq data (top) along with gene expression data from scRNA Seq (bottom) for selected pluripotency, neural and neuronal markers from mouse **(b)** and human **(e)** cells. **c,f)** Expression of selected pluripotency and neuronal marker genes for mouse **(c)**, and human **(f)** cells using estimate gene expression scores. Gene scores were imputed using *MAGIC* [201] for visualization purposes.

In the context of mouse cells, I observed distinct peaks in the chromatin accessibility profiles surrounding the region of pluripotency gene Pou5f1, which exhibited a rapid decline over time during differentiation (Figure 5.5a). Similarly, in human cells, chromatin accessibility around the POU5F1 gene demonstrated distinct peaks at earlier timepoints, gradually diminishing as differentiation progressed (Figure 5.5b). Notably, in mouse cells, the presence of peaks around

the Pou5f1 gene persisted until 8 hours into NPC differentiation, while human DNA surrounding the POU5F1 gene remained accessible until 2 days of NPC differentiation.

Notably, UMAP plots of chromatin accessibility profiles obtained from differentiating cells accurately reflected the linear trajectory of NPC differentiation for mouse and human (Figure 5.6a,d). To facilitate the visualization and interpretation of scATAC Seq data, I leveraged chromatin accessibility patterns to estimate gene expression profiles for cell state-specific marker genes. I calculated gene scores that estimate the level of gene expression based on the local accessibility of the gene region, including the promoter and gene body, across all cells in the data, adjusting for gene distances and large differences in gene size using *ArchR* [194] (Figure 5.6b-c,e-f).

Remarkably, estimated gene expression profiles for mouse (Figure 5.6b,c) and human (Figure 5.6e,f) cells closely resembled the actual gene expression profiles obtained from the scRNA Seq analysis (Figure 5.3b). The gradual downregulation of pluripotency genes such as POU5F1 and NANOG, along with the upregulation of neural marker genes such as SOX2 and PAX6, were evident during the course of differentiation (Figure 5.6b-c,e-f). Notably, the mouse neuronal marker Map2 exhibited expression primarily in the last two timepoints (day 7 and day 10), consistent with the findings from scRNA Seq analysis (Figure 5.6b,c).

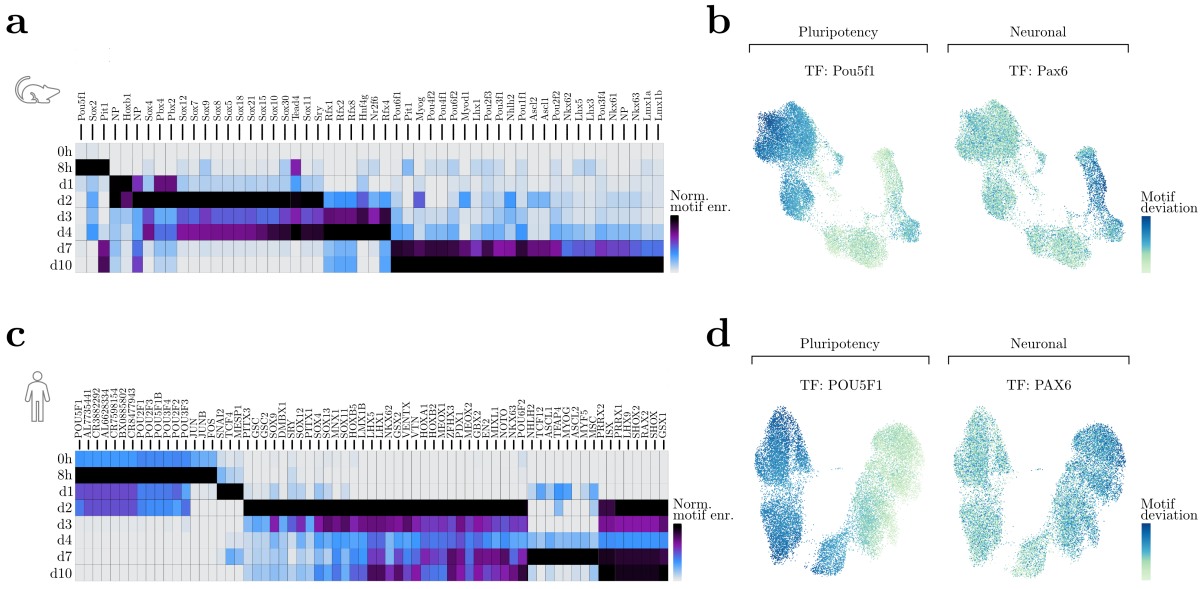

Figure 5.7: Transcription factor (TF) activity follows specified-specific dynamics. **a,c)** Heatmap of TF motif hypergeometric enrichment-adjusted p values within the marker peaks of each sampling timepoint for mouse **(a)** and human **(c)** cells. **b,d)** UMAP projection of TF motif deviation scores for pluripotency TF POU5F1 (left) and neuronal TF PAX6 (right) for mouse **(b)** and human **(d)** cells.

To explore variations in accessible genomic regions, I identified characteristic peaks of open chromatin of cells from each timepoint representing variations in accessible genomic regions. I

then scrutinized these sets of peaks for enriched DNA binding motifs of transcription factors (TFs) (Figure 5.7a,c). In mouse cells, motifs for Pou5f1 TFs were enriched during the initial 8 hours of differentiation (Figure 5.7a), while human cells exhibited enriched TF motifs for POU5F1 until 2 days into the differentiation protocol (Figure 5.7c). Consistent with the scRNA Seq results, Sox2/SOX2 motifs were enriched from around day 1 for mouse cells and from around day 2 for human cells during the differentiation timecourse (Figure 5.7a,c).

To ease visualization and interpretation of TF activity, I analyzed TF activity per cell by calculating the deviation of per-cell accessibility of a given motif from the expected accessibility based on the average of all cells using *chromVar* [206] (Figure 5.7b,d). Similarly to TF motif enrichment, the TF deviation score for pluripotency marker Pou5f1/POU5F1 was highest in cells during early NPC differentiation and ended around day 1 for mouse cells (Figure 5.7b), while in human cells, it decreased around day 2 (Figure 5.7d). Estimated TF activity of Pax6/PAX6 gradually increased during differentiation for both mouse and human cells (Figure 5.7b,d).

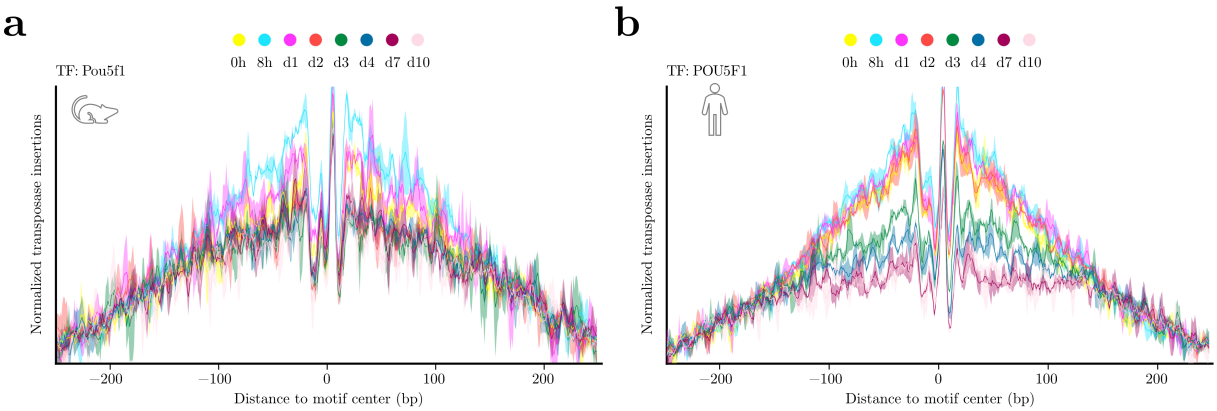

Figure 5.8: Tn5 bias-adjusted TF footprints for POU5F1. Lines are colored according to their sampling time for mouse **(a)** and human **(b)** cells.

Considering that the mere presence of a DNA binding motif does not necessarily imply actual binding by the corresponding TF, I performed TF footprinting analysis to identify specific TFs that were physically bound to particular genomic regions (Figure 5.8). Footprinting analysis confirmed the highest TF binding of POU5F1 during the first day in mouse cells (Figure 5.8a) and the first 2 days in human cells (Figure 5.8b).

In summary, both scRNA Seq and scATAC Seq data from differentiating PSCs of multiple species accurately reflected the NPC differentiation process. This descriptive analysis at a gene, chromatin, and TF level revealed a consistent alignment between gene expression and chromatin accessibility profiles, providing a more holistic understanding of cellular dynamics during NPC differentiation.

5.3 Unraveling pluripotent stem cell differentiation dynamics through single cell gene expression and chromatin accessibility

The cell cycle describes a series of events that exerts regulatory control over pivotal aspects of cellular dynamics, including cell growth, DNA replication, and cell division. Throughout the entirety of an organism's life cycle, cells undergo a perpetual progression through discrete and discernible phases of this cell cycle continuum. The G2/M phase marks the transition from the second gap phase (G2), where the cell prepares for cell division to mitosis (M), where nuclear division and segregation of genetic material occurs. Following the completion of mitosis, cells enter the G1 phase, the first gap phase. G1 serves as a critical decision-making point, where cells assess environmental cues and internal signals to determine whether to proceed with cell division or enter a non-dividing state. During the S phase, DNA synthesis occurs, resulting in the formation of two identical copies [92].

5.3.1 The impact of cell physiology on differentiation speed

By examining marker genes specific to each cell cycle phase, I calculated a cell cycle score for individual cells and assigned them to a particular phase (Figure 5.9a). Comparing the cell cycle phase compositions between species, I observed a significant increase in the G1 fraction of mouse cells over time during NPC differentiation, with 73% of cells at day 7 and 88% of cells at day 10 being in the G1 phase (Figure 5.9b). Additionally, mouse cells consistently exhibited the lowest number of captured genes across all three cell cycle phases when compared to the other species (Figure 5.9c).

To investigate this further, fluorescent ubiquitination-based cell cycle indicator (Fucci) cell lines were generated, enabling the determination of the cell cycle phase over time (Figure 5.9d). Absolute measurements of cell cycle length validated that mouse cells have a 1.4-fold shorter cell cycle duration compared to human cells (Figure 5.9e). Notably, in line with my measurements of captured genes, cell volume measurements using a Coulter Counter revealed that mouse cells exhibited the smallest overall cell size (Figure 5.9f). Note that the generation of Fucci cell lines and the experimental measurements of cell cycle length and cell volume (Figure 5.9d-f) was performed by Julia Schröder under supervision of Dr. Christian Schröter (see section 1.5 of this thesis for a detailed summary of individual contributions).

In summary, variations in cell cycle dynamics and cell size across species suggest that species-specific cell physiology might be an important factor in setting differentiation speed.

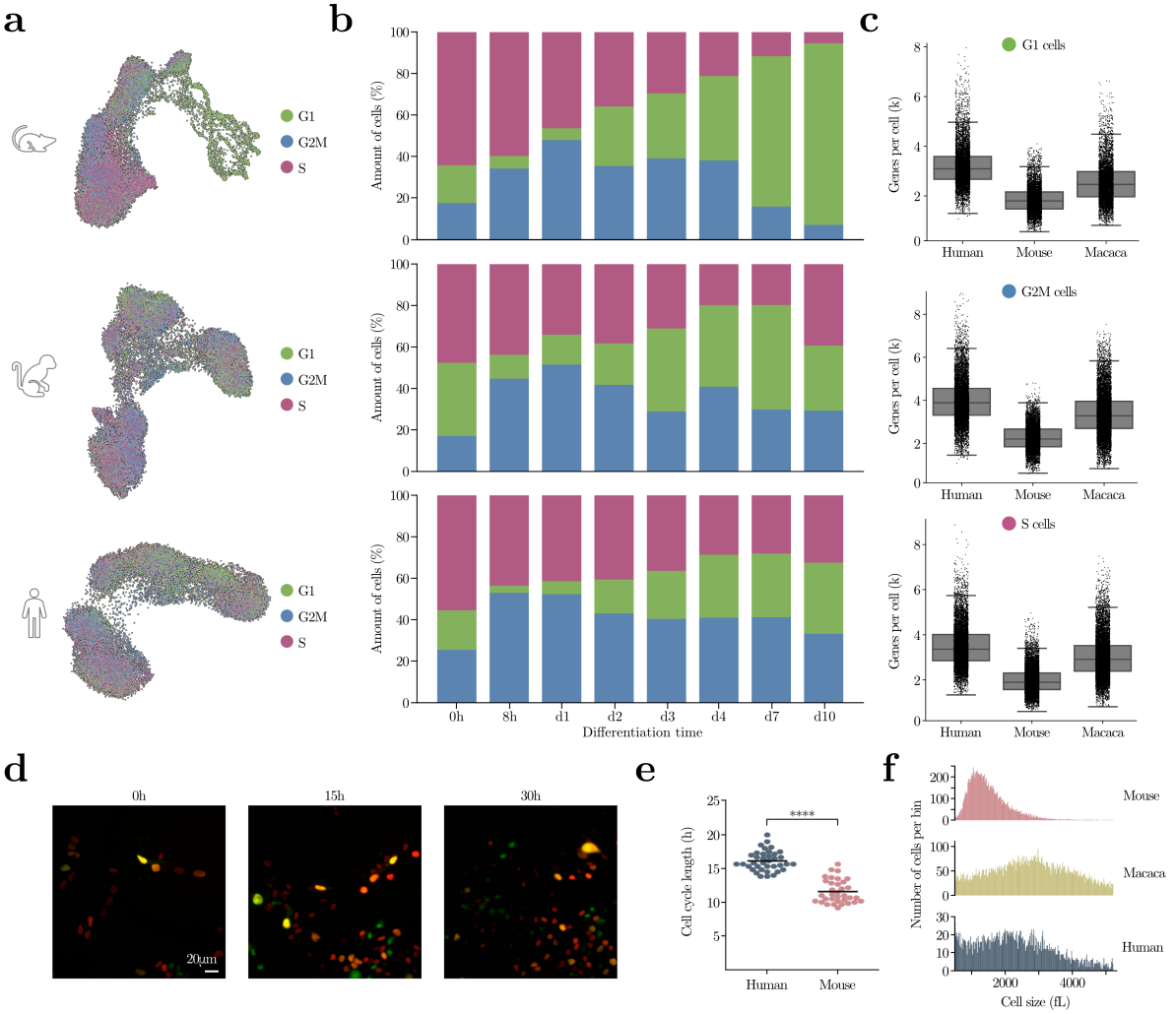

Figure 5.9: Differences among species in cell cycle phase composition and captured genes. a) UMAP projection of mouse (top), macaca (center), and human (bottom) cells according to their assigned cell cycle phase. **b)** Proportions of G1, G2M, and S cells over time for mouse (top), macaca (center), and human (bottom) cells. **c)** Amount of captured genes between species for G1 (top), G2M (center), and S (bottom) cells. **d)** Generation of fluorescent ubiquitination-based cell cycle indicator (Fucci) cell lines to determine the cell cycle phase of single cells over differentiation time. **e)** Absolute cell cycle length between human (16h) and mouse (11.5h) cells from Fucci cell tracking differs significantly (p value <0.01 using a t-test). **f)** Cell volume measurements for mouse (top), macaca (center), and human (bottom) cells.

5.3.2 Cross-species mapping unveils distinct global differentiation rates

To investigate the overall speed of differentiation in PSCs across the three species, I mapped cells from human, mouse and macaca into one common embedding (Figure 5.10). I performed integration of embeddings and annotations of macaca and human cells using Scanpy's *ingest* function, with mouse cells serving as the reference since they exhibited the fastest differentiation speed and therefore covered the most stages during the differentiation process.

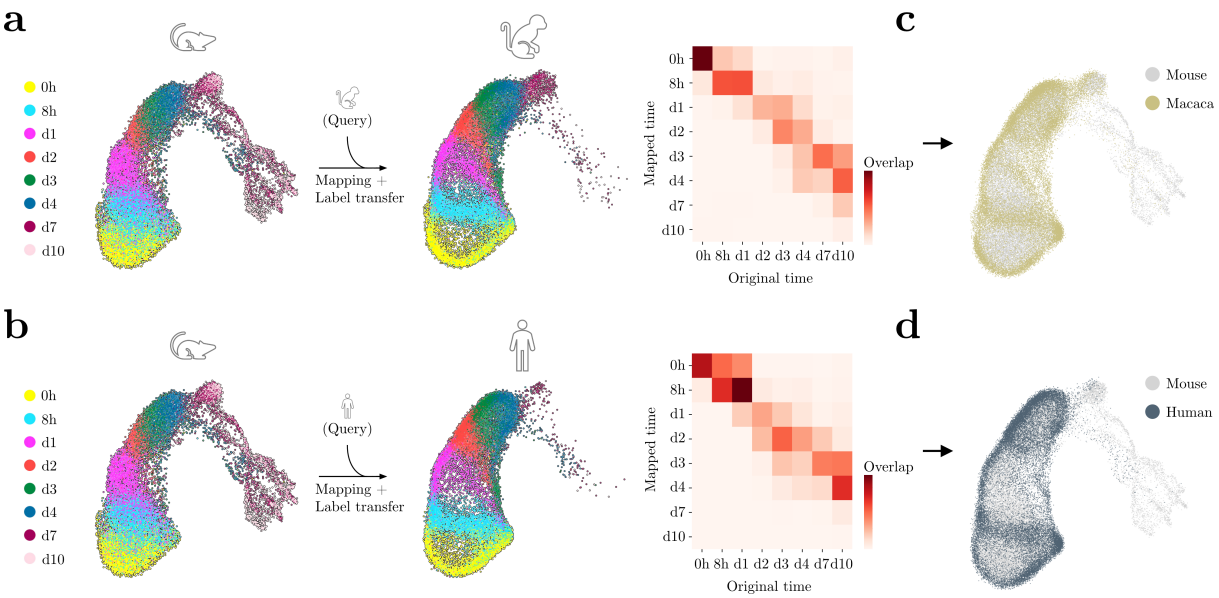

Figure 5.10: Projection of macaca and human cells to reference embedding of mouse cells. a-b) Macaca **(a)** and human **(b)** cells were projected onto mouse UMAP embedding, serving as a reference (left). Correlation between original sampling time and novel mapped time annotation (right). **c-d)** Common UMAP embedding of mouse reference and macaca **(c)** or human **(d)** cells.

It is noteworthy that instead of learning a joint representation of query and reference data, I projected query data onto the reference embedding, allowing us to analyze how cells from the slower species (macaca and human) correspond to cells from the fast mouse reference, thus characterizing differences in global differentiation speed between species.

Upon mapping, I observed that macaca cells from the 10-day differentiation course only corresponded to mouse cells until around day 4-7 (Figure 5.10a, left), leading to a global shift between the original time and the inferred mapped time (Figure 5.10a, right). Similarly, for human cells, this effect was even more pronounced, as human cells from the 10-day differentiation course correspond to mouse cells until around day 4 (Figure 5.10b). To facilitate visual interpretation, I projected cells from all three species into a common embedding, where I observed that macaca cells progressed slightly further along the differentiation time course compared to human cells, indicating a slightly higher differentiation speed in macaca cells compared to human cells (Figure 5.10c-d). Finally, I compared growth rates between species, demonstrating that mouse differentiation occurs at a pace 2.2 times faster than that of macaca, and 2.4 times faster than human differentiation.

Overall, cross-species gene expression mapping provided insights into the global differences in differentiation speed among the species, highlighting the unique progression and correspondence of cells from macaca and human species to cells from the faster mouse species.

5.3.3 Identifying temporal drivers of NPC differentiation

To analyze the temporal dynamics of gene expression and chromatin accessibility during NPC differentiation and uncover regulatory mechanisms involved, I utilized trajectory inference methods to find a gradual order of estimated gene expression and chromatin accessibility profiles from scATAC Seq data. This allowed us to establish a pseudotime that corresponds to the linear NPC differentiation process (Figure 5.11).

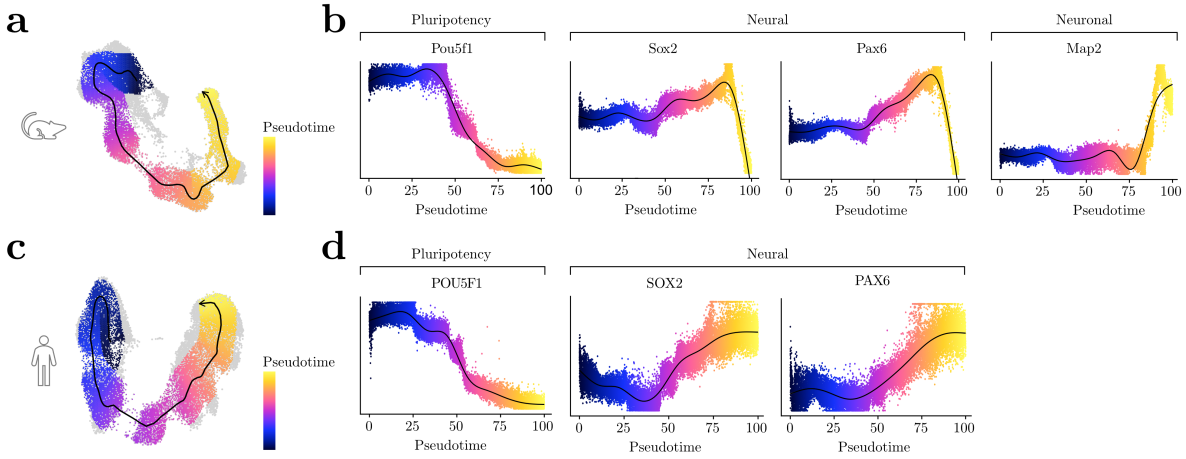

Figure 5.11: Inferred cellular trajectory approximates NPC differentiation. **a,c)** UMAP projection of inferred pseudotime for mouse **(a)** and human **(c)** cells. **b,d)** Gene scores of pluripotency, neural and neuronal markers against inferred pseudotime values for mouse **(b)** and human **(d)** cells. Cells were colored by their pseudotime.

In the case of mouse (Figure 5.11a-b) and human (Figure 5.11c-d) cells, the inferred pseudotime aligned well with the expected pattern of NPC differentiation. This alignment was evident from the gradual up- or downregulation of pluripotency, neural, and neuronal markers along the inferred pseudotime (Figure 5.11b,d).

Moving forward, I performed an integrative analysis across the cellular trajectory by integrating estimated gene expression (gene scores) and TF motif accessibility based on *chromVar* [206] from scATAC Seq data throughout the inferred pseudotime, revealing correlated regulatory dynamics and potential drivers of NPC differentiation (Figure 5.12).

Through this integrated analysis, I identified 50 (mouse) and 34 (human) TFs where gene expression is positively correlated to changes in the accessibility of their corresponding motif, thus exhibiting similar dynamics between gene expression and TF motif accessibility. Among these, eight transcription factors were conserved between species: POU5F1, GSC, NFE2, ZIC3, KLF11, NFYA, MEIS1 and RFX4. As expected, POU5F1 (Figure 5.12) appeared prominently in both mouse and human cells over time, indicating its involvement in regulating NPC differentiation.

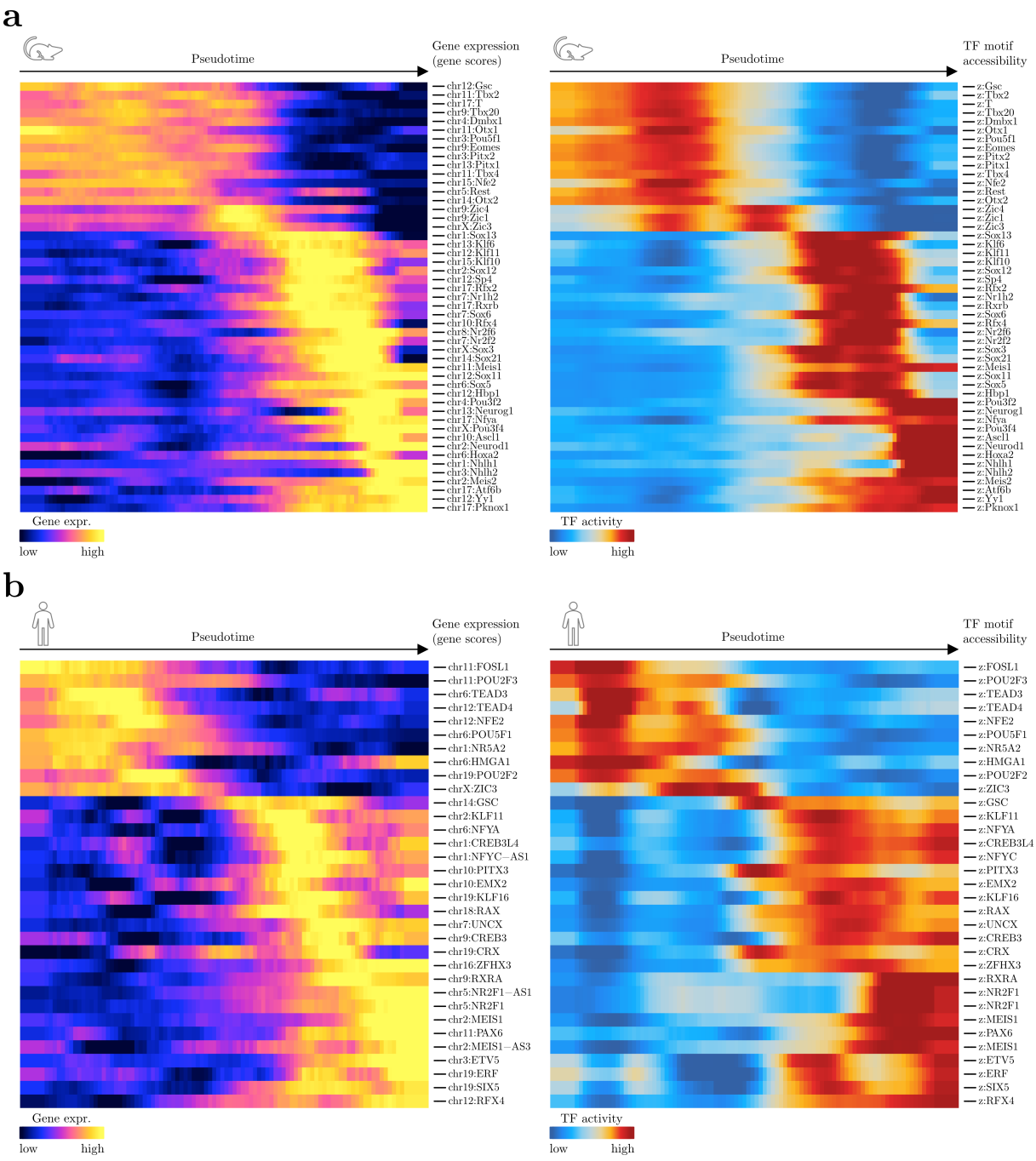

Figure 5.12: Correlated regulatory dynamics between gene scores and TF activity. **a-b)** Heatmap of TFs for which estimated gene expression is positively correlated with chromVAR TF motif deviation across the NPC differentiation trajectory for mouse **(a)** and human **(b)** cells.

Overall, by leveraging trajectory inference methods and integrating gene expression and chromatin accessibility data, I was able to uncover the temporal dynamics of NPC differentiation and identify potential regulatory mechanisms involved during cellular development.

In summary, I characterized differentiating PSCs from mouse, macaca and human under identical conditions using timelapsed combined scRNA and scATAC Seq. My analysis revealed

that species-wide differences during NPC differentiation were accurately captured in the gene expression and chromatin accessibility profiles of single cells, and were reflected in distinct patterns of gene expression and chromatin accessibility as cells transitioned from pluripotent to neural states. Furthermore, I observed variations in cell cycle distribution and cell size among the different species, indicating that species-specific cell physiology plays a crucial role in determining the speed of differentiation. Cross-species mapping identified species-specific global rates of NPC differentiation, revealing that mouse differentiation occurs at a speed 2.2 times faster than macaca differentiation and 2.4 times faster than human differentiation. Finally, integrating gene expression and chromatin accessibility along the NPC differentiation trajectory, I identified species-specific and conserved TFs exhibiting similar dynamics between gene expression and chromatin accessibility, revealing putative drivers of NPC differentiation.

The study provides a comprehensive characterization of NPC differentiation in PSCs from multiple species. I demonstrate the influence of species-specific factors on differentiation speed and highlight the utility of gene expression and chromatin accessibility data for unraveling regulatory dynamics. These findings contribute to a better understanding of the molecular processes and species-specific differences during NPC differentiation.

5.4 Discussion

Pluripotent stem cells (PSCs) hold immense potential for regenerating damaged or destroyed tissues by differentiating into specific cell types required for tissue repair. In particular, patient-specific iPSC lines provide a customizable and biocompatible platform and have become indispensable for disease modeling and high-throughput screening for drug discovery and toxicity testing [71–73]. However, the clinical use of PSCs for cell differentiation or tissue repair beyond investigational regenerative medicine remains limited, which, among others, is due to slow and inefficient differentiation processes, as well as immature characteristics of derived cell types [59, 63, 65, 71, 74, 75]. As differentiation of PSCs into functional cell types can take several weeks and the yield of functional cells is usually quite low, there is a growing interest in accelerating PSC differentiation to produce functional cell types more rapidly and efficiently to improve regenerative medicine approaches [14, 75]. The complex process of PSC differentiation can be better understood by studying the developmental timescales underlying cellular differentiation. Consequently, studying developmental timescales is critical for realizing the vast potential of PSCs in cellular therapies [78].

One interesting direction of recent regenerative medicine efforts where the timing of differentiation plays a key role, is the generation of human organs in animals, which has garnered considerable attention for transplantation purposes. One approach employed in this context is interspecies blastocyst complementation, where human PSCs are introduced into embryos of other species to generate chimeras [327]. While successful generation of mouse chimeras with rat pancreas has been achieved by genetically modifying the mice to prevent their own cells from developing into a pancreas [86], the consequences of species-specific developmental timing become evident when examining the engraftment of stem cells injected into non-related host blastocysts. For example, human stem cell engraftment in pig pre-implantation blastocysts showed limited contribution at post-implantation stages, with only one human cell per 100,000 host cells [328]. When injecting macaca cells into the same host, the proportion of monkey-origin cells ranges from one in 1,000 to one in 10,000 cells [329]. These disparities in developmental speed suggest that the slower-progressing cells are likely outcompeted by the more rapidly developing host cells. Therefore, the generation of human organs in animals requires an understanding of the origins and mechanisms underlying developmental timing to overcome the species barrier and achieve synchronized developmental progression between PSCs and the host organism [327]. Moreover, knowledge of species-specific developmental speed has the potential to assist in the selection of appropriate animal models for preclinical studies. By utilizing animal models that closely resemble the developmental speed of human cells, researchers can more accurately evaluate the safety and efficacy of cellular therapies before advancing to human clinical trials.

Previous studies that investigated species-specific development have often only focused on well-established human and mouse cells, and have employed distinct culture conditions for

each species, resulting in limited cross-species comparability. Additionally, these studies have typically lacked single cell resolution, which hinders a comprehensive understanding of cellular heterogeneity within populations [78, 81, 88]. By harmonizing culture media and optimizing differentiation methods, we established a robust experimental platform that facilitated direct and unbiased comparisons of differentiating PSCs across multiple species. This approach enabled us to overcome the confounding effects of distinct culture conditions, enabling a more accurate assessment of species-specific differences. I then investigated species-specific developmental speed underlying NPC differentiation by characterizing differentiating PSCs from human, mouse, and macaca utilizing timelapsed multiomic single cell gene expression and chromatin accessibility profiling.

This study unveiled distinct species-specific variations at various molecular levels during the process of NPC differentiation. By examining gene expression patterns and chromatin accessibility dynamics over time, I observed differences in the temporal expression and accessibility of key markers involved in the process of NPC differentiation. By mapping cells from different species into a common embedding space, I gained insights into the overall speed of differentiation, as reflected by the distinct cell stages reached by cells from each species. Notably, I found that the overall speed of differentiation varied across species, with mouse cells displaying a comparatively accelerated progression compared to macaca and human cells. Furthermore, my investigation of temporal dynamics enabled the identification of species-specific and conserved drivers of NPC differentiation. To enhance our understanding of cross-species differentiation dynamics, computational tools like *moscot* utilize multiomic data from multiple timepoints to map cellular states over time and space using optimal transport [330]. Applying *moscot* to cross-species multiomic data during NPC differentiation could help identify similarities and differences in species differentiation trajectories, offering additional insights into species-specific and conserved temporal progression of cellular states during NPC differentiation. Nontheless, the findings of this study elucidates the complex molecular mechanisms governing PSC differentiation and provide valuable insights into the species-specific developmental speed underlying cellular development.

At the cellular level, differences in species-specific timing were shown to be reflected in the cell cycle, with PSCs progressing quickly through cell division by minimizing the time of cell cycle gap phases [331]. Although core aspects of the cell cycle seem to be mostly conserved between species, varying cell division rates among species requires a flexible, physiology-coupled cell cycle [332], leading to differences in duration of both the cell cycle as a whole and respective phases among species [92, 331, 333]. Strikingly, I observed a rise in the G1 fraction of mouse cells as the differentiation progressed, with nearly all cells entering the G1 phase by day 7. Intriguingly, when compared to the other species, mouse cells consistently exhibited the lowest number of captured genes across all three cell cycle phases. Absolute measurements of cell cycle length based on tracking single cells through the respective cell cycle phases revealed a shorter cell cycle in mouse cells compared to human cells. This observation, combined with the differences in the

absolute number of captured genes per cell between species, suggests that species-specific cell physiology may play a crucial role in determining the speed of differentiation.

What ultimately drives species-specific developmental speed remains unclear [334, 335]. However, it is believed that differences in biochemical reaction rates, including the rates of protein and mRNA production and degradation, play a crucial role in determining species-specific developmental timing [81, 88, 336]. Indeed, as protein turnover affects protein abundance and activity and thereby impacts cellular processes, protein degradation rates were shown to influence developmental speed [81, 337]. In line with the increased cell cycle length we observed when comparing human against mouse cells, Rayon et al. correlated the observed slower rate of human differentiation with an increase in protein stability and cell cycle duration in human cells compared to mouse cells, as these factors directly influence the abundance and activity of key regulatory proteins involved in developmental processes [81].

The cause of these differential biochemical reaction rates between species has been attributed to the rate of metabolic activity [336, 338]. Measuring the period of the segmentation clock using human and mouse PSCs, Diaz-Cuadros et al. show that mass-specific metabolic rates scale with the developmental rate and are therefore higher in mouse cells than in human cells [336]. Furthermore, measurements in mitochondria oxidative activity and glucose metabolism in human and mouse developing cortical neurons revealed a species-specific timeline of functional mitochondria maturation, with mouse neurons exhibiting an accelerated increase in mitochondria-dependent oxidative activity compared to their human counterparts [338]. Therefore, studying how energy is processed in cells from the three species could help us understand the differences in how these species use mitochondria and glycolysis for energy production.

Despite the limited understanding regarding the precise molecular mechanisms governing developmental timing, studies have demonstrated the susceptibility of developmental timing to manipulation. For instance, by modifying culture conditions, researchers have successfully accelerated the differentiation of PSCs into neural cell types [339–341]. However, it is often challenging to discern whether developmental timing was altered per se or if the acceleration simply reflects an improvement in the culture conditions, allowing for a closer approximation of the in vivo differentiation rate [78, 342].

The preservation of species-specific developmental timing in PSCs therefore presents a practical challenge for regenerative medicine, but also offers a unique opportunity to utilize PSCs as a novel tissue culture model for investigating the intricate relationship between developmental timing in mammals [78]. Studying species-specific developmental timing of PSCs promises valuable insights into underlying molecular mechanisms to unlock potential strategies for enhancing cellular therapies, thus opening doors to improved personalized therapeutic approaches based on a deeper understanding of PSC differentiation.

6 Summary and outlook

Recent breakthroughs in single cell omics have opened the door to understanding cellular function, development, and disease dynamics at a single cell level, offering a paradigm shift in the development and optimization of cellular therapies. However, the data generated from these technologies is relatively novel, and consequently, data analysis remains a significant challenge. This thesis presents the use of single cell technologies to transform and advance cellular therapies from the perspective of computational data analysis. The thesis starts with introducing the wide field of cellular therapies and highlights the use of chimeric antigen receptor (CAR) T cells for cancer treatment approaches and pluripotent stem cells (PSCs) for regenerative medicine, before introducing general concepts of single cell biology (chapter 1). The methods chapter (chapter 2) addresses the variable properties of single cell gene expression (scRNA Seq) and single cell chromatin accessibility (scATAC Seq) data by describing computational workflows of scRNA Seq and scATAC Seq data analysis. Sections 3,4, and 5 present scientific contributions.

CAR T cell therapy faces challenges in translating its success to non-B cell malignancies due to a lack of safe targets, evident from the observed toxicities and limited efficacy of CAR T cells. Chapter 3 therefore introduces a computational approach to predict targets for CAR T cell therapy based on scRNA Seq data. Applying this approach to AML led to the discovery of two previously unrecognized targets. Extensive functional validation of these established CAR T cells shows robust efficacy with minimal off-target toxicity toward healthy human tissues, providing a strong rationale for further clinical development. With a rapid increase in clinical trials, careful consideration of the safety and potential risks associated with CAR target antigen selection is essential before initiating clinical testing. The second contribution is therefore an extensive analysis of scRNA Seq profiles of CAR targets from clinical trials in follicular lymphoma, multiple myeloma, and acute lymphoblastic leukemia tumors as well as healthy tissues across the human body (chapter 4). While I could not observe a clear association between global gene expression profiles of CAR targets and toxicities in the clinics, this chapter provides valuable high resolution CAR target expression profiles and proposes novel targets that exhibit a favorable gene expression pattern for the treatment of the above mentioned diseases.

The translational potential of PSCs for cell differentiation or tissue repair remains constrained by slow and inefficient differentiation processes and immature characteristics of derived cell types. Accelerating the differentiation of PSCs to yield functional cell types more rapidly and efficiently requires a fundamental understanding of developmental timescales underlying cellular differentiation. Chapter 5 therefore contributes a characterization of species-specific developmental timing of PSCs derived from three mammalian species. Timelapsed profiling of differentiating cells using combined scRNA Seq and scATAC Seq unveiled distinct species-specific variations in gene expression, global differentiation rates, cell cycle phases, chromatin accessibility, transcription factor activity and potential drivers of neural progenitor cell differentiation.

6.1 The impact of single cell omics on future cellular treatments

Breakthroughs in single cell technologies over the last decade have paved the way for the creation of comprehensive single cell and tissue atlases [343], which play a vital role in advancing our understanding of health and disease [104]. These technologies hold immense significance in medicine, as they enable the identification of underlying mechanisms of disease initiation and progression, novel disease signatures for diagnostic purposes, understanding the mechanisms underlying treatment resistance, and the discovery and optimization of molecular gene and cell therapies [104].

High throughput single cell technologies are instrumental for the discovery of rare, disease-relevant cell types, which may express critical disease-associated genes and offer potential targets for treatment and therapy [104, 344]. Furthermore, these technologies enable the characterization of shifts in cellular composition between healthy and diseased individuals or following therapeutic perturbations [104, 109] and can serve as vital tools in drug development by identifying cells and molecular pathways with high therapeutic potential for specific diseases [345]. In the realm of regenerative medicine, single cell data can facilitate the identification of regenerative mechanisms within human tissues that can serve as therapeutic targets. Additionally, they contribute to the development of more robust disease models by comparing healthy and diseased tissue and predicting potential mechanisms to enhance the accuracy of these models [104].

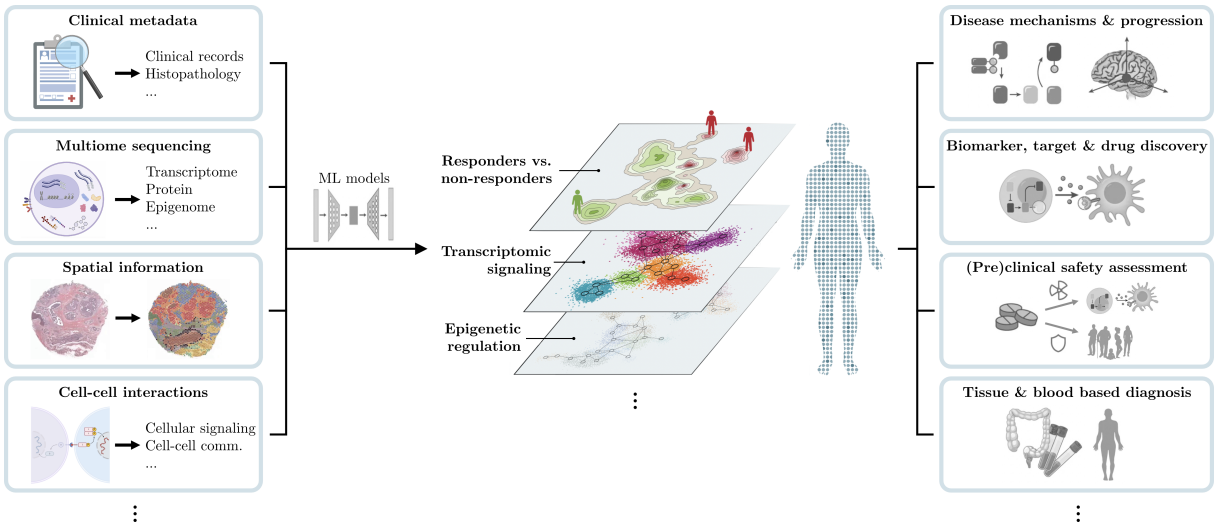

Figure 6.1: Towards a multi-layered understanding of health and disease. Machine learning models, such as foundation models, have the potential to comprehensively capture cellular variation across multi-layered data. These integrated approaches enable significant breakthroughs in cellular therapies and could revolutionize our understanding of cellular biology across health and disease. Adapted from Ginhoux et al. [121] and Rood et al. [104] using elements from BioRender.com (2023).

Initiatives like the Human Cell Atlas (HCA) and The Cancer Genome Atlas (TCGA) are creating extensive cross tissue single cell atlases, increasing our understanding of health and disease on

a single cell level. Foundation models could likely emerge as useful tools to decipher cellular diversity and molecular signatures [346]. These models integrate vast datasets from molecular, spatial and morphometric modalities to comprehensively learn cellular variation. As these models evolve and become more accessible, they could play a key role in understanding and predicting perturbation effects. Moreover, they could facilitate the targeted design of disease perturbations, potentially steering cells towards desired states. This has the potential to not only revolutionize cellular therapies, but offer significant breakthroughs in our understanding of cellular biology across health and disease (Figure 6.1) [104, 346].

Considerable breakthroughs have been made in targeting tumor-associated proteins with small molecules, antibodies or reprogrammed immune cells. However, a crucial realization is that, for most cancer types, there is no single ideal therapeutic target due to the inherent risk of cross-reactive toxicity with cells in healthy tissue. Moreover, the highly heterogeneous and constantly mutating nature of cancer commonly leads to resistance to targeted therapies through antigenic loss or escape mutations [347, 348].

In the future, research on targeted therapies will likely move away from focusing on individual molecules and shift towards recognizing patterns of molecular features present in tumors [347]. Utilizing single cell data from cancer and normal tissue databases will enable the identification of discriminative patterns that differentiate tumors from relevant normal tissues [347, 349]. Drug discovery efforts must also consider the probable mechanisms of tumor variability and escape to design strategies that encompass a wide range of therapeutic possibilities [347]. In the context of targeted cellular therapies, single cell data can be utilized to identify cell populations that are dispensable or targetable within the body, considering factors such as age or sex. For instance, the clinical success of CD19-directed therapies stems from the fact that patients can survive with B cell aplasia [23, 347]. Additionally, the ever-expanding collection of large single cell atlases allows for the analysis of differences between patients in cell populations and target expressing cells across diverse donor groups [347]. Lastly, it is imperative to shift the focus from disease onset to disease prevention, prognosis and progression, as most single cell data currently available is derived from the former [346].

6.2 Expanding the horizons of single cell omics

Despite the remarkable possibilities of single cell data, there are still many unexplored opportunities and areas for improvement. For example, how can we understand the function of newly identified cell types, their relation to the underlying dynamics of a biological system and their implications in disease? How can we characterize cell-cell interactions to unravel tissue organization in health and disease? How can we generate new computational tools that keep up with continuously increasing amounts of data and enable integration of multiple data modalities to unlock novel biological insights?

To address these challenges, the integration of additional omics types is essential to explore different layers of cellular and tissue information. Recently, spatial transcriptomics has garnered substantial attention due to its rapidly increasing resolution, offering insights into cellular composition, cell-cell interactions and molecular interactions. Commercial platforms such as Visium and the innovative Xenium platform from 10X Genomics offer targeted spatial profiling of genes and proteins at remarkable subcellular resolution. Computational approaches for analysis of spatial data include Squidpy, which combines tools from omics and image analysis to enable scalable storing, manipulation and interactive visualization of spatial molecular data [350]. These advancements holds great promise for targeted cellular therapies, as understanding cellular locations within heterogeneous tumors and their surrounding microenvironment becomes critical for optimizing treatment outcomes. For instance, in the case of CAR T cell therapy, restricted trafficking and tumor infiltration, coupled with the immunosuppressive tumor microenvironment, still limit its effectiveness in solid tumors [351]. To mount an effective response, CAR T cells must bypass this hostile tumor microenvironment, infiltrate the tumor, recognize their tumor-specific antigen, and then differentiate and persist as memory T cells that provide long-term protection [352]. Integrating spatial technologies can shed light on the tumor microenvironment, potentially explaining performance limitations of CAR T therapies in solid tumors.

However, integrating multiple layers of data creates computational challenges, and requires development of new computational tools and methods for seamless data integration and analysis. Therefore, combining single cell data with machine learning and more recently, deep learning algorithms, further empower emerging cellular therapies by guiding the design of highly actionable therapeutic approaches [347]. Deep learning methods hold the potential to integrate large-scale datasets to enable comparative studies across species, offering molecular insights into disease phenomena, simulating the effects of drugs or other perturbations, and facilitating drug discovery efforts in relevant model organisms [346]. Deep learning approaches have already shown promise in drug response prediction in cancer, such as a deep variational autoencoder for imputing drug response [353], and *CDRscan*, a convolutional neural net that predicts drug effectiveness [354]. Moreover, the emergence of large language models provides an exciting opportunity for biology. These generative neural networks trained on extensive datasets in an unsupervised manner could potentially offer a comprehensive understanding of cellular variation across all available covariates by connecting molecular, spatial, and phenotypic modalities [346, 355].

In addition to computational challenges, accessibility and usability of data from diverse fields pose significant challenges for researchers. Lack of documentation and standardization in and between clinical studies, as observed during the evaluation of clinical data from CAR T cell therapy patients (chapter 4), significantly hindered data utilization and sharing. To overcome these hurdles, building extensive biobank resources with rich and standardized metadata, alongside

large-scale profiling of samples from clinically annotated and diverse cohorts, is essential. Recent computational efforts like scverse [135] provide foundational support for single cell data modalities and analysis, offering computational frameworks that researchers can build upon with confidence in ongoing maintenance and improvement. Such initiatives play a vital role in extracting meaningful insights from the ever-growing and increasingly complex single cell data landscape.

6.3 Insights from clinical data to advance CAR T cell therapies

Based on unprecedented response rates and likely cures of various B cell malignancies, CAR T cells against CD19 and BCMA have been approved and are thus now part of the standard of care. Enormous effort has been invested in moving these advanced therapeutic products into other hematological malignancies and, importantly, solid tumors, with limited reported success [32, 56, 209, 270]. Since its first administration, over a decade of follow-up data on the first patients receiving CD19-targeted CAR T cells for B cell malignancies has become available, as well as data from numerous studies involving patients undergoing CAR T cell therapies [356, 357], which is crucial for identifying factors associated with successful outcomes. Leveraging this wealth of clinical data has shown that, despide many repeated forms of toxicities, the overarching challenge remains achieving high response rates and minimizing relapse rates [22, 357, 358].

Efforts to delineate factors influencing the success of CAR T cell therapy have been extensive. High tumor burden, elevated baseline inflammation levels, a higher Eastern Cooperative Oncology Group (ECOG) score, a measure of a patient's functional status in serious illness, and the gut microbiome have unsurprisingly been linked to reduced therapy efficacy [357, 359]. Notably, absolute levels of CAR T cells emerged as a significant predictor of treatment response. CAR T cells undergo rapid expansion post-infusion, reaching peak levels, and subsequently persisting at lower levels for an extended period. Higher peak levels and sustained CAR expression during the initial month post-infusion have consistently correlated with improved treatment responses [24, 356, 357, 360]. Therefore, promoting CAR T cell expansion and mitigating exhaustion represent promising strategies for future advancements [361].

Integration of single cell sequencing at various stages of CAR T cell therapy holds significant promise. T cells exhibit a range of differentiation states with diverse functions [362], and these heterogeneous characteristics have been associated with CAR T cell responses [363]. However, the optimal T cell composition for CAR T cell therapy remains uncertain, although evidence suggests that the presence of less-differentiated naive T cells or central memory T cells is critical for favorable adoptive cell therapy responses [357]. Single cell technologies can aid in characterizing T cell heterogeneity in patients and identifying specific cell signaling pathways that can be targeted to induce desired shifts in T cell phenotypes. Additionally, early screening of a patient's tumor landscape using single cell sequencing could serve as a predictive indicator of treatment-responsive antigen-expressing cells, offering valuable insights in treatment planning and patient outcomes.

6.4 Mitigating antigen escape through dual targeting CARs

Chapter 3 of this thesis has introduced a computational approach aimed at identifying novel targets for CAR T cell therapy, addressing the critical issue of tumor-specific and safe targets. I have identified two previously unrecognized targets for CAR T cells in AML, resulting in the filing of two patents. These targets are currently undergoing preclinical testing to ensure compliance with Good Manufacturing Practice (GMP) standards. The ultimate goal is to progress these targets into clinical trials for potential therapeutic application. However, it is apparent that the choice of target antigen represents just one determinant of the success of CAR T cell therapy. Even if the tumor cells initially express the target antigen, there is a potential for subpopulations of cancer cells to emerge during treatment that exhibit reduced or lost antigen expression, a phenomenon termed antigen loss or antigen escape. This poses a formidable challenge and constitutes one of the major limitations of CAR T cell therapy [22]. Antigen-negative or low-antigen-expressing cancer cells can evade CAR T cell recognition, thereby enabling tumor evasion, leading to possible disease relapse. For example, CD19-targeted CAR T cell therapy demonstrates durable responses in 70-90% of acute lymphoblastic leukemia patients; however, follow-up data indicate that 30-70% of patients who experience disease recurrence after treatment exhibit downregulation or loss of the CD19 antigen [348, 364].

Dual-targeting approaches represent a promising strategy to overcome these limitations. Leveraging multi-antigen-targeting techniques, boolean logic has been applied to "gate" the activity of CAR T cells, optimizing their efficiency while minimizing toxicity [365]. For instance, T cells equipped with two independent CAR molecules or a mixture of distinct specific CAR T cells can utilize the "OR" logic gate, which allows CAR T cells to exert antitumor effects in the presence of either targeted antigen. By targeting two tumor-associated antigens with the "OR" logic gate, CAR T cells can surmount the issue of antigen escape, as the presence of an alternative target ensures continued efficacy even if one target is lost. Conversely, "AND" logic-gated CAR T cells require simultaneous presence of both antigens for activation, achieving highly selective antitumor efficacy. Moreover, the "NOT" logic gate can be harnessed to enable engineered T cells to discriminate between target and non-target cells, preventing CAR-mediated killing of healthy cells and thereby enhancing the safety of CAR T cells. Notably, this approach demands careful calibration of the expression levels of both the "NOT" receptor and CAR receptor, as well as relatively high expression levels of the "NOT" antigen to achieve full inhibition of CAR T cell killing [347].

Still, the first study using dual-targeting CD19/CD22 CAR T cells for treatment of acute lymphoblastic leukemia has reported antigen loss in over 25% of patients [366, 367]. The utilization of single cell technologies assumes a pivotal role in this context. Since single targets often possess limited discriminatory capabilities in distinguishing most tumors from normal tissue, large-scale screening of tumor and healthy tissues can facilitate the identification of suitable

antigens and the prediction of optimal combinations of all logic gates [233, 368]. For instance, Ahmadi and colleagues harnessed publicly available single cell tumor transcriptomic data to predict the most suitable target combinations for precise and selective cancer targeting [368]. In addition, generative models capable of predicting cellular responses to various perturbations could be employed to predict synergistic or antagonistic interactions of different combinations of CAR targets alongside other therapeutic interventions [321, 369]. This could enable the identification of optimal combinations of antigens and logic gates, potentially enhancing the efficacy of CAR T cell therapy. While these approaches seem promising, further data is needed to determine whether targeting multiple antigens can indeed overcome the challenge of antigen escape, or if it necessitates combination with other enhancements in CAR therapies.

6.5 Potential off-the-shelf approaches of adoptive cellular therapy

Despite the significant advancements in CAR T cell therapy, its high cost of several hundred thousand dollars per patient, and the prolonged manufacturing process of approximately 1-2 months remain major hurdles [35, 36]. Furthermore, the labor-intensive and expensive cell manufacturing processes, along with inadequate commercial scaling, have hindered the widespread implementation of CAR T cell therapies to meet clinical demands [32]. Although ongoing research and development hold promise for cost reduction over time, there is an urgent need for more accessible and "off-the-shelf" universal therapeutics for diverse patient populations. One potential solution lies in the utilization of lipid nanoparticles carrying mRNA instructions to reprogram T lymphocytes into functional CAR T cells in vivo, offering transiently engineered T cells, thereby minimizing toxicities associated with lymphodepletion and enabling precise dosing [370].

In pursuit of enhanced controllability, flexibility, and selectivity, adapter CARs (AdCARs) have emerged as a recent innovation [371]. These AdCAR T cells use a linker-label-epitope specific to biotin-labeled adapter molecules, which, in turn, bind to tumor-specific antigens. Rather than binding directly to their target cancer cells, AdCAR T cells employ an intermediate linker, providing a versatile antigen selection strategy through infusion of multiple target-specific adapter molecules into patients. Additionally, AdCARs can be switched on or off with ease, by withholding or administering the adapter molecules, making them a promising and convenient off-the-shelf therapeutic solution [371].

Besides CARs, Bi-specific T cell engagers (BiTEs) represent another compelling approach to adoptive cell therapy. BiTEs are recombinant bispecific proteins designed to tether a T cell to a cancer cell and trigger an immune response, comprising two linked single-chain variable fragments from distinct antibodies, with one fragment targeting CD3 on T cells and the other targeting antigens on malignant cells [372]. Unlike CARs, which necessitate the extraction and modification of a patient's T cells, BiTE antibodies directly target existing circulating T cells in

the patient's blood and lymph nodes, rendering them a readily available off-the-shelf treatment option. However, it is important to note that targeting CD3 with BiTEs may inadvertently recruit certain "counterproductive" CD3-positive T cell subsets, such as naive or exhausted T cells [373]. Single cell sequencing could be employed to identify alternative potential T cell targets for BiTEs, mitigating any unwanted side effects.

In 1900, the eminent German chemist and Nobel Laureate, Paul Ehrlich, set forth a visionary notion: the conceptualization of a compound capable of selectively targeting disease-causing agents, thereby obliterating pathogens while sparing the host from harm [374]. Ehrlich aptly referred to these miraculous entities as "Zauberkugel" - magic bullets, a term that has since evolved into a synonymous expression for groundbreaking therapeutic approaches employing agents that precisely target the very structural foundations of afflicted cells. As our understanding of the complexities involved in this endeavor has deepened, continuous advancements in cellular therapies and single cell technologies has brought us ever closer than ever to realizing the visionary concepts originally proposed by Ehrlich more than a century ago.

Bibliography

[1] P. S. Ajmani. "History of Blood Transfusion". *Immunohematology and Blood banking: Principles and Practice*. Ed. by P. S. Ajmani. Singapore: Springer Singapore, 2020, pp. 119–123.

[2] E. Fastag, J. Varon, and G. Sternbach. "Richard Lower: the origins of blood transfusion". en. *J. Emerg. Med.* 44.6 (June 2013), pp. 1146–1150.

[3] P. L. Giangrande. "The history of blood transfusion". en. *Br. J. Haematol.* 110.4 (Sept. 2000), pp. 758–767.

[4] D. D. Farhud and M. Zarif Yeganeh. "A brief history of human blood groups". en. *Iran. J. Public Health* 42.1 (Jan. 2013), pp. 1–6.

[5] E. D. Thomas et al. "Intravenous infusion of bone marrow in patients receiving radiation and chemotherapy". en. *N. Engl. J. Med.* 257.11 (Sept. 1957), pp. 491–496.

[6] N. Granot and R. Storb. "History of hematopoietic cell transplantation: challenges and progress". en. *Haematologica* 105.12 (Dec. 2020), pp. 2716–2729.

[7] E. Hatzimichael and M. Tuthill. "Hematopoietic stem cell transplantation". en. *Stem Cells Cloning* 3 (Aug. 2010), pp. 105–117.

[8] I. Henig and T. Zuckerman. "Hematopoietic stem cell transplantation-50 years of evolution and future perspectives". en. *Rambam Maimonides Med J* 5.4 (Oct. 2014), e0028.

[9] V. Ramachandran, S. S. Kolli, and L. C. Strowd. "Review of Graft-Versus-Host Disease". en. *Dermatol. Clin.* 37.4 (Oct. 2019), pp. 569–582.

[10] M. C. Pasquini. "Impact of graft-versus-host disease on survival". en. *Best Pract. Res. Clin. Haematol.* 21.2 (June 2008), pp. 193–204.

[11] S. J. Lee et al. "Severity of chronic graft-versus-host disease: association with treatment-related mortality and relapse". en. *Blood* 100.2 (July 2002), pp. 406–414.

[12] J. S. Marshall et al. "An introduction to immunology and immunopathology". en. *Allergy Asthma Clin. Immunol.* 14.Suppl 2 (Sept. 2018), p. 49.

[13] E. W. Weber, M. V. Maus, and C. L. Mackall. "The Emerging Landscape of Immune Cell Therapies". en. *Cell* 181.1 (Apr. 2020), pp. 46–62.

[14] A. E.-H. El-Kadiry, M. Rafei, and R. Shammaa. "Cell Therapy: Types, Regulation, and Clinical Benefits". en. *Front. Med.* 8 (Nov. 2021), p. 756029.

[15] K. Khalid et al. "Stem Cell Therapy and Its Significance in HIV Infection". en. *Cureus* 13.8 (Aug. 2021), e17507.

[16] S. Reardon. *Third patient free of HIV after receiving virus-resistant cells*. en. http://dx.doi.org/10.1038/d41586-023-00479-2. Accessed: 2023-4-24. Feb. 2023.

[17] Grand View Research, Inc. *Cell Therapy Market Size & Trends Analysis Report, 2030.pdf.* https://www.grandviewresearch.com/industry-analysis/cell-therapy-market. Accessed: 2023-3-17. 2021.

[18] D. J. Irvine et al. "The future of engineered immune cell therapies". en. *Science* 378.6622 (Nov. 2022), pp. 853–858.

[19] N. M. Mount et al. "Cell-based therapy technology classifications and translational challenges". en. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370.1680 (Oct. 2015), p. 20150017.

[20] I. Kim. "A brief overview of cell therapy and its product". en. *J Korean Assoc Oral Maxillofac Surg* 39.5 (Oct. 2013), pp. 201–202.

[21] C. H. June and M. Sadelain. "Chimeric Antigen Receptor Therapy". en. *N. Engl. J. Med.* 379.1 (July 2018), pp. 64–73.

[22] R. C. Sterner and R. M. Sterner. "CAR-T cell therapy: current limitations and potential strategies". en. *Blood Cancer J.* 11.4 (Apr. 2021), p. 69.

[23] S. L. Maude et al. "Chimeric antigen receptor T cells for sustained remissions in leukemia". en. *N. Engl. J. Med.* 371.16 (Oct. 2014), pp. 1507–1517.

[24] S. S. Neelapu et al. "Axicabtagene Ciloleucel CAR T-Cell Therapy in Refractory Large B-Cell Lymphoma". en. *N. Engl. J. Med.* 377.26 (Dec. 2017), pp. 2531–2544.

[25] S. Yamanaka. "Pluripotent Stem Cell-Based Cell Therapy-Promise and Challenges". en. *Cell Stem Cell* 27.4 (Oct. 2020), pp. 523–531.

[26] M. Wysoczynski and R. Bolli. "A realistic appraisal of the use of embryonic stem cell-based therapies for cardiac repair". en. *Eur. Heart J.* 41.25 (July 2020), pp. 2397–2404.

[27] M. R. Rickels and R. P. Robertson. "Pancreatic Islet Transplantation in Humans: Recent Progress and Future Directions". en. *Endocr. Rev.* 40.2 (Apr. 2019), pp. 631–668.

[28] S. A. Rosenberg and N. P. Restifo. "Adoptive cell transfer as personalized immunotherapy for human cancer". en. *Science* 348.6230 (Apr. 2015), pp. 62–68.

[29] M. W. Rohaan, S. Wilgenhof, and J. B. A. G. Haanen. "Adoptive cellular therapies: the current landscape". en. *Virchows Arch.* 474.4 (Apr. 2019), pp. 449–461.

[30] D. Li et al. "Genetically engineered T cells for cancer immunotherapy". en. *Signal Transduct Target Ther* 4 (Sept. 2019), p. 35.

[31] J. A. Marin-Acevedo et al. "Cancer immunotherapy beyond immune checkpoint inhibitors". en. *J. Hematol. Oncol.* 11.1 (Jan. 2018), p. 8.

[32] L. Labanieh and C. L. Mackall. "CAR immune cells: design principles, resistance and the next generation". en. *Nature* 614.7949 (Feb. 2023), pp. 635–648.

[33] J. Wang, Y. Hu, and H. Huang. "Current development of chimeric antigen receptor T-cell therapy". en. *Stem Cell Investig* 5 (Dec. 2018), p. 44.

[34] Y. Kuwana et al. “Expression of chimeric receptor composed of immunoglobulin-derived V regions and T-cell receptor-derived C regions”. en. *Biochem. Biophys. Res. Commun.* 149.3 (Dec. 1987), pp. 960–968.

[35] G. Choi, G. Shin, and S. Bae. “Price and Prejudice? The Value of Chimeric Antigen Receptor (CAR) T-Cell Therapy”. en. *Int. J. Environ. Res. Public Health* 19.19 (Sept. 2022).

[36] S. Fiorenza et al. “Value and affordability of CAR T-cell therapy in the United States”. en. *Bone Marrow Transplant.* 55.9 (Sept. 2020), pp. 1706–1715.

[37] S. L. Maude et al. “Tisagenlecleucel in Children and Young Adults with B-Cell Lymphoblastic Leukemia”. en. *N. Engl. J. Med.* 378.5 (Feb. 2018), pp. 439–448.

[38] S. J. Schuster et al. “Tisagenlecleucel in Adult Relapsed or Refractory Diffuse Large B-Cell Lymphoma”. en. *N. Engl. J. Med.* 380.1 (Jan. 2019), pp. 45–56.

[39] N. Raje et al. “Anti-BCMA CAR T-Cell Therapy bb2121 in Relapsed or Refractory Multiple Myeloma”. en. *N. Engl. J. Med.* 380.18 (May 2019), pp. 1726–1737.

[40] R. J. Brentjens et al. “Eradication of systemic B-cell tumors by genetically targeted human T lymphocytes co-stimulated by CD80 and interleukin-15”. en. *Nat. Med.* 9.3 (Mar. 2003), pp. 279–286.

[41] S. S. Neelapu. “Managing the toxicities of CAR T-cell therapy”. en. *Hematol. Oncol.* 37 Suppl 1 (June 2019), pp. 48–52.

[42] S. S. Neelapu et al. “Chimeric antigen receptor T-cell therapy - assessment and management of toxicities”. en. *Nat. Rev. Clin. Oncol.* 15.1 (Jan. 2018), pp. 47–62.

[43] C. L. Bonifant et al. “Toxicity and management in CAR T-cell therapy”. en. *Mol Ther Oncolytics* 3 (Apr. 2016), p. 16011.

[44] D. C. Fajgenbaum and C. H. June. “Cytokine Storm”. en. *N. Engl. J. Med.* 383.23 (Dec. 2020), pp. 2255–2273.

[45] M. L. Davila et al. “Efficacy and toxicity management of 19-28z CAR T cell therapy in B cell acute lymphoblastic leukemia”. en. *Sci. Transl. Med.* 6.224 (Feb. 2014), 224ra25.

[46] C. J. Turtle et al. “Immunotherapy of non-Hodgkin’s lymphoma with a defined ratio of CD8+ and CD4+ CD19-specific chimeric antigen receptor-modified T cells”. en. *Sci. Transl. Med.* 8.355 (Sept. 2016), 355ra116.

[47] O. Castaneda-Puglianini and J. C. Chavez. “Assessing and Management of Neurotoxicity After CAR-T Therapy in Diffuse Large B-Cell Lymphoma”. en. *J. Blood Med.* 12 (Aug. 2021), pp. 775–783.

[48] U. H. Acharya et al. “Management of cytokine release syndrome and neurotoxicity in chimeric antigen receptor (CAR) T cell therapy”. en. *Expert Rev. Hematol.* 12.3 (Mar. 2019), pp. 195–205.

[49] C. L. Mackall and D. B. Miklos. “CNS Endothelial Cell Activation Emerges as a Driver of CAR T Cell-Associated Neurotoxicity”. en. *Cancer Discov.* 7.12 (Dec. 2017), pp. 1371–1373.

[50] F. Perna et al. “Integrating Proteomics and Transcriptomics for Systematic Combinatorial Chimeric Antigen Receptor Therapy of AML”. en. *Cancer Cell* 32.4 (Oct. 2017), 506–519.e5.

[51] R. A. Morgan et al. “Case report of a serious adverse event following the administration of T cells transduced with a chimeric antigen receptor recognizing ERBB2”. en. *Mol. Ther.* 18.4 (Apr. 2010), pp. 843–851.

[52] K. R. Parker et al. “Single-Cell Analyses Identify Brain Mural Cells Expressing CD19 as Potential Off-Tumor Targets for CAR-T Immunotherapies”. en. *Cell* 183.1 (Oct. 2020), 126–142.e17.

[53] O. Van Oekelen et al. “Neurocognitive and hypokinetic movement disorder with features of parkinsonism after BCMA-targeting CAR-T cell therapy”. en. *Nat. Med.* 27.12 (Dec. 2021), pp. 2099–2103.

[54] C. H. Lamers et al. “Treatment of metastatic renal cell carcinoma with CAIX CAR-engineered T cells: clinical evaluation and management of on-target toxicity”. en. *Mol. Ther.* 21.4 (Apr. 2013), pp. 904–912.

[55] C. H. J. Lamers et al. “Treatment of metastatic renal cell carcinoma (mRCC) with CAIX CAR-engineered T-cells-a completed study overview”. en. *Biochem. Soc. Trans.* 44.3 (June 2016), pp. 951–959.

[56] U. Patel et al. “CAR T cell therapy in solid tumors: A review of current clinical trials”. en. *EJHaem* 3.Suppl 1 (Jan. 2022), pp. 24–31.

[57] M. Castellarin et al. “A rational mouse model to detect on-target, off-tumor CAR T cell toxicity”. en. *JCI Insight* 5.14 (July 2020).

[58] V. K. Singh et al. “Describing the Stem Cell Potency: The Various Methods of Functional Assessment and In silico Diagnostics”. en. *Front Cell Dev Biol* 4 (Nov. 2016), p. 134.

[59] J. H. Hanna, K. Saha, and R. Jaenisch. “Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues”. en. *Cell* 143.4 (Nov. 2010), pp. 508–525.

[60] J. Poulos. “The limited application of stem cells in medicine: a review”. en. *Stem Cell Res. Ther.* 9.1 (Jan. 2018), p. 1.

[61] K. Takahashi and S. Yamanaka. “Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors”. en. *Cell* 126.4 (Aug. 2006), pp. 663–676.

[62] K. Takahashi et al. “Induction of pluripotent stem cells from adult human fibroblasts by defined factors”. en. *Cell* 131.5 (Nov. 2007), pp. 861–872.

[63] W. Zakrzewski et al. “Stem cells: past, present, and future”. en. *Stem Cell Res. Ther.* 10.1 (Feb. 2019), p. 68.

[64] Y. Shi et al. “Induced pluripotent stem cell technology: a decade of progress”. en. *Nat. Rev. Drug Discov.* 16.2 (Feb. 2017), pp. 115–130.

[65] V. Tabar and L. Studer. “Pluripotent stem cells in regenerative medicine: challenges and recent progress”. en. *Nat. Rev. Genet.* 15.2 (Feb. 2014), pp. 82–92.

[66] D. Ilic and C. Ogilvie. “Concise Review: Human Embryonic Stem Cells-What Have We Done? What Are We Doing? Where Are We Going?” en. *Stem Cells* 35.1 (Jan. 2017), pp. 17–25.

[67] E. Kroon et al. “Pancreatic endoderm derived from human embryonic stem cells generates glucose-responsive insulin-secreting cells in vivo”. en. *Nat. Biotechnol.* 26.4 (Apr. 2008), pp. 443–452.

[68] B. J. Hering et al. “Phase 3 Trial of Transplantation of Human Islets in Type 1 Diabetes Complicated by Severe Hypoglycemia”. en. *Diabetes Care* 39.7 (July 2016), pp. 1230–1240.

[69] P. Menasché et al. “Human embryonic stem cell-derived cardiac progenitors for severe heart failure treatment: first clinical case report”. en. *Eur. Heart J.* 36.30 (Aug. 2015), pp. 2011–2017.

[70] J. Liu. “Induced pluripotent stem cell-derived neural stem cells: new hope for stroke?” en. *Stem Cell Res. Ther.* 4.5 (2013), p. 115.

[71] J. Bragança et al. “Induced pluripotent stem cells, a giant leap for mankind therapeutic applications”. en. *World J. Stem Cells* 11.7 (July 2019), pp. 421–430.

[72] M. S. Elitt, L. Barbar, and P. J. Tesar. “Drug screening for human genetic diseases using iPSC models”. en. *Hum. Mol. Genet.* 27.R2 (Aug. 2018), R89–R98.

[73] C. Kavyasudha et al. “Clinical Applications of Induced Pluripotent Stem Cells – Stato Attuale”. *Cell Biology and Translational Medicine, Volume 1: Stem Cells in Regenerative Medicine: Advances and Challenges*. Ed. by K. Turksen. Cham: Springer International Publishing, 2018, pp. 127–149.

[74] K. Takahashi and S. Yamanaka. “A decade of transcription factor-mediated reprogramming to pluripotency”. en. *Nat. Rev. Mol. Cell Biol.* 17.3 (Mar. 2016), pp. 183–193.

[75] J. Brown et al. “Interspecies chimeric conditions affect the developmental rate of human pluripotent stem cells”. en. *PLoS Comput. Biol.* 17.3 (Mar. 2021), e1008778.

[76] M. Borowiak. “The new generation of beta-cells: replication, stem cell differentiation, and the role of small molecules”. en. *Rev. Diabet. Stud.* 7.2 (Aug. 2010), pp. 93–104.

[77] C. Salinno et al. “β-Cell Maturation and Identity in Health and Disease”. en. *Int. J. Mol. Sci.* 20.21 (Oct. 2019).

[78] C. Barry et al. “Species-specific developmental timing is maintained by pluripotent stem cells ex utero”. en. *Dev. Biol.* 423.2 (Mar. 2017), pp. 101–110.

[79] E. M. Otis and R. Brent. “Equivalent ages in mouse and human embryos”. en. *Anat. Rec.* 120.1 (Sept. 1954), pp. 33–63.

[80] G. La Manno et al. “Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells”. en. *Cell* 167.2 (Oct. 2016), 566–580.e19.

[81] T. Rayon et al. “Species-specific pace of development is associated with differences in protein stability”. en. *Science* 369.6510 (Sept. 2020).

[82] M. Diaz-Cuadros et al. “In vitro characterization of the human segmentation clock”. en. *Nature* 580.7801 (Apr. 2020), pp. 113–118.

[83] N. E. Lewis and J. Rossant. “Mechanism of size regulation in mouse embryo aggregates”. en. *J. Embryol. Exp. Morphol.* 72 (Dec. 1982), pp. 169–181.

[84] M. A. Power and P. P. Tam. “Onset of gastrulation, morphogenesis and somitogenesis in mouse embryos displaying compensatory growth”. en. *Anat. Embryol.* 187.5 (May 1993), pp. 493–504.

[85] M. B. Renfree and J. C. Fenelon. “The enigma of embryonic diapause”. en. *Development* 144.18 (Sept. 2017), pp. 3199–3210.

[86] T. Kobayashi et al. “Generation of rat pancreas in mouse by interspecific blastocyst injection of pluripotent stem cells”. en. *Cell* 142.5 (Sept. 2010), pp. 787–799.

[87] K. Bożyk et al. “Mouse-rat aggregation chimaeras can develop to adulthood”. en. *Dev. Biol.* 427.1 (July 2017), pp. 106–120.

[88] M. Matsuda et al. “Species-specific segmentation clock periods are due to differential biochemical reaction speeds”. en. *Science* 369.6510 (Sept. 2020), pp. 1450–1455.

[89] L. Weinberger et al. “Dynamic stem cell states: naive to primed pluripotency in rodents and humans”. en. *Nat. Rev. Mol. Cell Biol.* 17.3 (Mar. 2016), pp. 155–169.

[90] J. Geuder et al. “A non-invasive method to generate induced pluripotent stem cells from primate urine”. en. *Sci. Rep.* 11.1 (Feb. 2021), p. 3516.

[91] D. Ribatti. “An historical note on the cell theory”. en. *Exp. Cell Res.* 364.1 (Mar. 2018), pp. 1–4.

[92] B. Alberts et al. *Molecular Biology of the Cell - Seventh Edition*. WW Norton & Co, 2022.

[93] J. D. Watson and F. H. Crick. “The structure of DNA”. en. *Cold Spring Harb. Symp. Quant. Biol.* 18 (1953), pp. 123–131.

[94] X. Dai, S. Zhang, and K. Zaleta-Rivera. “RNA: interactions drive functionalities”. en. *Mol. Biol. Rep.* 47.2 (Feb. 2020), pp. 1413–1434.

[95] F. Sanger and A. R. Coulson. “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. en. *J. Mol. Biol.* 94.3 (May 1975), pp. 441–448.

[96] F. Sanger, S. Nicklen, and A. R. Coulson. “DNA sequencing with chain-terminating inhibitors”. en. *Proc. Natl. Acad. Sci. U. S. A.* 74.12 (Dec. 1977), pp. 5463–5467.

[97] S. C. Schuster. “Next-generation sequencing transforms today’s biology”. en. *Nat. Methods* 5.1 (Jan. 2008), pp. 16–18.

[98] R. A. Gibbs. “The Human Genome Project changed everything”. en. *Nat. Rev. Genet.* 21.10 (Oct. 2020), pp. 575–576.

[99] B. E. Slatko, A. F. Gardner, and F. M. Ausubel. “Overview of Next-Generation Sequencing Technologies”. en. *Curr. Protoc. Mol. Biol.* 122.1 (Apr. 2018), e59.

[100] J. M. Heather and B. Chain. “The sequence of sequencers: The history of sequencing DNA”. en. *Genomics* 107.1 (Jan. 2016), pp. 1–8.

[101] L. D. Stein. “The case for cloud computing in genome informatics”. en. *Genome Biol.* 11.5 (May 2010), p. 207.

[102] Wetterstrand, KA. *DNA Sequencing Costs: Data*. https : / / www . genome . gov / sequencingcostsdata. Accessed: 2023-12-29. 2022.

[103] A. Tanay and A. Regev. “Scaling single-cell genomics from phenomenology to mechanism”. en. *Nature* 541.7637 (Jan. 2017), pp. 331–338.

[104] J. E. Rood et al. “Impact of the Human Cell Atlas on medicine”. en. *Nat. Med.* 28.12 (Dec. 2022), pp. 2486–2496.

[105] A. Wagner, A. Regev, and N. Yosef. “Revealing the vectors of cellular identity with single-cell genomics”. en. *Nat. Biotechnol.* 34.11 (Nov. 2016), pp. 1145–1160.

[106] J. Cao et al. “A human cell atlas of fetal gene expression”. en. *Science* 370.6518 (Nov. 2020).

[107] Z. Miao et al. “Multi-omics integration in the age of million single-cell data”. en. *Nat. Rev. Nephrol.* 17.11 (Nov. 2021), pp. 710–724.

[108] S. R. Quake. “A decade of molecular cell atlases”. en. *Trends Genet.* 38.8 (Aug. 2022), pp. 805–810.

[109] L. W. Plasschaert et al. “A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte”. en. *Nature* 560.7718 (Aug. 2018), pp. 377–381.

[110] D. T. Montoro et al. “A revised airway epithelial hierarchy includes CFTR-expressing ionocytes”. en. *Nature* 560.7718 (Aug. 2018), pp. 319–324.

[111] C. A. Jackson and C. Vogel. “New horizons in the stormy sea of multimodal single-cell data integration”. en. *Mol. Cell* 82.2 (Jan. 2022), pp. 248–259.

[112] H. Li et al. “Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma”. en. *Cell* 176.4 (Feb. 2019), 775–789.e18.

[113] M. Sade-Feldman et al. “Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma”. en. *Cell* 176.1-2 (Jan. 2019), p. 404.

[114] J. Liu et al. “Single-cell RNA sequencing of psoriatic skin identifies pathogenic Tc17 cell subsets and reveals distinctions between CD8+ T cells in autoimmunity and cancer”. en. *J. Allergy Clin. Immunol.* 147.6 (June 2021), pp. 2370–2380.

[115] X. Hua et al. “Single-Cell RNA Sequencing to Dissect the Immunological Network of Autoimmune Myocarditis”. en. *Circulation* 142.4 (July 2020), pp. 384–400.

[116] S. Ogbeide et al. “Into the multiverse: advances in single-cell multiomic profiling”. en. *Trends Genet.* (May 2022).

[117] S. A. Morris. “The evolving concept of cell identity in the single cell era”. en. *Development* 146.12 (June 2019).

[118] D. Arendt et al. “The origin and evolution of cell types”. en. *Nat. Rev. Genet.* 17.12 (Dec. 2016), pp. 744–757.

[119] F. Tang et al. “mRNA-Seq whole-transcriptome analysis of a single cell”. en. *Nat. Methods* 6.5 (May 2009), pp. 377–382.

[120] X. Tang et al. “The single-cell sequencing: new developments and medical applications”. en. *Cell Biosci.* 9 (June 2019), p. 53.

[121] F. Ginhoux et al. “Single-cell immunology: Past, present, and future”. en. *Immunity* 55.3 (Mar. 2022), pp. 393–404.

[122] J. Ding et al. “Systematic comparison of single-cell and single-nucleus RNA-sequencing methods”. en. *Nat. Biotechnol.* 38.6 (June 2020), pp. 737–746.

[123] C. Ziegenhain et al. “Comparative Analysis of Single-Cell RNA Sequencing Methods”. en. *Mol. Cell* 65.4 (Feb. 2017), 631–643.e4.

[124] F. C. Grandi et al. “Chromatin accessibility profiling by ATAC-seq”. en. *Nat. Protoc.* 17.6 (Apr. 2022), pp. 1518–1552.

[125] M. J. Boland, K. L. Nazor, and J. F. Loring. “Epigenetic regulation of pluripotency and differentiation”. en. *Circ. Res.* 115.2 (July 2014), pp. 311–324.

[126] J. D. Buenrostro et al. “Single-cell chromatin accessibility reveals principles of regulatory variation”. en. *Nature* 523.7561 (July 2015), pp. 486–490.

[127] A. Gottschlich et al. “Single-cell transcriptomic atlas-guided development of CAR-T cells for the treatment of acute myeloid leukemia”. en. *Nat. Biotechnol.* (Mar. 2023), pp. 1–15.

[128] T. S. Andrews et al. “Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data”. en. *Nat. Protoc.* 16.1 (Jan. 2021), pp. 1–9.

[129] L. Zappia and F. J. Theis. “Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape”. en. *Genome Biol.* 22.1 (Oct. 2021), p. 301.

[130] L. Heumos et al. “Best practices for single-cell analysis across modalities”. en. *Nat. Rev. Genet.* (Mar. 2023), pp. 1–23.

[131] H. Heaton et al. “Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes”. en. *Nat. Methods* 17.6 (May 2020), pp. 615–620.

[132] H. M. Kang et al. “Multiplexed droplet single-cell RNA-sequencing using natural genetic variation”. en. *Nat. Biotechnol.* 36.1 (Jan. 2018), pp. 89–94.

[133] A. Butler et al. “Integrating single-cell transcriptomic data across different conditions, technologies, and species”. en. *Nat. Biotechnol.* 36.5 (June 2018), pp. 411–420.

[134] F. A. Wolf, P. Angerer, and F. J. Theis. “SCANPY: large-scale single-cell gene expression data analysis”. en. *Genome Biol.* 19.1 (Feb. 2018), p. 15.

[135] I. Virshup et al. “The scverse project provides a computational ecosystem for single-cell omics data analysis”. en. *Nat. Biotechnol.* (Apr. 2023).

[136] G. X. Y. Zheng et al. “Massively parallel digital transcriptional profiling of single cells”. en. *Nat. Commun.* 8 (Jan. 2017), p. 14049.

[137] T. Ilicic et al. “Classification of low quality cells from single-cell RNA-seq data”. en. *Genome Biol.* 17 (Feb. 2016), p. 29.

[138] M. D. Luecken and F. J. Theis. “Current best practices in single-cell RNA-seq analysis: a tutorial”. en. *Mol. Syst. Biol.* 15.6 (June 2019), e8746.

[139] C. A. Lareau et al. “Inference and effects of barcode multiplets in droplet-based single-cell assays”. en. *Nat. Commun.* 11.1 (Feb. 2020), p. 866.

[140] J. D. Bloom. “Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments”. en. *PeerJ* 6 (Sept. 2018), e5578.

[141] N. M. Xi and J. J. Li. “Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data”. en. *Cell Syst* 12.2 (Feb. 2021), 176–194.e6.

[142] S. L. Wolock, R. Lopez, and A. M. Klein. “Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data”. en. *Cell Syst* 8.4 (Apr. 2019), 281–291.e9.

[143] C. S. McGinnis, L. M. Murrow, and Z. J. Gartner. “DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors”. en. *Cell Syst* 8.4 (Apr. 2019), 329–337.e4.

[144] A. T. L. Lun et al. “EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data”. en. *Genome Biol.* 20.1 (Mar. 2019), p. 63.

[145] M. D. Young and S. Behjati. “SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data”. en. *Gigascience* 9.12 (Dec. 2020).

[146] G. K. Marinov et al. “From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing”. en. *Genome Res.* 24.3 (Mar. 2014), pp. 496–510.

[147] S. Picelli et al. “Full-length RNA-seq from single cells using Smart-seq2”. en. *Nat. Protoc.* 9.1 (Jan. 2014), pp. 171–181.

[148] A. T. L. Lun, K. Bach, and J. C. Marioni. “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts”. en. *Genome Biol.* 17 (Apr. 2016), p. 75.

[149] B. Vieth et al. “A systematic evaluation of single cell RNA-seq analysis pipelines”. en. *Nat. Commun.* 10.1 (Oct. 2019), p. 4667.

[150] J. Baran-Gale, T. Chandra, and K. Kirschner. “Experimental design for single-cell RNA sequencing”. en. *Brief. Funct. Genomics* 17.4 (July 2018), pp. 233–239.

[151] M. D. Luecken et al. “Benchmarking atlas-level data integration in single-cell genomics”. en. *Nat. Methods* 19.1 (Jan. 2022), pp. 41–50.

[152] R. Argelaguet et al. “Computational principles and challenges in single-cell data integration”. en. *Nat. Biotechnol.* 39.10 (Oct. 2021), pp. 1202–1215.

[153] H. T. N. Tran et al. “A benchmark of batch-effect correction methods for single-cell RNA sequencing data”. en. *Genome Biol.* 21.1 (Jan. 2020), p. 12.

[154] C. K. Stein et al. “Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat”. en. *BMC Bioinformatics* 16 (Feb. 2015), p. 63.

[155] I. Korsunsky et al. “Fast, sensitive and accurate integration of single-cell data with Harmony”. en. *Nat. Methods* 16.12 (Dec. 2019), pp. 1289–1296.

[156] K. Polański et al. “BBKNN: fast batch alignment of single cell transcriptomes”. en. *Bioinformatics* 36.3 (Feb. 2020), pp. 964–965.

[157] R. Lopez et al. “Deep generative modeling for single-cell transcriptomics”. en. *Nat. Methods* 15.12 (Dec. 2018), pp. 1053–1058.

[158] A. Sarkar and M. Stephens. “Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis”. en. *Nat. Genet.* 53.6 (June 2021), pp. 770–777.

[159] P. Brennecke et al. “Accounting for technical noise in single-cell RNA-seq experiments”. en. *Nat. Methods* 10.11 (Nov. 2013), pp. 1093–1095.

[160] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers. “How many principal components? stopping rules for determining the number of non-trivial axes revisited”. *Comput. Stat. Data Anal.* 49.4 (June 2005), pp. 974–997.

[161] L. McInnes, J. Healy, and J. Melville. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction” (Feb. 2018). arXiv: 1802.03426 [stat.ML].

[162] T. Chari and L. Pachter. “The Specious Art of Single-Cell Genomics”. en. Dec. 2022.

[163] V. D. Blondel et al. “Fast unfolding of communities in large networks”. en. *J. Stat. Mech.* 2008.10 (Oct. 2008), P10008.

[164] V. A. Traag, L. Waltman, and N. J. van Eck. “From Louvain to Leiden: guaranteeing well-connected communities”. en. *Sci. Rep.* 9.1 (Mar. 2019), p. 5233.

[165] A. Duò, M. D. Robinson, and C. Soneson. “A systematic performance evaluation of clustering methods for single-cell RNA-seq data”. en. *F1000Res.* 7 (July 2018), p. 1141.

[166] S. Freytag et al. “Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data”. en. *F1000Res.* 7 (Aug. 2018), p. 1297.

[167] V. Y. Kiselev, T. S. Andrews, and M. Hemberg. “Challenges in unsupervised clustering of single-cell RNA-seq data”. en. *Nat. Rev. Genet.* 20.5 (May 2019), pp. 273–282.

[168] P. V. Kharchenko. “The triumphs and limitations of computational methods for scRNA-seq”. en. *Nat. Methods* 18.7 (July 2021), pp. 723–732.

[169] T. Wang et al. “Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data”. en. *BMC Bioinformatics* 20.1 (Jan. 2019), p. 40.

[170] J. W. Squair et al. “Confronting false discoveries in single-cell differential expression”. en. *Nat. Commun.* 12.1 (Sept. 2021), p. 5692.

[171] C. Soneson and M. D. Robinson. “Bias, robustness and scalability in single-cell differential expression analysis”. en. *Nat. Methods* 15.4 (Apr. 2018), pp. 255–261.

[172] M. E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. en. *Nucleic Acids Res.* 43.7 (Apr. 2015), e47.

[173] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. en. *Bioinformatics* 26.1 (Jan. 2010), pp. 139–140.

[174] U. Raudvere et al. “g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)”. en. *Nucleic Acids Res.* 47.W1 (July 2019), W191–W198.

[175] M. Ashburner et al. “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium”. en. *Nat. Genet.* 25.1 (May 2000), pp. 25–29.

[176] M. Kanehisa et al. “KEGG: new perspectives on genomes, pathways, diseases and drugs”. en. *Nucleic Acids Res.* 45.D1 (Jan. 2017), pp. D353–D361.

[177] I. Tirosh et al. “Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq”. en. *Science* 352.6282 (Apr. 2016), pp. 189–196.

[178] E. Z. Macosko et al. “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets”. en. *Cell* 161.5 (May 2015), pp. 1202–1214.

[179] M. Lotfollahi et al. “Mapping single-cell data to reference atlases by transfer learning”. en. *Nat. Biotechnol.* 40.1 (Jan. 2022), pp. 121–130.

[180] C. Xu et al. “Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models”. *Mol. Syst. Biol.* 17.1 (Jan. 2021), e9620.

[181] J. M. W. Slack. “Conrad Hal Waddington: the last Renaissance biologist?” en. *Nat. Rev. Genet.* 3.11 (Nov. 2002), pp. 889–895.

[182] V. Bergen et al. “Generalizing RNA velocity to transient cell states through dynamical modeling”. en. *Nat. Biotechnol.* 38.12 (Dec. 2020), pp. 1408–1414.

[183] S. Tritschler et al. “Concepts and limitations for learning developmental trajectories from single cell genomics”. en. *Development* 146.12 (June 2019).

[184] W. Saelens et al. “A comparison of single-cell trajectory inference methods”. en. *Nat. Biotechnol.* 37.5 (May 2019), pp. 547–554.

[185] C. Weinreb et al. “Fundamental limits on dynamic inference from single-cell snapshots”. en. *Proc. Natl. Acad. Sci. U. S. A.* 115.10 (Mar. 2018), E2467–E2476.

[186] G. La Manno et al. “RNA velocity of single cells”. en. *Nature* 560.7719 (Aug. 2018), pp. 494–498.

[187] M. Lange et al. “CellRank for directed single-cell fate mapping”. en. *Nat. Methods* 19.2 (Jan. 2022), pp. 159–170.

[188] G. Schiebinger et al. “Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming”. en. *Cell* 176.4 (Feb. 2019), 928–943.e22.

[189] S. J. Clark et al. “Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity”. en. *Genome Biol.* 17 (Apr. 2016), p. 72.

[190] R. Jaenisch and A. Bird. “Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals”. en. *Nat. Genet.* 33 Suppl (Mar. 2003), pp. 245–254.

[191] S. Baek and I. Lee. “Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation”. en. *Comput. Struct. Biotechnol. J.* 18 (June 2020), pp. 1429–1439.

[192] F. Yan et al. “From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis”. en. *Genome Biol.* 21.1 (Feb. 2020), p. 22.

[193] H. Chen et al. “Assessment of computational methods for the analysis of single-cell ATAC-seq data”. en. *Genome Biol.* 20.1 (Nov. 2019), p. 241.

[194] J. M. Granja et al. “ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis”. en. *Nat. Genet.* 53.3 (Mar. 2021), pp. 403–411.

[195] R. Fang et al. “Comprehensive analysis of single cell ATAC-seq data with SnapATAC”. en. *Nat. Commun.* 12.1 (Feb. 2021), p. 1337.

[196] Y. Zhang et al. “Model-based analysis of ChIP-Seq (MACS)”. en. *Genome Biol.* 9.9 (Sept. 2008), R137.

[197] J. Ou et al. “ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data”. en. *BMC Genomics* 19.1 (Mar. 2018), p. 169.

[198] H. M. Amemiya, A. Kundaje, and A. P. Boyle. “The ENCODE Blacklist: Identification of Problematic Regions of the Genome”. en. *Sci. Rep.* 9.1 (June 2019), p. 9354.

[199] D. A. Cusanovich et al. “Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing”. en. *Science* 348.6237 (May 2015), pp. 910–914.

[200] T. Stuart et al. “Single-cell chromatin state analysis with Signac”. en. *Nat. Methods* 18.11 (Nov. 2021), pp. 1333–1341.

[201] D. van Dijk et al. “Recovering Gene Interactions from Single-Cell Data Using Data Diffusion”. en. *Cell* 174.3 (July 2018), 716–729.e27.

[202] H. A. Pliner et al. “Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data”. en. *Mol. Cell* 71.5 (Sept. 2018), 858–871.e8.

[203] J. A. Castro-Mondragon et al. “JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles”. en. *Nucleic Acids Res.* 50.D1 (Nov. 2021), pp. D165–D173.

[204] S. A. Lambert et al. “Similarity regression predicts evolution of transcription factor sequence specificity”. en. *Nat. Genet.* 51.6 (June 2019), pp. 981–989.

[205] P. Kheradpour and M. Kellis. “Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments”. en. *Nucleic Acids Res.* 42.5 (Mar. 2014), pp. 2976–2987.

[206] A. N. Schep et al. “chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data”. en. *Nat. Methods* 14.10 (Oct. 2017), pp. 975–978.

[207] S. Baek, I. Goldstein, and G. L. Hager. “Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity”. en. *Cell Rep.* 19.8 (May 2017), pp. 1710–1722.

[208] R. Pique-Regi et al. “Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data”. en. *Genome Res.* 21.3 (Mar. 2011), pp. 447–455.

[209] S. Lesch et al. “Determinants of response and resistance to CAR T cell therapy”. en. *Semin. Cancer Biol.* 65 (Oct. 2020), pp. 80–90.

[210] K. D. Cummins and S. Gill. “Will CAR T cell therapy have a role in AML? Promises and pitfalls”. en. *Semin. Hematol.* 56.2 (Apr. 2019), pp. 155–163.

[211] M. L. Suvà and I. Tirosh. “Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges”. en. *Mol. Cell* 75.1 (July 2019), pp. 7–12.

[212] Y. Jing et al. “Expression of chimeric antigen receptor therapy targets detected by single-cell sequencing of normal cells may contribute to off-tumor toxicity”. en. *Cancer Cell* (Oct. 2021).

[213] D. Türei et al. “Integrated intra- and intercellular signaling knowledge for multicellular omics analysis”. en. *Mol. Syst. Biol.* 17.3 (Mar. 2021), e9923.

[214] D. Bausch-Fluck et al. “A mass spectrometric-derived cell surface protein atlas”. en. *PLoS One* 10.3 (Apr. 2015), e0121314.

[215] D. Bausch-Fluck et al. “The in silico human surfaceome”. en. *Proc. Natl. Acad. Sci. U. S. A.* 115.46 (Nov. 2018), E10988–E10997.

[216] M. Efremova et al. “CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes”. en. *Nat. Protoc.* 15.4 (Apr. 2020), pp. 1484–1506.

[217] P. J. Thul et al. “A subcellular map of the human proteome”. en. *Science* 356.6340 (May 2017).

[218] N. Habib et al. “Massively parallel single-nucleus RNA-seq with DroNc-seq”. en. *Nat. Methods* 14.10 (Oct. 2017), pp. 955–958.

[219] N. Kim et al. “Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma”. en. *Nat. Commun.* 11.1 (May 2020), p. 2285.

[220] B. J. Stewart et al. “Spatiotemporal immune zonation of the human kidney”. en. *Science* 365.6460 (Sept. 2019), pp. 1461–1466.

[221] K. J. Travaglini et al. “A molecular cell atlas of the human lung from single-cell RNA sequencing”. en. *Nature* 587.7835 (Nov. 2020), pp. 619–625.

[222] E. Madissoon et al. “scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation”. en. *Genome Biol.* 21.1 (Dec. 2019), p. 1.

[223] P. A. Reyfman et al. “Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis”. en. *Am. J. Respir. Crit. Care Med.* 199.12 (June 2019), pp. 1517–1536.

[224] S. A. MacParland et al. “Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations”. en. *Nat. Commun.* 9.1 (Oct. 2018), p. 4383.

[225] P. Ramachandran et al. “Resolving the fibrotic niche of human liver cirrhosis at single-cell level”. en. *Nature* 575.7783 (Nov. 2019), pp. 512–518.

[226] J. B. Cheng et al. “Transcriptional Programming of Normal and Inflamed Human Epidermis at Single-Cell Resolution”. en. *Cell Rep.* 25.4 (Oct. 2018), pp. 871–883.

[227] K. R. James et al. “Distinct microbial and immune niches of the human colon”. en. *Nat. Immunol.* 21.3 (Mar. 2020), pp. 343–353.

[228] X. Han et al. “Construction of a human cell landscape at single-cell level”. en. *Nature* 581.7808 (May 2020), pp. 303–309.

[229] D. S. Wishart et al. “DrugBank 5.0: a major update to the DrugBank database for 2018”. en. *Nucleic Acids Res.* 46.D1 (Jan. 2018), pp. D1074–D1082.

[230] Cancer Genome Atlas Research Network et al. “Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia”. en. *N. Engl. J. Med.* 368.22 (May 2013), pp. 2059–2074.

[231] F. Thol and A. Ganser. “Treatment of Relapsed Acute Myeloid Leukemia”. en. *Curr. Treat. Options Oncol.* 21.8 (June 2020), p. 66.

[232] K. D. Cummins and S. Gill. “Chimeric antigen receptor T-cell therapy for acute myeloid leukemia: how close to reality?” en. *Haematologica* 104.7 (July 2019), pp. 1302–1308.

[233] M. MacKay et al. “The therapeutic landscape for cells engineered with chimeric antigen receptors”. en. *Nat. Biotechnol.* 38.2 (Feb. 2020), pp. 233–244.

[234] F. P. Tambaro et al. “Autologous CD33-CAR-T cells for treatment of relapsed/refractory acute myelogenous leukemia”. en. *Leukemia* 35.11 (Nov. 2021), pp. 3282–3286.

[235] T. Sauer et al. “CD70-specific CAR T-cells have potent activity against Acute Myeloid Leukemia (AML) without HSC toxicity”. en. *Blood* (Mar. 2021).

[236] H. Jetani et al. “Siglec-6 is a novel target for CAR T-cell therapy in acute myeloid leukemia (AML)”. en. *Blood* (July 2021).

[237] R. Myburgh et al. “Anti-human CD117 CAR T-cells efficiently eliminate healthy and malignant CD117-expressing hematopoietic cells”. en. *Leukemia* 34.10 (Oct. 2020), pp. 2688–2703.

[238] H. Tashiro et al. “Treatment of Acute Myeloid Leukemia with T Cells Expressing Chimeric Antigen Receptors Directed to C-type Lectin-like Molecule 1”. en. *Mol. Ther.* 25.9 (Sept. 2017), pp. 2202–2213.

[239] M. Casucci et al. “CD44v6-targeted T cells mediate potent antitumor effects against acute myeloid leukemia and multiple myeloma”. en. *Blood* 122.20 (Nov. 2013), pp. 3461–3472.

[240] P. van Galen et al. “Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity”. en. *Cell* 176.6 (Mar. 2019), 1265–1281.e24.

[241] E. Breman et al. “Overcoming Target Driven Fratricide for T Cell Therapy”. en. *Front. Immunol.* 9 (Dec. 2018), p. 2940.

[242] D. S. Fischer et al. “Sfaira accelerates data and model reuse in single cell genomics”. en. *Genome Biol.* 22.1 (Aug. 2021), p. 248.

[243] C. Muus et al. “Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics”. en. *Nat. Med.* 27.3 (Mar. 2021), pp. 546–559.

[244] A. M. Jurga, M. Paleczna, and K. Z. Kuter. “Overview of General and Discriminating Markers of Differential Microglia Phenotypes”. en. *Front. Cell. Neurosci.* 14 (Aug. 2020), p. 198.

[245] B. Erblich et al. “Absence of colony stimulation factor-1 receptor results in loss of microglia, disrupted brain development and olfactory deficits”. en. *PLoS One* 6.10 (Oct. 2011), e26317.

[246] A. A. Petti et al. “A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing”. en. *Nat. Commun.* 10.1 (Aug. 2019), p. 3660.

[247] D. K. Edwards 5th et al. “CSF1R inhibitors exhibit antitumor activity in acute myeloid leukemia by blocking paracrine signals from support cells”. en. *Blood* 133.6 (Feb. 2019), pp. 588–599.

[248] K. Y. Sletta, O. Castells, and B. T. Gjertsen. “Colony Stimulating Factor 1 Receptor in Acute Myeloid Leukemia”. en. *Front. Oncol.* 11 (Mar. 2021), p. 654817.

[249] A. Haque et al. “A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications”. en. *Genome Med.* 9.1 (Aug. 2017), p. 75.

[250] A. Griciuc et al. “TREM2 Acts Downstream of CD33 in Modulating Microglial Pathology in Alzheimer’s Disease”. en. *Neuron* 103.5 (Sept. 2019), 820–835.e7.

[251] P. A. Cassier et al. “CSF1R inhibition with emactuzumab in locally advanced diffuse-type tenosynovial giant cell tumours of the soft tissue: a dose-escalation and dose-expansion phase 1 study”. en. *Lancet Oncol.* 16.8 (Aug. 2015), pp. 949–956.

[252] D. M. O’Rourke et al. “A single dose of peripherally infused EGFRvIII-directed CAR T cells mediates antigen loss and induces adaptive resistance in patients with recurrent glioblastoma”. en. *Sci. Transl. Med.* 9.399 (July 2017).

[253] J. Gust et al. “Endothelial Activation and Blood-Brain Barrier Disruption in Neurotoxicity after Adoptive Immunotherapy with CD19 CAR-T Cells”. en. *Cancer Discov.* 7.12 (Dec. 2017), pp. 1404–1419.

[254] R. M. Sterner et al. “GM-CSF inhibition reduces cytokine release syndrome and neuroinflammation but enhances CAR-T cell function in xenografts”. en. *Blood* 133.7 (Feb. 2019), pp. 697–709.

[255] A. H. J. Tan, N. Vinanica, and D. Campana. “Chimeric antigen receptor-T cells with cytokine neutralizing capacity”. en. *Blood Adv* 4.7 (Apr. 2020), pp. 1419–1431.

[256] H. Tamura et al. “Expression of functional B7-H2 and B7.2 costimulatory molecules and their prognostic implications in de novo acute myeloid leukemia”. en. *Clin. Cancer Res.* 11.16 (Aug. 2005), pp. 5708–5717.

[257] F. Re et al. “Expression of CD86 in acute myelogenous leukemia is a marker of dendritic/monocytic lineage”. en. *Exp. Hematol.* 30.2 (Feb. 2002), pp. 126–134.

[258] Z. Zheng et al. “Expression patterns of costimulatory molecules on cells derived from human hematological malignancies”. en. *J. Exp. Clin. Cancer Res.* 17.3 (Sept. 1998), pp. 251–258.

[259] C. M. Gavile et al. “CD86 regulates myeloma cell survival”. en. *Blood Adv* 1.25 (Nov. 2017), pp. 2307–2319.

[260] Ł. Sędek et al. “Differential expression of CD73, CD86 and CD304 in normal vs. leukemic B-cell precursors and their utility as stable minimal residual disease markers in childhood B-cell precursor acute lymphoblastic leukemia”. en. *J. Immunol. Methods* 475 (Dec. 2019), p. 112429.

[261] E. C. Guinan et al. “Pivotal role of the B7:CD28 pathway in transplantation tolerance and tumor immunity”. en. *Blood* 84.10 (Nov. 1994), pp. 3261–3282.

[262] L. J. Zhou and T. F. Tedder. “CD14+ blood monocytes can differentiate into functionally mature CD83+ dendritic cells”. en. *Proc. Natl. Acad. Sci. U. S. A.* 93.6 (Mar. 1996), pp. 2588–2592.

[263] C. Smyth et al. “Identification of a dynamic intracellular reservoir of CD86 protein in peripheral blood monocytes that is not associated with the Golgi complex”. en. *J. Immunol.* 160.11 (June 1998), pp. 5390–5396.

[264] H. A. Blair and E. D. Deeks. “Abatacept: A Review in Rheumatoid Arthritis”. en. *Drugs* 77.11 (July 2017), pp. 1221–1233.

[265] P. S. Adusumilli et al. “A Phase I Trial of Regional Mesothelin-Targeted CAR T-cell Therapy in Patients with Malignant Pleural Disease, in Combination with the Anti-PD-1 Agent Pembrolizumab”. en. *Cancer Discov.* 11.11 (Nov. 2021), pp. 2748–2763.

[266] R. G. Majzner and C. L. Mackall. “Clinical lessons learned from the first leg of the CAR T cell journey”. en. *Nat. Med.* 25.9 (Sept. 2019), pp. 1341–1355.

[267] H. Jetani et al. “CAR T-cells targeting FLT3 have potent activity against FLT3-ITD+ AML and act synergistically with the FLT3-inhibitor crenolanib”. en. *Leukemia* 32.5 (May 2018), pp. 1168–1179.

[268] H. Döhner, A. H. Wei, and B. Löwenberg. “Towards precision medicine for AML”. en. *Nat. Rev. Clin. Oncol.* 18.9 (Sept. 2021), pp. 577–590.

[269] S. Huang et al. “Deciphering and advancing CAR T-cell therapy with single-cell sequencing technologies”. en. *Mol. Cancer* 22.1 (May 2023), p. 80.

[270] C. L. Flugel et al. “Overcoming on-target, off-tumour toxicity of CAR T cell therapy for solid tumours”. en. *Nat. Rev. Clin. Oncol.* 20.1 (Jan. 2023), pp. 49–62.

[271] A. V. Hirayama and C. J. Turtle. “Toxicities of CD19 CAR-T cell immunotherapy”. en. *Am. J. Hematol.* 94.S1 (May 2019), S42–S49.

[272] S. Fried et al. “Early and late hematologic toxicity following CD19 CAR-T cells”. en. *Bone Marrow Transplant.* 54.10 (Oct. 2019), pp. 1643–1650.

[273] J. N. Brudno and J. N. Kochenderfer. “Recent advances in CAR T-cell toxicity: Mechanisms, manifestations and management”. en. *Blood Rev.* 34 (Mar. 2019), pp. 45–55.

[274] H. Chen, F. Ye, and G. Guo. “Revolutionizing immunology with single-cell RNA sequencing”. en. *Cell. Mol. Immunol.* 16.3 (Mar. 2019), pp. 242–249.

[275] P. Guruprasad et al. “The current landscape of single-cell transcriptomics for cancer immunotherapy”. en. *J. Exp. Med.* 218.1 (Jan. 2021).

[276] HuBMAP Consortium. “The human body at cellular resolution: the NIH Human Biomolecular Atlas Program”. en. *Nature* 574.7777 (Oct. 2019), pp. 187–192.

[277] Y. Zhang et al. “Single-Cell Analysis of Target Antigens of CAR-T Reveals a Potential Landscape of “On-Target, Off-Tumor Toxicity””. en. *Front. Immunol.* 12 (Dec. 2021), p. 799206.

[278] Tabula Sapiens Consortium* et al. “The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans”. en. *Science* 376.6594 (May 2022), eabl4896.

[279] C. Domínguez Conde et al. “Cross-tissue immune cell analysis reveals tissue-specific features in humans”. en. *Science* 376.6594 (May 2022), eabl5197.

[280] P. A. Szabo et al. “Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease”. en. *Nat. Commun.* 10.1 (Oct. 2019), p. 4706.

[281] J. Berg et al. “Human cortical expansion involves diversification and specialization of supragranular intratelencephalic-projecting neurons”. en. Apr. 2020.

[282] T. E. Bakken et al. “Comparative cellular analysis of motor cortex in human, marmoset and mouse”. en. *Nature* 598.7879 (Oct. 2021), pp. 111–119.

[283] S. W. Lukowski et al. “A single-cell transcriptome atlas of the adult human retina”. en. *EMBO J.* 38.18 (Sept. 2019), e100811.

[284] Y. Muto et al. “Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney”. en. *Nat. Commun.* 12.1 (Apr. 2021), p. 2190.

[285] J.-E. Park et al. “A cell atlas of human thymic development defines T cell repertoire formation”. en. *Science* 367.6480 (Feb. 2020).

[286] M. Xiang et al. “A Single-Cell Transcriptional Roadmap of the Mouse and Human Lymph Node Lymphatic Vasculature”. en. *Front Cardiovasc Med* 7 (Apr. 2020), p. 52.

[287] D. Fawkner-Corbett et al. “Spatiotemporal analysis of human intestinal development at single-cell resolution”. en. *Cell* 184.3 (Feb. 2021), 810–826.e23.

[288] Y. Wang et al. “Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine”. en. *J. Exp. Med.* 217.2 (Feb. 2020).

[289] T. S. Andrews et al. “Single-Cell, Single-Nucleus, and Spatial RNA Sequencing of the Human Liver Identifies Cholangiocyte and Mesenchymal Heterogeneity”. en. *Hepatol Commun* 6.4 (Apr. 2022), pp. 821–840.

[290] M. Baron et al. “A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure”. en. *Cell Syst* 3.4 (Oct. 2016), 346–360.e4.

[291] J. Peng et al. “Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma”. en. *Cell Res.* 29.9 (Sept. 2019), pp. 725–738.

[292] M. Enge et al. “Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns”. en. *Cell* 171.2 (Oct. 2017), 321–330.e14.

[293] M. J. Muraro et al. “A Single-Cell Transcriptome Atlas of the Human Pancreas”. en. *Cell Syst* 3.4 (Oct. 2016), 385–394.e3.

[294] M. Litviňuková et al. “Cells of the adult human heart”. en. *Nature* 588.7838 (Dec. 2020), pp. 466–472.

[295] G. H. Henry et al. “A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra”. en. *Cell Rep.* 25.12 (Dec. 2018), 3530–3542.e5.

[296] P. Bhat-Nakshatri et al. “A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells”. en. *Cell Rep Med* 2.3 (Mar. 2021), p. 100219.

[297] G. Han et al. “Immune microenvironment subtypes and association with tumor cell mutations and antigen expression in follicular lymphoma”. en. Apr. 2021.

[298] S. M. Tirier et al. “Subclone-specific microenvironmental impact and drug response in refractory multiple myeloma revealed by single-cell transcriptomics”. en. *Nat. Commun.* 12.1 (Nov. 2021), p. 6960.

[299] E. Khabirova et al. “Single-cell transcriptomics reveals a distinct developmental state of KMT2A-rearranged infant B-cell acute lymphoblastic leukemia”. en. *Nat. Med.* 28.4 (Apr. 2022), pp. 743–751.

[300] I. Virshup et al. “anndata: Annotated data”. en. Dec. 2021.

[301] L. Jardine et al. “Blood and immune development in human fetal bone marrow and Down syndrome”. en. *Nature* 598.7880 (Oct. 2021), pp. 327–331.

[302] H. M. Levitin, J. Yuan, and P. A. Sims. “Single-Cell Transcriptomic Analysis of Tumor Heterogeneity”. en. *Trends Cancer Res.* 4.4 (Apr. 2018), pp. 264–268.

[303] A. Freedman and E. Jacobsen. “Follicular lymphoma: 2020 update on diagnosis and management”. en. *Am. J. Hematol.* 95.3 (Mar. 2020), pp. 316–327.

[304] T. Wen et al. “Inhibitors targeting Bruton’s tyrosine kinase in cancers: drug development advances”. en. *Leukemia* 35.2 (Feb. 2021), pp. 312–332.

[305] J. L. Munoz et al. “BTK Inhibitors and CAR T-Cell Therapy in Treating Mantle Cell Lymphoma-Finding a Dancing Partner”. en. *Curr. Oncol. Rep.* 24.10 (Oct. 2022), pp. 1299–1311.

[306] M. S. Magee et al. “GUCY2C-directed CAR-T cells oppose colorectal cancer metastases without autoimmunity”. en. *Oncoimmunology* 5.10 (Sept. 2016), e1227897.

[307] M. S. Magee et al. “Human GUCY2C-Targeted Chimeric Antigen Receptor (CAR)-Expressing T Cells Eliminate Colorectal Cancer Metastases”. en. *Cancer Immunol Res* 6.5 (May 2018), pp. 509–516.

[308] S. Sleiman et al. “Anti-FLT3 CAR T Cells in Acute Myeloid Leukemia”. *Blood* 138 (Nov. 2021), p. 1703.

[309] C. H. J. Lamers et al. “Treatment of metastatic renal cell carcinoma with autologous T-lymphocytes genetically retargeted against carbonic anhydrase IX: first clinical experience”. en. *J. Clin. Oncol.* 24.13 (May 2006), e20–2.

[310] E. Drent et al. “A Rational Strategy for Reducing On-Target Off-Tumor Effects of CD38-Chimeric Antigen Receptors by Affinity Optimization”. en. *Mol. Ther.* 25.8 (Aug. 2017), pp. 1946–1958.

[311] K. Feng et al. “Phase I study of chimeric antigen receptor modified T cells in treating HER2-positive advanced biliary tract cancers and pancreatic cancers”. en. *Protein Cell* 9.10 (Oct. 2018), pp. 838–847.

[312] S. Gill et al. “Preclinical targeting of human acute myeloid leukemia and myeloablation using chimeric antigen receptor-modified T cells”. en. *Blood* 123.15 (Apr. 2014), pp. 2343–2354.

[313] A. Mardiros et al. “T cells expressing CD123-specific chimeric antigen receptors exhibit specific cytolytic effector functions and antitumor effects against human acute myeloid leukemia”. en. *Blood* 122.18 (Oct. 2013), pp. 3138–3148.

[314] S. Lesch et al. “T cells armed with C-X-C chemokine receptor type 6 enhance adoptive cell therapy for pancreatic tumours”. en. *Nat Biomed Eng* (June 2021).

[315] B. L. Cadilha et al. “Combined tumor-directed recruitment and protection from immune suppression enable CAR T cell efficacy in solid tumors”. en. *Sci Adv* 7.24 (June 2021).

[316] F. Märkl et al. “Bispecific antibodies redirect synthetic agonistic receptor modified T cells against melanoma”. en. *J Immunother Cancer* 11.5 (May 2023).

[317] C. B. Steen et al. “The landscape of tumor cell states and ecosystems in diffuse large B cell lymphoma”. en. *Cancer Cell* 39.10 (Oct. 2021), 1422–1437.e10.

[318] S. C. van den Brink et al. “Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations”. en. *Nat. Methods* 14.10 (Sept. 2017), pp. 935–936.

[319] L. Pan et al. “Isoform-level quantification for single-cell RNA sequencing”. en. *Bioinformatics* 38.5 (Feb. 2022), pp. 1287–1294.

[320] L. T. Senbanjo and M. A. Chellaiah. “CD44: A Multifunctional Cell Surface Adhesion Receptor Is a Regulator of Progression and Metastasis of Cancer Cells”. en. *Front Cell Dev Biol* 5 (Mar. 2017), p. 18.

[321] K. Kanemaru et al. “Spatially resolved multiomics of human cardiac niches”. en. *Nature* 619.7971 (July 2023), pp. 801–810.

[322] P. Nieto et al. “A single-cell tumor immune atlas for precision oncology”. en. *Genome Res.* 31.10 (Oct. 2021), pp. 1913–1926.

[323] L. Wang et al. “A single-cell atlas of glioblastoma evolution under therapy reveals cell-intrinsic and cell-extrinsic therapeutic targets”. en. *Nat Cancer* 3.12 (Dec. 2022), pp. 1534–1552.

[324] D. Xiong et al. “A comparative study of multiple instance learning methods for cancer detection using T-cell receptor sequences”. en. *Comput. Struct. Biotechnol. J.* 19 (May 2021), pp. 3255–3268.

[325] M. Stauske et al. “Non-Human Primate iPSC Generation, Cultivation, and Cardiac Differentiation under Chemically Defined Conditions”. en. *Cells* 9.6 (May 2020).

[326] S. M. Chambers et al. “Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling”. en. *Nat. Biotechnol.* 27.3 (Mar. 2009), pp. 275–280.

[327] T. Rayon and J. Briscoe. “Cross-species comparisons and in vitro models to study tempo in development and homeostasis”. en. *Interface Focus* 11.3 (June 2021), p. 20200069.

[328] J. Wu et al. “Interspecies Chimerism with Mammalian Pluripotent Stem Cells”. en. *Cell* 168.3 (Jan. 2017), 473–486.e15.

[329] R. Fu et al. “Domesticated cynomolgus monkey embryonic stem cells allow the generation of neonatal interspecies chimeric pigs”. en. *Protein Cell* 11.2 (Feb. 2020), pp. 97–107.

[330] D. Klein et al. “Mapping cells through time and space with moscot”. en. May 2023.

[331] L. Zaveri and J. Dhawan. “Cycling to Meet Fate: Connecting Pluripotency to the Cell Cycle”. en. *Front Cell Dev Biol* 6 (June 2018), p. 57.

[332] D. Schwabe et al. “The transcriptome dynamics of single cells during the cell cycle”. en. *Mol. Syst. Biol.* 16.11 (Nov. 2020), e9946.

[333] J. Boonstra and J. A. Post. “Molecular events associated with reactive oxygen species and cell cycle progression in mammalian cells”. en. *Gene* 337 (Aug. 2004), pp. 1–13.

[334] M. Ebisuya and J. Briscoe. “What does time mean in development?” en. *Development* 145.12 (June 2018).

[335] T. Rayon. “Cell time: How cells control developmental timetables”. en. *Sci Adv* 9.10 (Mar. 2023), eadh1849.

[336] M. Diaz-Cuadros et al. “Metabolic regulation of species-specific developmental rates”. en. *Nature* 613.7944 (Jan. 2023), pp. 550–557.

[337] K. Swovick et al. “Interspecies Differences in Proteome Turnover Kinetics Are Correlated With Life Spans and Energetic Demands”. en. *Mol. Cell. Proteomics* 20 (Jan. 2021), p. 100041.

[338] R. Iwata et al. “Mitochondria metabolism sets the species-specific tempo of neuronal development”. en. *Science* 379.6632 (Feb. 2023), eabn4705.

[339] S. M. Chambers et al. “Combined small-molecule inhibition accelerates developmental timing and converts human pluripotent stem cells into nociceptors”. en. *Nat. Biotechnol.* 30.7 (July 2012), pp. 715–720.

[340] N. Sasaki et al. “Chemical inhibition of sulfation accelerates neural differentiation of mouse embryonic stem cells and human induced pluripotent stem cells”. en. *Biochem. Biophys. Res. Commun.* 401.3 (Oct. 2010), pp. 480–486.

[341] M. W. Amoroso et al. “Accelerated high-yield generation of limb-innervating motor neurons from human stem cells”. en. *J. Neurosci.* 33.2 (Jan. 2013), pp. 574–586.

[342] A. M. B. Tadeu et al. “Transcriptional profiling of ectoderm specification to keratinocyte fate in human embryonic stem cells”. en. *PLoS One* 10.4 (Apr. 2015), e0122493.

[343] L. Sikkema et al. “An integrated cell atlas of the lung in health and disease”. en. *Nat. Med.* 29.6 (June 2023), pp. 1563–1577.

[344] B. Van de Sande et al. “Applications of single-cell RNA sequencing in drug discovery and development”. en. *Nat. Rev. Drug Discov.* 22.6 (June 2023), pp. 496–520.

[345] I. Yofe, R. Dahan, and I. Amit. “Single-cell genomic approaches for developing the next generation of immunotherapies”. en. *Nat. Med.* 26.2 (Feb. 2020), pp. 171–177.

[346] M. Polychronidou et al. “Single-cell biology: what does the future hold?” en. *Mol. Syst. Biol.* 19.7 (July 2023), e11799.

[347] G. M. Allen and W. A. Lim. “Rethinking cancer targeting strategies in the era of smart cell therapeutics”. en. *Nat. Rev. Cancer* 22.12 (Dec. 2022), pp. 693–702.

[348] R. G. Majzner and C. L. Mackall. “Tumor Antigen Escape from CAR T-cell Therapy”. en. *Cancer Discov.* 8.10 (Oct. 2018), pp. 1219–1226.

[349] C. A. Lareau, K. R. Parker, and A. T. Satpathy. “Charting the tumor antigen maps drawn by single-cell genomics”. en. *Cancer Cell* 39.12 (Dec. 2021), pp. 1553–1557.

[350] G. Palla et al. “Squidpy: a scalable framework for spatial omics analysis”. en. *Nat. Methods* 19.2 (Feb. 2022), pp. 171–178.

[351] A. Daei Sorkhabi et al. “The current landscape of CAR T-cell therapy for solid tumors: Mechanisms, research progress, challenges, and counterstrategies”. en. *Front. Immunol.* 14 (Mar. 2023), p. 1113882.

[352] Z.-Z. Zhang et al. “Improving the ability of CAR-T cells to hit solid tumors: Challenges and strategies”. en. *Pharmacol. Res.* 175 (Jan. 2022), p. 106036.

[353] P. Jia et al. “Deep generative neural network for accurate drug response imputation”. en. *Nat. Commun.* 12.1 (Mar. 2021), p. 1740.

[354] Y. Chang et al. “Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature”. en. *Sci. Rep.* 8.1 (June 2018), p. 8857.

[355] K. Singhal et al. “Large language models encode clinical knowledge”. en. *Nature* (July 2023).

[356] K. M. Cappell et al. “Long-Term Follow-Up of Anti-CD19 Chimeric Antigen Receptor T-Cell Therapy”. en. *J. Clin. Oncol.* 38.32 (Nov. 2020), pp. 3805–3815.

[357] K. M. Cappell and J. N. Kochenderfer. “Long-term outcomes following CAR T cell therapy: what we know so far”. en. *Nat. Rev. Clin. Oncol.* (Apr. 2023), pp. 1–13.

[358] T. Haslauer et al. “CAR T-Cell Therapy in Hematological Malignancies”. en. *Int. J. Mol. Sci.* 22.16 (Aug. 2021).

[359] C. K. Stein-Thoeringer et al. “A non-antibiotic-disrupted gut microbiome is associated with clinical responses to CD19-CAR-T cell cancer immunotherapy”. en. *Nat. Med.* 29.4 (Apr. 2023), pp. 906–916.

[360] J. S. Abramson et al. “Lisocabtagene maraleucel for patients with relapsed or refractory large B-cell lymphomas (TRANSCEND NHL 001): a multicentre seamless design study”. en. *Lancet* 396.10254 (Sept. 2020), pp. 839–852.

[361] G. López-Cantillo et al. “CAR-T Cell Performance: How to Improve Their Persistence?” en. *Front. Immunol.* 13 (Apr. 2022), p. 878209.

[362] L. Gattinoni, C. A. Klebanoff, and N. P. Restifo. “Paths to stemness: building the ultimate antitumour T cell”. en. *Nat. Rev. Cancer* 12.10 (Oct. 2012), pp. 671–684.

[363] J. A. Fraietta et al. “Determinants of response and resistance to CD19 chimeric antigen receptor (CAR) T cell therapy of chronic lymphocytic leukemia”. en. *Nat. Med.* 24.5 (May 2018), pp. 563–571.

[364] S. L. Maude et al. “CD19-targeted chimeric antigen receptor T-cell therapy for acute lymphoblastic leukemia”. en. *Blood* 125.26 (June 2015), pp. 4017–4023.

[365] B. Xie et al. “Current Status and Perspectives of Dual-Targeting Chimeric Antigen Receptor T-Cell Therapy for the Treatment of Hematological Malignancies”. en. *Cancers* 14.13 (June 2022).

[366] S. Cordoba et al. “CAR T cells with dual targeting of CD19 and CD22 in pediatric and young adult patients with relapsed or refractory B cell acute lymphoblastic leukemia: a phase 1 trial”. en. *Nat. Med.* 27.10 (Oct. 2021), pp. 1797–1805.

[367] H. Shalabi et al. “CD19/22 CAR T cells in children and young adults with B-ALL: phase 1 results and development of a novel bicistronic CAR”. en. *Blood* 140.5 (Aug. 2022), pp. 451–463.

[368] S. Ahmadi et al. “The landscape of receptor-mediated precision cancer combination therapy via a single-cell perspective”. en. *Nat. Commun.* 13.1 (Mar. 2022), p. 1613.

[369] M. Lotfollahi et al. “Predicting cellular responses to complex perturbations in high-throughput screens”. en. *Mol. Syst. Biol.* 19.6 (June 2023), e11517.

[370] J. G. Rurik et al. “CAR T cells produced in vivo to treat cardiac injury”. en. *Science* 375.6576 (Jan. 2022), pp. 91–96.

[371] C. M. Seitz et al. “Novel adapter CAR-T cell technology for precisely controllable multiplex cancer targeting”. en. *Oncoimmunology* 10.1 (Dec. 2021), p. 2003532.

[372] S. Zhou et al. “The landscape of bispecific T cell engager in cancer treatment”. en. *Biomark Res* 9.1 (May 2021), p. 38.

[373] A. Singh, S. Dees, and I. S. Grewal. “Overcoming the challenges associated with CD3+ T-cell redirection in cancer”. en. *Br. J. Cancer* 124.6 (Mar. 2021), pp. 1037–1048.

[374] K. Strebhardt and A. Ullrich. “Paul Ehrlich’s magic bullet concept: 100 years of progress”. en. *Nat. Rev. Cancer* 8.6 (June 2008), pp. 473–480.