Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Using RNA barcoding and sequencing to study cellular differentiation on a single-cell and population level

Gunnar Kuut
aus
Tallinn, Estland

2021

Ekrlärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Veit Hornung betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

Gunnar Kuut

München, den 23.04.2021

Dissertation eingereicht am: 26.04.2021

1. Gutachter: Prof. Dr. Veit Hornung
2. Gutachter: Prof. Dr. Lucas Jae

Mündliche Prüfung am : 28.06.2021

Table of Contents

1 Introduction

1.1 Genes, gene expression, and transcriptomics

A cell is the basic structural and biological unit of life. Its structure, function, and how it interacts with its environment are dictated by its genetic material, which consists of deoxyribonucleic acid (DNA). The smallest functional unit of DNA is defined as a gene[1]. This definition only included protein-coding regions, a DNA segment transcribed into messenger RNA (mRNA) and translated into protein via the ribosome. Today, the characterization of a gene is more complex and encompassing. It includes all transcribed RNA with any function or observable trait (phenotype). Examples of non-protein-coding (non-coding) RNA include regulatory RNA (lncRNA, miRNA), structural RNA (ribozymes and ribosomal RNA), and transport RNA (tRNA). The sum-total of all the RNA in a cell or a population of cells is called the transcriptome[2–4].

The central dogma in biology defines the directionality of information flow, from DNA to RNA to protein[5]. Multicellular organisms (where all cells contain the same set of DNA) therefore require complex regulation at each step of this central dogma to produce a wide variety of cells with distinct functions. This varied regulation begins with RNA transcription, also termed gene expression. Gene expression is regulated on many levels, starting with how the cell packages its DNA in the nucleus (eukaryotic cells), also called chromatin. Gene expression can also be affected by modifications to the DNA (methylation) and by other expressed genes (transcription factors) as well as specific sequences of DNA preceding and sometimes following a gene (enhancers)[6]. In order to understand how cells function and differ from each other, there is no better place to start than to look for a way to measure gene expression.

Several crucial discoveries and technological advances were necessary to make gene expression (transcriptomics) analysis a routine laboratory method (Fig.1.1 A). Arguably one of the most important breakthroughs for the field of RNA research happened in 1970 when two scientists independently reported the discovery of an enzyme that can synthesize DNA from RNA – a reverse transcriptase[7]. Its name was derived from its function and it opened up a whole avenue of research by allowing RNA to be transcribed into complementary DNA (cDNA), which is more stable than the original RNA and can be used with DNA-specific methods such as restriction digest. If synthesized from eukaryotic mRNA, cDNA will also not contain any introns, allowing the study and use of only the protein-coding sequences of DNA. The first laboratory method to measure or quantify RNA

in any meaningful way in regards to gene expression was developed by Alwine et al. In 1977, they showed that it is possible to separate the total RNA of a sample by gel electrophoresis then transfer it to a specially modified paper, hybridized the sample to a radioactively tagged cDNA or RNA probe and develop it. This method allows the detection of a specific RNA in a given sample[8]. Initially, as mentioned, all probes were radioactively labeled, but new methods became available over time, such as chemiluminescent, fluorescent, or even antibody-tagged (immuno-Northern blots) probes[9,10]. This novel method allowed accurate detection and quantification of any RNA with a known sequence. Northern blotting is still in use today due to the accuracy and sensitivity, and ability to directly detect RNA (Fig.1.1 B). The limitations of this method are its low throughput and the requirement to know the sequence of the RNA transcript of interest. Additionally, quantification is relative to either a control RNA or another sample which can be error-prone.

The next giant leap in our ability to detect and measure RNA transcripts came from the development of one of molecular biology's most common technique – the Polymerase Chain Reaction or PCR.

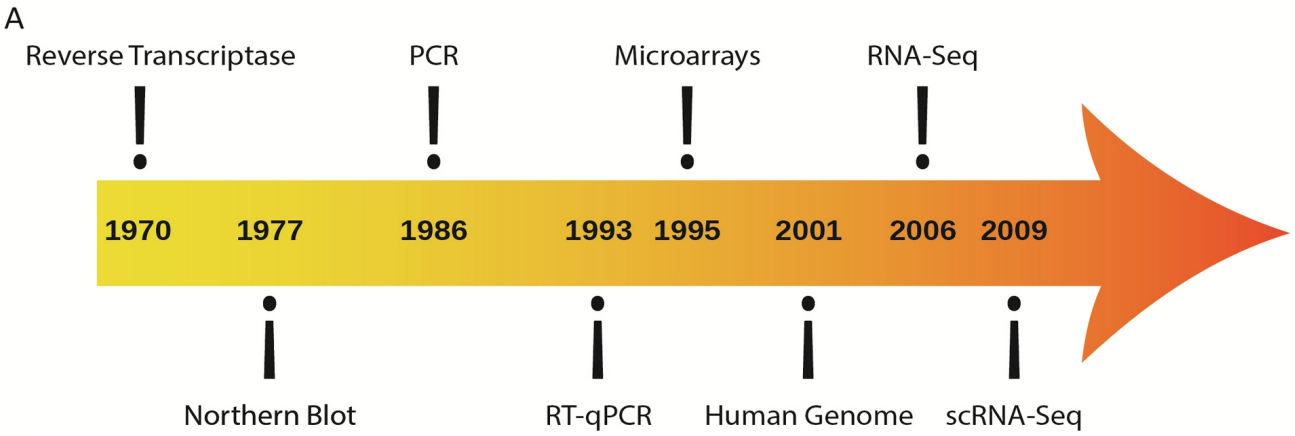

Figure 1.1 (**A**) Major developments in gene expression analysis

In 1986, Mullis et al. developed a method that allowed them to amplify any known DNA sequence exponentially. This ground-breaking technique was initially quite laborious, involving tedious sample handling steps. After each amplification step (cycle) in the protocol, it was necessary to add a fresh enzyme to the reaction as the old one was being denatured due to high temperatures[11]. Only after a thermostable version of DNA polymerase from *T. aquaticus* was combined with this novel technique did PCR became routine, high-throughput, and commonplace in all laboratories[12,13]. The next evolution of

PCR was pioneered by Higuchi et al. through utilizing double-stranded DNA (dsDNA) dye and video camera technology. This setup allowed the real-time quantification of PCR reaction[14], also called qPCR. Combination of reverse transcription (RT), PCR, and real-time monitoring (RT-qPCR) allowed for relative quantification of any mRNA species with a known sequence. Later, as the technology developed, it became possible to measure multiple genes in one reaction with fluorescently labeled probes. Still, RT-qPCR had some shortcomings. It is a low-throughput method and limited to just a handful of genes at a time, and as the Northern blot method, it cannot be used to detect novel transcripts. These techniques are invaluable in targeted studies with known genes yet leave the rest of the vast transcriptome unstudied.

The initial estimate for the number of genes in the human genome ranged from 50,000 – 90,000[15]. Since the first draft of the human genome in 2001, the more accurate estimates have put that number to be around 20,000[16]. As of February 2021, Ensembl notes 20,437 protein-coding gene transcripts, 16,900 lncRNAs, 4,867 small non-coding RNAs, and an additional 2221 miscellaneous RNAs[17].

Need drives innovation; a higher throughput method was required to analyze the whole transcriptome; the solution came from using the hybridization method in an arrayed setup. The basic principle is as follows: short DNA sequences (probes) of the genes of interest were attached to a glass or silicon chip in a grid-like fashion. RNA or cDNA of the sample of interest is applied and visualized with dsDNA binding dye. More copies of one gene equate to a brighter signal. In 1995 the first study was published, mentioning “microarray” in its title[18], and it measured 45 *A. thaliana* genes simultaneously[18]. Two years later, Lashkari et al. developed a microarray that contained probes for all of the open reading frames (ORFs) in yeast (2,479 in total)[19]. The technology was subsequently commercialized, and whole-genome transcriptome studies in humans and model organisms sky-rocketed (Fig.1.1 B). Many different microarrays were made to profile all known genes, SNPs, and alternative splicing of mRNA[20]. Microarrays enabled high throughput and standardized workflows that are cost-effective and remain competitive to date. Despite all the possibilities that microarrays offer, they are not without their drawbacks. They only allow measuring known transcripts (known sequences). As such, it is impossible to discover completely novel transcripts or unexpected polymorphisms, alternative splicing, or editing. Also, the sensitivity of microarrays is significantly lower compared to RT-qPCR, which is especially relevant when analyzing genes with low levels of expression[21].

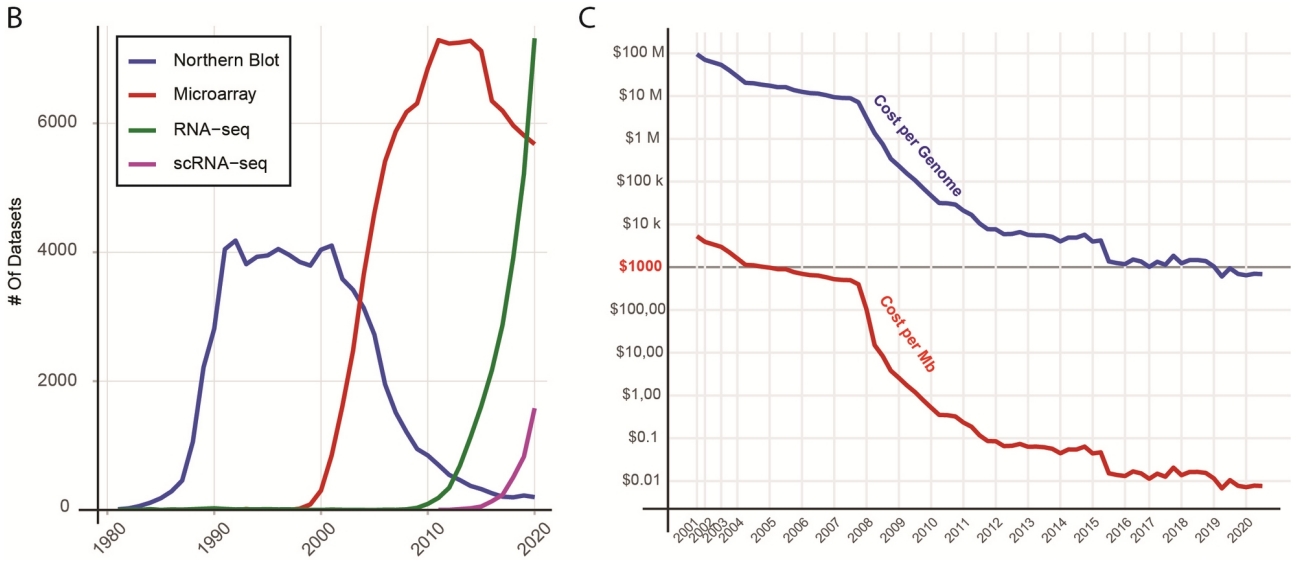

Figure 1.1 (**B**) Popularity of methods. The total number of datasets found per year in PubMed.gov. (**C**) Sequencing cost per one human genome in blue and per one megabase (Mb) in red[22].

A solution to all of these problems came from another combination of technologies. Combining reverse transcription with PCR and DNA sequencing, it became possible to use any DNA sequencing platform to get accurate and sensitive gene expression information. Now, only the cost was an issue, since in RNA-sequencing (RNA-seq), unlike DNA-sequencing, it is not only the sequences themselves but also the abundances of those sequences that are of interest.

This problem was solved thanks to another mega-project. In the early to mid-2000s, driven by the Human Genome Project, the next generation of sequencing technologies was coming of age[16]. Until then, RNA-seq was limited by the enormous cost of sequencing in general (Fig.1.1 C). Next-generation sequencing (NGS), also called 2nd generation, uses microfluidic chips to run and measure millions of simultaneous reactions. Also called sequencing by synthesis, it was developed initially by Solexa and later acquired by Illumina. NGS offers cheaper and much higher throughput sequencing than conventional Sanger sequencing[23].

The first RNA-seq paper was published in 2006, analyzing the prostate cancer cell line LNCaP with a combination of expressed sequence tag (EST) and sequencing by synthesis approach[24]. By 2008, sequencing price had dropped 100-fold since 2001 (Fig.1.1 C), and Illumina offered more than a gigabase of data in one sequencing run. These developments positioned RNA-seq to become a widely used technique to this day. RNA-seq offers far more accurate and sensitive transcriptome profiling than any other method, and it overcomes all the main limitations of Northern blots, qPCRs, and microarrays since it can detect completely

novel transcripts, alternative splicing, SNPs, polymorphisms, and RNA editing[4]. Despite the falling cost of sequencing and all the advantages, it took more than a decade for RNA-seq to overtake microarrays in the number of datasets produced per year (Fig.1.1 B). Currently, the most significant issues with RNA-seq are all associated with the method of sequencing. NGS offers a very high throughput and fidelity DNA sequencing, but it is limited to short read lengths (50-300bp). Short read lengths make it necessary to sequence more deeply to later combine the reads into one continuous transcript or contig during the data analysis, increasing the costs. Also, because of short read lengths, it is challenging to align repetitive sequences, although that specific drawback rarely affects RNA-seq[4,25].

Another problem with current DNA sequencing methods is the need for amplification before sequencing, leading to several additional difficulties. First, no direct RNA sequencing is possible, as PCR works only on DNA. Second, the amplification of different sequences is not uniform and, therefore, can cause amplification bias. Here, high numbers of short reads with the same sequence could mean either a very abundant transcript or an uneven amplification[4]. Third, increasing the input material (number of cells in a given sample) reduces the number of cycles needed for a PCR amplification but comes at the cost of cellular resolution. RNA-seq is often performed on a collection of cells either from a tissue or cell culture, and as a read-out, we get an average expression level of all the genes in the sample. To achieve a single-cell resolution in RNA-seq, more sensitive sequencers and innovations in library preparation were needed.

1.2 Single-cell RNA sequencing

The NGS platforms have seen significant improvement since their inception. The sequencing accuracy, sensitivity, and throughput, along with molecular biology methods, have all gotten to the point that it has been possible to analyze whole transcriptomes of single cells for over a decade now. The first single-cell dataset was published by Tang et al. in 2009 when they sequenced four mouse blastomeres, the cells that form after a fertilized egg cell divides. The choice of these was deliberate as they are one of the biggest cells in a mammalian body and contain a lot of RNA, making working with these cells much easier. Within five years, the average number of cells captured and sequenced in a single-cell study grew to hundreds, and after a decade, it was in the hundreds of thousands[26] (Fig.1.2 A). Every year more and more single-cell RNA-seq (scRNA-seq) datasets are published. At this rate, in another five years, it is likely to overtake bulk RNA-seq as the most common transcriptomics tool (Fig.1.1 B).

Despite the myriad of different scRNA-seq protocols, the main parts of these methods fall into two categories. The first and the oldest methods involve some way of separating and choosing cells (FACS, laser dissection, etc.) and then running the library preparation reactions in microtiter plates or other small vessels. The second main group of methods is a lot more automated and requires less hands-on time. It utilizes microfluidic technologies and specialized microchips to separate the reactions from each other, while they are usually still carried out on the same microfluidic chip.

An offshoot of that technology uses microfluidic chips to separate single cells into small lipid droplets containing the required reagents and primers for the first steps of library preparation. After lysis and RT, part of which is also cDNA barcoding in the droplet, all the samples are mixed and handled together to save on reagents and time. The first such protocol was aptly named Drop-seq, and it offers one of the highest throughputs of all scRNA-seq methods[28].

This need to make scRNA-seq more affordable also brings us to the next big division in the library preparation methods; full-length mRNA vs. tag-based sequencing. All high throughput single-cell methods are tag-based end-counting (usually 3´ end of mRNA molecules) methods that generate partial cDNA transcripts. Sequencing these tags results in a count table of genes, where the numbers signify how many times each tag (gene) was captured per cell[29–31]. This method, also called digital gene expression (DGE), is an excellent and cost-effective way to capture and sequence mRNA from vast numbers of cells simultaneously. The first method to do this was single-cell tagged reverse transcription (STRT-seq). They prepared single-cell libraries of 96 cells and at the same time also only captured the 5´ ends of mRNAs for sequencing[32]. However, by focusing only on a portion of any given transcript, it is impossible to derive information about other gene expression mechanics, including alternative splicing or RNA editing.

To overcome this limitation, one would need to produce full-length cDNA transcripts of mRNAs. The first protocol to achieve this for single cells was based on the Switching Mechanism at the 5′ end of RNA Template or SMART method[33–35]. Full-length methods have much-reduced throughput and increased sequencing cost. However, some collaborative efforts have produced full-length cDNA libraries of single cells and sequenced them in comparable numbers to tag-based methods.

One of these was the Tabula Muris Consortium. In 2018 they published two single-cell datasets in parallel; one full-length, containing 44,949 single cells and one tag-based, containing 55,656 single cells produced using a commercial solution from 10X Genomics[36]. Still, the scope of projects like these is beyond the majority of standard laboratories, as costs are prohibitively high.

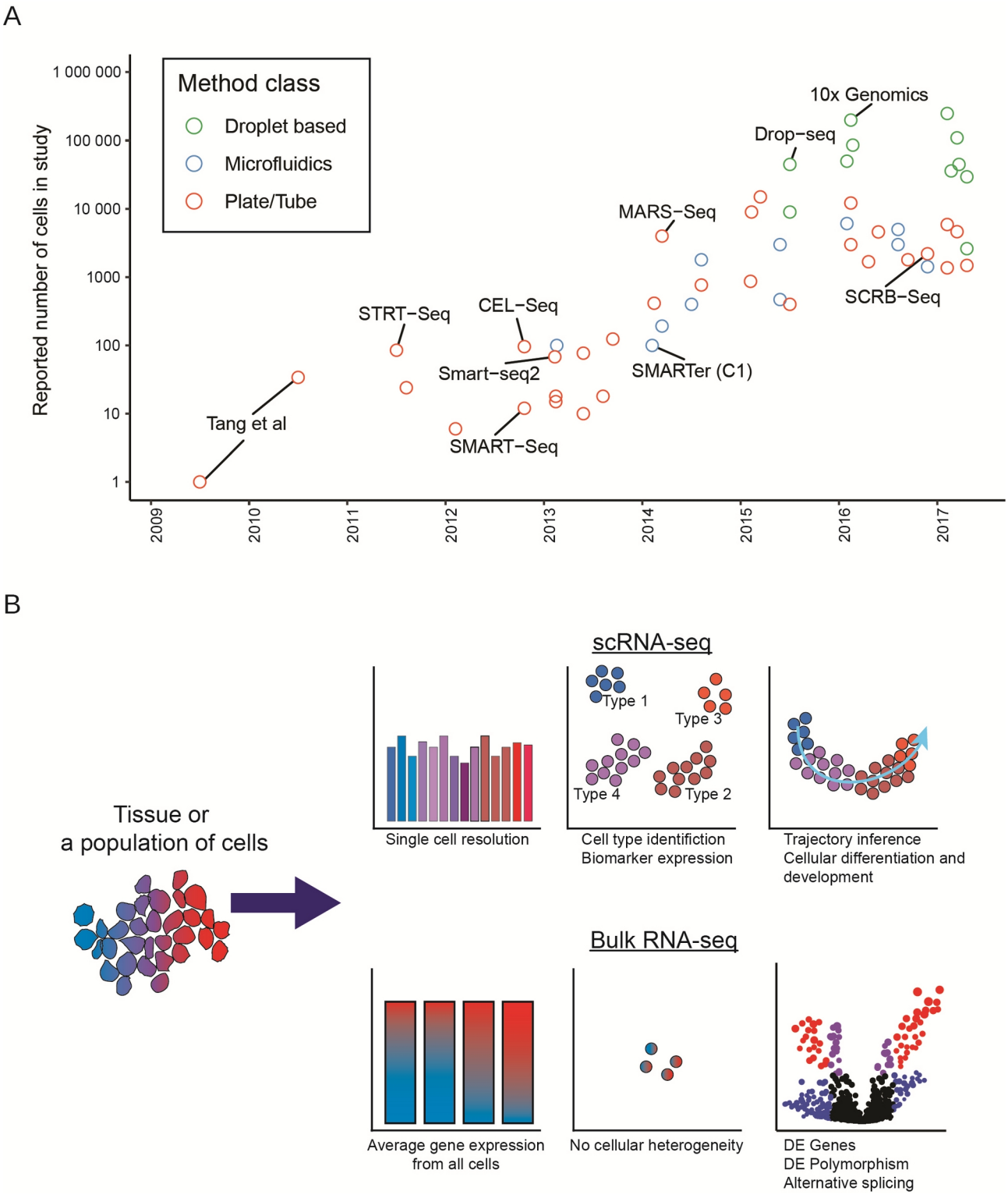

Figure 1.2 (**A**) Reported numbers of single cells in publications ordered by date. Data points are colored by the type of library preparation. Major milestones are indicated with labels. Figure adapted from[27]. (**B**) Single-cell RNA-seq vs. bulk RNA-seq. scRNA-seq can reveal cellular heterogeneity in a population of cells. It allows the detection of novel cell types and cell development inference.

One more major difference to note is the way cDNA is amplified in a given protocol. Again, two options exist; the amplification can be either linear or exponential. The latter is always based on PCR, but the former uses in vitro transcription or IVT to produce enough RNA transcripts based of (barcoded) cDNA that can be reverse transcribed yet again and sequenced like in any other method. Although IVT protocols for scRNA-seq were among the first (CEL-seq), they were quickly overtaken by PCR-based methods. The main reason to use IVT is to avoid amplification bias[37], but the bulk of that problem has been solved by incorporating Unique Molecular Identifiers (UMIs) during the RT step. There will be more on that in chapter 4.1.

Interestingly only microfluidics platforms have achieved commercial success. The first one on the market was Fluidigm's C1 system, followed by 10X Genomics Chromium, only two years later. The C1 system offers full-length coverage and a choice between different library preparation methods STRT-seq or SMART-seq, but by its design, it is limited to ~1000 cells per run[38]. 10X Chromium is directly based on Drop-seq, and like Drop-seq, it has one of the highest throughputs of all the platforms. 10X has also managed to bring a certain standardization and quality control to the market. It is exceedingly simple to use, and it offers a combination of multi-omics approaches such as cell surface protein capture and constant improvements to keep this platform competitive[39,40].

Regardless of how the library preparation is performed, all of them are still sequenced on second-generation (NGS) platforms, which dictates specific requirements for the libraries (see chapter 1.1). These will inevitably change as technologies develop. Already since 2019, single-cell datasets have been published[41–43] that used what is called the third-generation sequencing platforms. The main difference between the NGS and the third-generation is the read length. The two leading platforms: Single-Molecule Real-Time sequencing SMRT-sequencing by PacBio and Nanopore sequencing by ONT, offer reads that are kilobases long, compared to 200-300 bases offered by Illumina. PacBio's approach involves modified polymerase and fluorescently tagged dNTPs. DNA to be sequenced is circularized and attached to the polymerase, which proceeds to synthesize a new complementary strand. With each addition of a dNTP, a light signal is given off, and the bases are read in real-time. To achieve higher accuracy, the same sequence of circularized DNA is "read-through" multiple times (up to 20X)[44] (Fig. 1.2 C).

ONT takes a slightly different approach to PacBio. A nano-sized pore is used in conjunction with a motor protein that feeds a single-stranded DNA (or RNA) through it under an electric field. Unlike any other sequencing method, Nanopore sequencing can be used for direct RNA sequencing and direct genomic DNA sequencing that can reveal

epigenetic modifications[41] (Fig. 1.2 C).

Both of these platforms are highly modular, and by repeating the main elements (pores or polymerase chambers), it is possible to increase the throughput, although currently, they are still only now catching up to Illumina. Both technologies had similar troubles in the beginning, the speed of polymerase for SMRT-seq and the speed of motor protein for ONT was extremely high, and it was necessary to figure out ways to slow them down to get a more accurate read-out. Currently, the third-generation sequencers are all troubled by high error rates[45,46]. One way to overcome this problem is to combine the second and third-generation sequencing platforms.

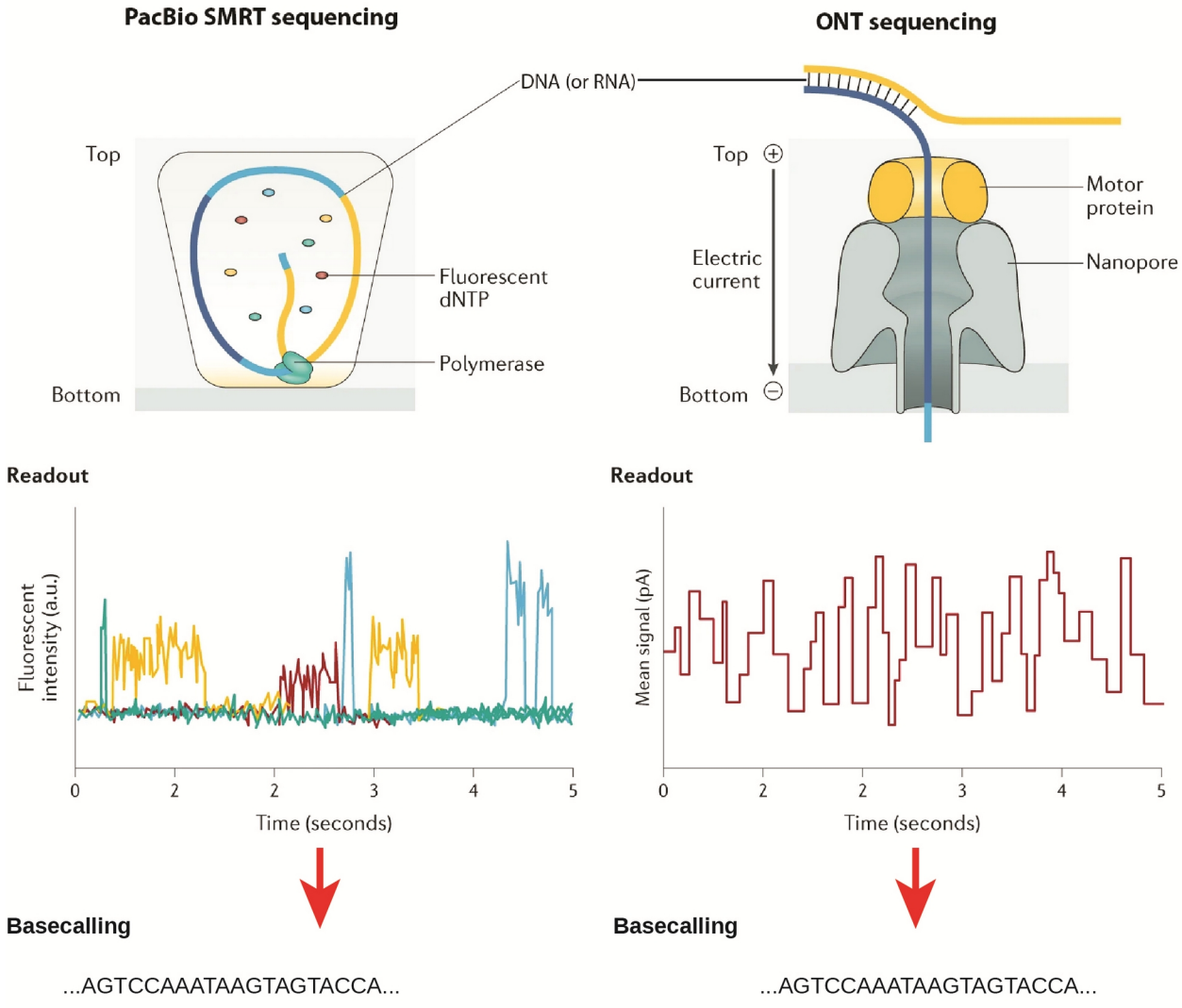

Figure 1.2 (**C**) Overview of 3rd generation sequencing platforms. SMRT-sequencing from PacBio and Nanopore sequencing from Oxford Nanopore Technologies. Figure adapted from[47].

A protocol called ScNaUmi-seq does just that by using the Chromium from 10X Genomics with NGS for high accuracy barcode reads and nanopore from ONT for full-length reads[41]. One drawback of ScNaUmi-seq is still the sequencing cost, as it effectively doubles it by sequencing the same libraries on two different platforms.

In principle, it is possible to apply scRNA-seq in place of regular bulk RNA-seq. Still, because of certain limitations and differences in scRNA-seq methods, it is more common to use it for experiments where high sensitivity is not required. Although single-cell technologies have improved immensely, the low mRNA capture rate remains the main limitation. Droplet-based platforms capture only about 10-30% of the total mRNA in a cell. This number is slightly higher for plate-based methods (up to 60%)[48,49].

The benefits of scRNA-seq are manyfold. It can detect gene expression changes on a single-cell level, reveal the cellular heterogeneity of tissues, detect novel and rare subtypes of cells, shed light on cellular differentiation and identify other population dynamics like cell cycle[50] (Fig.1.2 B).

Further developments, especially those related to nanopore sequencing, seem very promising – not only would high-accuracy Nanopore sequencing offer cheap and high throughput full-length RNA-seq, but it also could be used for direct RNA sequencing and real-time mapping of the reads[51,52]. The implications of such technological achievements are profound and exciting.

1.3 Cellular differentiation and scRNA-seq

All somatic cells in a multicellular organism contain essentially the same DNA, yet as the organism develops, gene expression changes dictate how the cell will differentiate and acquire new functions[53]. Most often, this process is described as a pebble rolling down the hill. At the beginning of the journey, a cell has many diverging paths ahead, but as it progresses, its choices become more restricted and the path more defined, as do its functions and cell type. The general shape of this landscape is defined by gene expression[54]. Cells differ from each other morphologically and functionally, and both differences are primarily determined by the proteins they produce. Therefore, the fundamental difference comes down to gene regulation and expression. Genes are regulated at many levels: transcription, mRNA splicing, mRNA editing, mRNA export from

the nucleus, translation, post-translational modification, but the most apparent phenotypic regulation occurs at the level of transcription. This is also the most energy-efficient way for a cell to regulate its function, which is also favored by evolution[6]. If a gene is not required, it is best to avoid production rather than recycle an unused product.

Gene expression is affected by both internal programs as well as external stimuli. Cells have to find a balance between reacting to these perturbations whilst maintaining their course of development[55,56]. With the advent of scRNA-seq, it has become possible to follow these changes and responses on a single-cell level. The opportunities provided by the scRNA-seq are unprecedented. The ability to sequence thousands of cells as they undergo developmental processes[57,58] and detect changes in transcriptional programs in response to stimuli or treatments[59] will continue to impact medicine and biotechnology significantly.

To better deal with this data, a whole branch of analytical methods has been developed since 2014, specifically for analyzing the dynamic processes in single-cell data[60]. The so-called pseudotime analysis, or more generally – trajectory inference (TI), is a set of computational techniques that model how cells go through the cell cycle, differentiation, and activation. Within a few years, these methods have become one of the most popular tools for single-cell-omics[61].

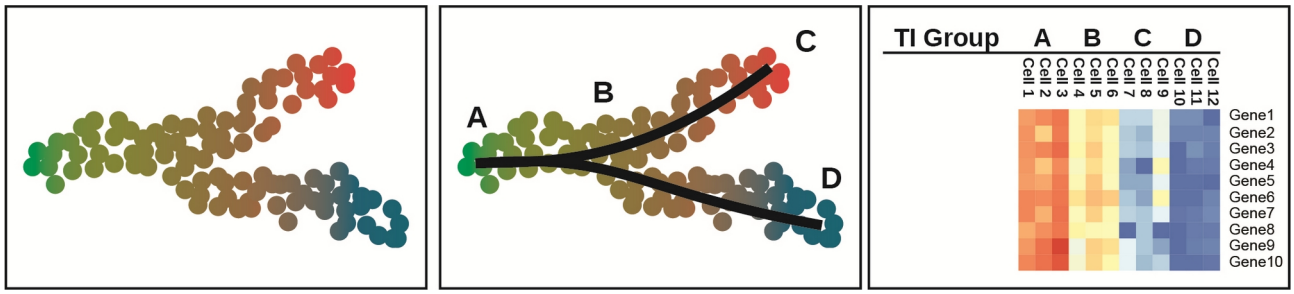

Figure 1.3 Main steps in pseudotime analysis workflow. In the first panel, the algorithm positions cells based on their similarity to other cells. Middle panel, the algorithm builds a trajectory to capture the dynamic processes in the dataset. Last panel, finding differentially expressed genes along the identified trajectories.

The main idea of any TI analysis is to order the cells along a trajectory based on their gene expression similarity. This trajectory can be linear, bifurcating, tree-shaped, or even circular. TI analysis creates a (pseudo)temporal order in the data through which it is possible to follow gene expression patterns as they change along the trajectory. Ultimately, TI aims to rebuild the sequence that cells took in the developmental process, reveal the transcriptional changes during that process and describe the potential cell fates (subsets)[60] (Fig.1.3).

1.4 Human T cell biology

T lymphocytes or T cells are an integral part of the vertebrate adaptive immune system. They help to fight infections and tumors, but they can also respond to allergens or self-antigens. T cells develop from hematopoietic stem cells that move from bone marrow to thymus, an organ where they were first discovered and received their name from. In the thymus, T cells undergo further maturation and selection. As one of the first steps of maturation, T cells undergo T cell receptor (TCR) rearrangement. Following that, the immature T cells develop into double-positive (DP) cells for both TCR co-receptors (CD4 and CD8), after which the selection process starts. How exactly the T cell selection occurs in humans is still up for debate, but in mice, the DP cells undergo two selection rounds. First, only the cells that interact with major histocompatibility complex class I or II (MHC-I or MHC-II) on an antigen-presenting cell (APC) will be selected for and receive a pro-survival signal. The second round is the negative selection to remove the cells that react too strongly to self-antigens presented by dendritic cells (DC). After the selection process, either one or the other TCR co-receptor will be down-regulated, and naive $CD4^+$ or $CD8^+$ T cells are ready to leave the thymus[62].

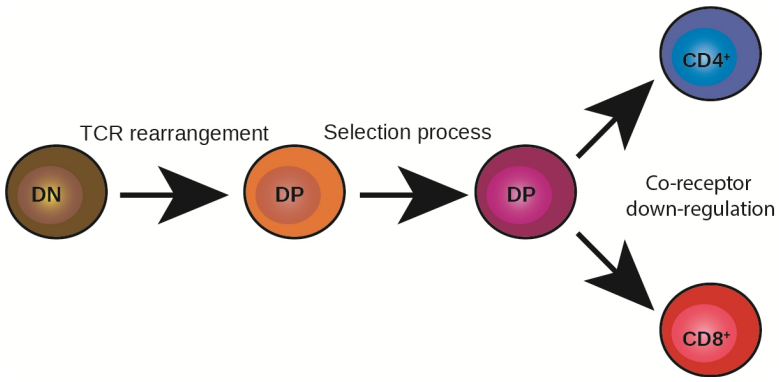

Figure 1.4 T cell development in the thymus. Figure adapted from[63].

The main difference between $CD4^+$ and $CD8^+$ T cells is how they function. $CD8^+$ T cells, also called cytotoxic T cells, destroy pathogenic cells (virus or bacteria-infected cells and tumorigenic cells) by first identifying them through antigens presented to them on MHC-I molecules[6]. After identifying target cells, $CD8^+$ T cells release perforin, a protein capable of forming pores in the target cell membrane, followed by the release of granzymes that enter the affected cell through the newly formed pores. In the cell, the granzymes trigger a

signaling cascade that eventually leads to cell death by apoptosis (programmed cell death)[64]. On the other hand, $CD4^+$ T cells or helper T cells do not directly kill pathogenic cells; instead, they help to direct the immunological processes that fight infections[65].

In humans, unlike mice, mature T cells are present in the fetus from the early months of pregnancy. Additionally, the percentage of naive vs. memory T cells changes throughout the human lifetime. From birth to early adulthood, the majority of T cells are new and naive. During these formative years, the organism is constantly encountering new antigens. Memory T cells start to accumulate from childhood, and the numbers level off when a person enters adulthood. This new status quo is kept until the late stages of life when immuno-senescence sets in, reducing T cell functionality and increasing inflammation[66]. Recent studies have revealed an intriguing phenomenon in supercentenarians (people who survive to a very old age of 110+ years in good health and cognitive state). They have elevated numbers of $CD4^+$ T cells compared to a control group of 50- to 80-year-old people. Those extra $CD4^+$ T cells express very high levels of cytotoxic genes and are primarily (up to 70%) composed of only ten clonotypes, and just one clonotype can make up to 35% of the entire T cell population[67].

1.4.1 $CD4^+$ T cell activation and differentiation

Unlike $CD8^+$ T cells that only require clonal expansion to accomplish their function, $CD4^+$ T cells need an additional round of differentiation into specific “effector” types to fulfill their role. To achieve this, a naive $CD4^+$ T cell first needs to come in contact with a professional antigen-presenting cell that is exhibiting an antigen on its MHC-II. TCR, together with its co-receptor CD4, binds the MHC-II as the first of three “classical” signals[68]. The second is the costimulatory signal. CD28, one of the best-characterized receptors for co-stimulation on T cells, binds CD80 and CD86 on APCs. Receiving the first two signals is enough for the naive $CD4^+$ T to cell leave quiescence and start the clonal expansion up to millions of T cells that all recognize the same antigen. To limit and control this expansion, the proliferating T cells themselves transiently produce CTLA-4, a molecule that binds to CD28, and instead of the activation signal, transmits an inhibitory signal[69]. To avoid inappropriate signaling, APCs only express CD28 ligands a short time after coming in contact with a pathogen. When TCR binds the MHC-II without a costimulatory signal, the

activation is instead terminated[70,71].

Lastly, the third signal is transmitted to the activated cells through the immediate microenvironment. Cytokines released by APCs will determine the cell fate of T helper cells[72] (Fig. 1.41). Another factor that might play a role in T cell fate decisions is the affinity by which antigens bind to the TCR. For example, Constant et al. showed that low-affinity vs. high-affinity binding of TCR resulted in different T helper subsets[73]. On the transcriptional level, the cytokine signaling activates a family of STAT proteins (Signal Transducer and Activator of Transcription) that, together with lineage-specific master transcription factors, regulate the differentiation into memory T cells (T effector cells)[74–77] (Table 4.1.2).

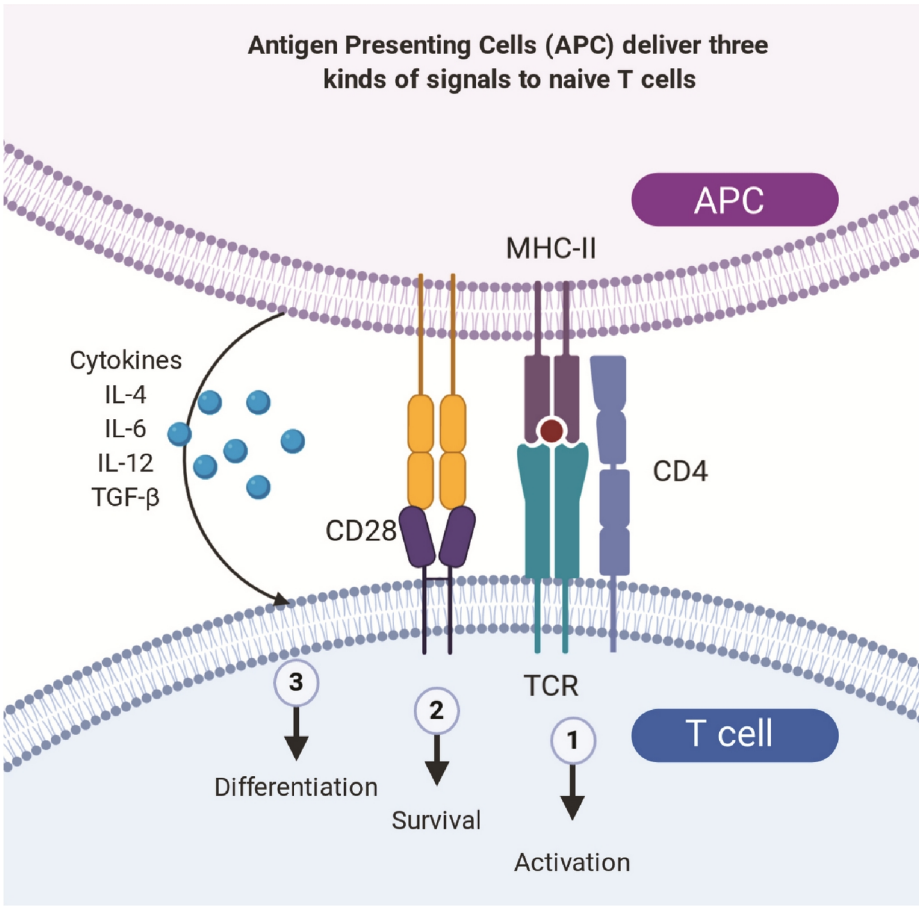

Figure 1.4.1 T cell activation. APC has an antigen on the MHC-II (the red dot). TCR recognizes the antigen. The CD28 molecule recognizes the costimulatory signal. The release of cytokines transmits the third polarizing signal into the microenvironment. Created with BioRender.com. Modified from[64].

Another important part of the T cell activation is an apoptotic process called activation- induced cell death (AICD). AICD has multiple functions. Premature T cells

undergo AICD during the normal developmental selection process[78]. AICD is also part of the T cell response and helps to limit the expansion of T cells after the activation and possibly plays a role in the differentiation of T helper subsets[79,80]. AICD is also responsible for the contraction phase, which is part of the T cell immune response. The normal course of T cell immunity has three phases: expansion, contraction, and memory phase. In the contraction phase, AICD causes a massive decline in T cell numbers, after which only around 5% of the T cells that went through activation and proliferation remain and make up the memory T cells. This is the beginning of the memory phase[81].

1.4.2 Types of $CD4^+$ T cells

As with T cell development, most of what we know about differentiation and types of T helper cells comes from the murine system. The pioneering work by Mosmann et al. in 1986 led to the discovery of two functionally different types of $CD4^+$ T cells in mice. They named these two T_H1 and T_H2, and they are distinguished from each other based on the cytokine expression profile in response to various stimuli. Further studies showed that specific transcriptional programs are responsible for these cell fates. In 1997 Zhang et al. showed the dependence of T_H2 lineage on GATA3, a transcription factor (TF), and a known IL-5 promoter activator. Followed by Szabo et al. in 2000, when they described a novel TF called T-bet (TBX21), responsible for T_H1 lineage commitment.

The function of T_H1 cells is to signal to phagocytic and cytotoxic cells, to involve them in a fight against intracellular pathogens. T_H1 cells can also help to fight intracellular bacteria that have managed to survive in macrophages. T_H1 cells can activate those macrophages to fight the intracellular pathogens after detecting the bacterial antigens on their surface.

T_H2 cells, on the other hand, help to deal with bacterial toxins and extracellular pathogens by facilitating antibody production[82]. T_H2 cells are specifically required for B cells be able to start producing Immunoglobulin E (IgE) antibodies, whose main role is the defense against parasites. Because IgEs are also responsible for many types of allergic reactions, such as conjunctivitis, asthma, rhinitis, and food allergies, that makes the T_H2 cells doubly interesting[6,83]. This T_H1-T_H2 dichotomy lasted until 2005 when the third subset of T helper cells was discovered – T_H17 cells. To this end, it was shown by Harrington et al. that T_H17 cells develop in a T_H1 and T_H2 independent manner. Although T_H17 cells seem to have

some role in early defense mechanisms and they are thought to be positive regulators of immune response (increased inflammation), there has been a lot more research into T_H17 cells that are implicated in autoimmune diseases, such as psoriasis, MS, Crohn's disease, and arthritis[84,85].

These classical T_H types of the adaptive immune system also have their counterpart in the innate branch of immunity. The innate lymphoid cells or ILCs are an integral part of innate immune response, and similarly to T helper cells, there are three main subtypes of them.

The first group contains natural killer (NK) cells and ILC1 cells. They produce IFN-γ when activated, and their transcriptional program is controlled by TBX-21, similarly to T_H1s. The second group, ILC2, produce T_H2 type cytokines, such as IL-4, IL5, and IL-13, and like T_H2, the transcription factor GATA3 controls their development. The last group of ILCs, the ILC3, are similar to T_H17 cells in that their development depends on the TF RORγt, and they produce IL-22 when activated[86,87]. However, these groups were first defined in mouse studies, and there is less conclusive evidence of ILC groups that are distinctly different from NK cells in humans. The study of ILCs has been made more difficult due to the lack of any specific lineage markers. The few single-cell studies so far have not been able to find distinct ILC types from a mixed population of lymphocytes[88–90], but there seems to be some evidence of tissue specificity[91].

Around the same time as T_H17 discovery, additional evidence had accumulated that a fourth group of T helper cells could be added to the list[92]. A specialized set of $CD4^+$ T cells was described in the context of B cell maturation and high-affinity antibody production that takes place in germinal centers. This new subset was named follicular B helper T cells or T_{FH}[93]. Another addition to T helper lineages was the T_H9 cells characterized by their expression of IL-9, and they are implicated in allergies, inflammation, and anti-tumor immunity[94,95] (Table 1.4.1).

Corresponding T helper subsets have also been found in humans, and the cell fates seem to be similarly controlled by master transcription factors. GATA3 (T_H2), TBX21 (T_H1), and RORC (T_H17) all control similar pathways during T helper cell differentiation in humans as they do in mice[96–98]. However, from the very beginning of T helper research, there were some notable differences. For example this, T_H1-T_H2 dichotomy was not so clearly present in humans as it was in mice. Many human T helper cells produce some or all of the cytokines from both T_H1-T_H2 lineage[99,100]. Not only that, but many immune cells such as

monocytes, B-cells, eosinophils, NK-cells are capable of producing the "classical" T_H1-T_H2 cytokines as well[101]. Many recent studies have revealed a lot more plasticity in T helper cell populations in humans, as well as in mice. In this regard, it is important to note that T_H1, T_H2, T_H17, and T_{FH} all have been successfully reprogrammed to produce cytokines of some other T helper lineage[102–105]. This reprogramming is usually achieved by changing the culture conditions to that of the desired lineage. Some of these plasticity experiments have also been carried out in living organisms. In 2012, Panzer et al. showed that in vitro polarized T_H1 and T_H17 can be turned towards T_H2 fate in vivo. It has also been shown that T_H1 and T_H2 cells commonly express both of the master regulators: GATA3 and TBX21 in the mouse system and in the human system[106]. The discovery of several other tentative T helper cell lineages such as GM-CSF (CSF2) producing T_HGM cells[107,108], IL-22 producing T_H22 cells[109] and TGF-β, IL-4, IL-10 producing T_H3 cells[110], has made it increasingly more apparent that well-defined subsets of T helper cells as such might not exist in vivo. At least not in clear and discrete sub-populations. The entire T helper system might be more complex than simple low parametric read-outs suggest.

Table 1.4.2 T helper cell subsets with cell fate-specific master regulators and their role in immunity.

	T_H1	**T_H2**	**T_H9**	**T_H17**	**T_{FH}**
Cytokine Profile	IFN-γ, IL-12, TNF	IL-4. IL-5, IL-13	IL-9	IL-17A, IL-17F, IL-22	IL-4, IL-21
Master TF	TBX21, STAT4, STAT1	GATA3, STAT6	STAT6, PU.1, IRF4	RORγt/RORC, STAT3	BCL-6, STAT3
Role in defence	Intracellular viral and bacterial defence	AB production, eosinophil activation	Anti-tumor immunity	Mucosal barriers	B cell development
Role in autoimmunity	MS, Diabetes	Asthma, chronic inflammation	Allergies, inflammation	Arthritis, IBD, Asthma	Lupus, GPA

Therefore, it is no surprise that scRNA-seq technologies hold the promise to finally clarify the enigma on whether discrete, terminally differentiated T helper subsets exist or whether the system contains more plasticity than originally thought. So far, most single-cell studies have revealed a heterogeneous and continuous T cells population, regardless of whether in mice or men[67,111–114]. It turns out that defining cell types in a very high-dimensional dataset is more complicated than in more conventional read-outs like ELISA and FACS. It is relatively easy to tell cell types apart based on the expression of a few cytokines or receptors, but if the entire transcriptomes of those cells are compared, these differences will become less apparent[115].

Moreover, cell type assignment in scRNA-seq data is commonly performed on dimensionality reduced data. Dimensionality reduction (DR) methods fall into two categories, linear dimensionality reduction (LDR), such as PCA or LDA, or nonlinear dimensionality reduction (NDR) methods like tSNE, UMAP. For single-cell data, NDRs are preferred as they are better at capturing the structure of data in lower dimensions. The problem with this approach is that regardless of the method used, reducing the dimension from tens of thousands to just a two will introduce substantial distortions[116]. It is also possible to perform the clustering (cell type assignment) in higher dimensions, but for visualization purposes, two or a maximum of three dimensions is all that is practical to depict. The current understanding is that the gene expression data, especially scRNA-seq data, is inherently low dimensional. As Heimberg et al. showed, genes are co-expressed in modules, or transcriptional programs, which vastly reduces the number of dimensions from the number of all genes down to the number of modules in a given cell type. The significant variation and dominant transcriptional programs should be preserved in reduced dimensions[117], but the question remains, how to determine the optimal number of dimensions for DR methods and how reliable are such methods for cell-type assignment. For now, very little research has been put into discovering the inherent pitfalls in using DR data for downstream analysis[116,118,119].

2 Aims of the Study

As discussed above, cellular differentiation is a complex topic, but recent advances in transcriptomics technologies have made the study of these processes easier than ever before. T helper cells are responsible for a great many immune responses and functions in humans and other vertebrates, but in order to function properly, they need to go through many rounds of differentiation. Therefore, it is vital to have an in-depth knowledge of helper T cell activation and differentiation processes, to develop therapies and cure associated diseases. However, despite decades of research, there are still several open questions concerning helper T cell differentiation. When and how exactly is the cell fate determined, how do cells retain their effector functions, and how do T cell subsets differ from each other transcriptionally on a single-cell level.

Our first objective was to set up and optimize a single-cell sequencing platform and test its capabilities in biologically meaningful situations. The second objective was to use this platform to study human T-helper cells. More specifically, we wanted to study the activation and differentiation of helper T cells to identify and describe the types of memory T cells and unravel time-dependent regulators of cell fate.

While Chapter 4 focuses on the first aim, the second aim is explored in chapter 5.

3 Materials and Methods

3.1 Materials

The consumables for sterile and non-sterile laboratory work were purchased from the following manufacturers: Bioplastics, Bio-Rad, Biozym, Corning, Greiner, Labomedic, Neolab, Sarstedt and VWR. The consumables for sterile and non-sterile laboratory work came from the following manufacturers: Bioplastics, Biorad, Biozym, Corning, Greiner, Labomedic, Neolab, Sarstedt, Starlab, *and VWR.*

3.1.1 Antibodies and FACS reagents

Name	**Supplier**	**Application**	**Dilution**
CCR7-BV421, G043H7, mouse IgG2a	BioLegend	FACS	1:25
CD14-FITC, mouse BLD-301804	BioLegend	FACS	1:50
CD19-BV711, mouse BLD-302246	BioLegend	FACS	1:50
CD3-PerCP, clone UCHT1, mouse IgG1	BioLegend	FACS	1:100
CD4-Qdot605, clone S3.5, mouse IgG2a	Thermo Fisher	FACS	1:100
CD45RA-PE, clone HI100, mouse IgG2b	BioLegend	FACS	1:200
IFN-γ-BV711, clone B27, mouse, IgG1	BD Bioscience	FACS	1:100
IL-4-PE, clone 8D4-8, mouse, IgG1	BD Bioscience	FACS	1:100
LIVE/DEAD™ Fixable Near-IR Dead Cell Stain	Thermo Fisher	FACS	1:1000
NucBlue™ Live ReadyProbes™	Thermo Fisher	FACS	2 drops/mL

3.1.2 Buffers and solution

Homemade Buffers	**Components (end conc.)**
50X TAE Buffer	242g Tris 57.1 ml Acetic acid 18.6g EDTA 2Na-2H2O add water to 1L
Direct Lysis Buffer	0,2 mg/ml Proteinase K 1 mM CaCl2 3 mM MgCl2 1 mM EDTA

	1% Triton X-100 10 mM Tris pH 7,5
FACS buffer	2mM EDTA 2% FCS in PBS
TE Buffer	10 mM Tris-Hcl 1 mM EDTA Na2

3.1.3 Cell culture media and reagents

Cell Culture (media and reagents)	**Supplier**
Advanced RPMI 1640	Gibco
B-estradiol	Sigma-Aldrich
CD28	Gibco
CD3	Gibco
DPBS	Gibco
Fetal Cow Serum (FCS)	Gibco
GeneJuice	Merck
HEPES	Gibco
hIL-3	PeptoTech
hM-CSF	PeptoTech
Ionomycin	Alomone Labs
LPS (E.coli)	Invivogen
Penicillin/streptomycin	Gibco
Phorbol 12-myristate 13-acetate (PMA)	Enzo Life Sciences
Rabbit Anti-Human IFN-gamma	PeproTech
Recombinant human IL-12	R&D Systems
Recombinant human IL-2	R&D Systems
RPMI 1640	Gibco
Sodium Pyruvate	Gibco

3.1.4 Chemicals and reagents

Chemical and Reagent	**Supplier**
6X Loading Dye	Thermo Fisher

Agarose Ultrapure	Biozym
Aqua	Braun
Biocoll	Merck Millipore
DDT	Carl Roth
DEPC H2O	Invitrogen
DNA AWAY	Thermo Fisher
DNA Stain G	Serva
dNTPs	Genaxxon
EDTA 0.5M	Invitrogen
Ethanol Ultrapure	Carl Roth
Gene Ruler 100 bp, 1 Kb, and plus	Thermo Fisher
Igepal	Sigma-Aldrich
Isopropanol	Carl Roth
NaCl	Sigma-Aldrich
PEG 8000	Sigma-Aldrich
RNAprotect Cell Reagent	Qiagen
RNase Zap	Sigma-Aldrich
Sodium Azide	Sigma-Aldrich
Tris 1M	Invitrogen
TritonX-100	Carl Roth
UltraPure Distilled Water	Invitrogen

3.1.5 Enzymes and enzyme buffers

Enzymes and buffers	**Supplier**
5x DNase Buffer	Thermo Fisher
5x Reverse Transcriptase buffer	Thermo Fisher
5X SuperScript IV RT buffer	Invitrogen
DNase I	Thermo Fisher
ERCC	Thermo Fisher

Exonuklease I	NEB
KAPA	Kapa Biosystems
Maxima H- Reverse Transcriptase	Thermo Fisher
Proteinase K	VWR
RNasine Plus	Promega
SuperScript IV	Invitrogen
Terra Polymerase	Terra
UHRR	Thermo Fisher

3.1.6 Kits

Kits	**Supplier**
Cell Line NucleofectorTM Kit T	Lonza
High Sensitivity DNA Kit DNA Kit	Agilent
IFN-gamma ELISA	R&D Systems
IL-13 ELISA	R&D Systems
IL-4 ELISA	R&D Systems
MiSeq Reagent Kit v2 (300-cycles)	Illumina
Naive CD4 T cell isolation kit II, human	Miltenyi Biotec
Nextera XT DNA Library Preparation Kit	Illumina
Pan T cell isolation kit human	Miltenyi Biotec
Quant-iT™ PicoGreen™ dsDNA Assay Kit	Invitrogen
RNAdvance Viral Extraction Kit	Beckman Coulter
TNF ELISA	BD Biosciences

3.1.7 Primers

SCRB-seq Barcoded Primers (96) contain unique 8 bp long barcodes (indicated with **XXXXXXX** in the table below) and unique molecular identifiers UMI indicated with N – Any base with a V – Any base except T to make them anchored.

Name	Sequence	Used for
Barcoded oligo(dT)	/5Biosg/ACACTCTTTCCCTACACGACGCTCTTCCGATCT **XXXXXXXXNNNNNNNNNN**TTTTTTTTTTTTTTTTTTTTTTTTTTTVN	SCRB-seq
Modified i5	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCC AGTGTAAATATGAGAGGTAT	SCRB-seq
Single Primer PCR	ACACTCTTTCCCTACACGACGC	SCRB-seq
TSO	ACACTCTTTCCCTACACGACGCrGrGrG	SCRB-seq
IRF3 Fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGG TCCATATGAAGTCTCCAGA	Genotyping
IRF3 Rev	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAAC AGCCGCTTCAGTGGGTTCT	Genotyping
IRF7 Fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCT TCCAGGGCACGCGGAAACA	Genotyping
IRF7 Rev	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAAC ACCTGACCGCCACCTAACT	Genotyping
MYD88 Fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGCA CGTTCAAGAACAGAGACAG	Genotyping
MYD88 Rev	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTT GCTCGGGGCTCCAGATTGT	Genotyping
TBK1 Fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGA ATTTTGTCCATGTGGGA	Genotyping
TBK1 Rev	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGT GTCATATCTAATGAAGCATTGCA	Genotyping
TICAM1 Fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTA GAGGCATTGAAGGCCGATG	Genotyping
TICAM1 Rev	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTC GGGCCTCATCCTGAAGTTC	Genotyping
RELA Fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAATG GGGCTGCGGTGTCCCCTG	Genotyping
RELA Rev	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAGA CATCCAAACCTGACTCCCA	Genotyping
CHUK Fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCAA AGACACCAAAGCTCAAGGA	Genotyping
CHUK Rev	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAG	Genotyping

	CATCAGAGTAGATTTGTACA	
IKBKB Fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTCA GGGGCATGCGGCATTTATC	Genotyping
IKBKB Rev	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGC AGAGTGTGCTCCTTTCCTC	Genotyping
IKBKE Fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCGA GACGAACTTCTCATCATCA	Genotyping
IKBKE Rev	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAAC TCCTGTCTCTCTGGATGCA	Genotyping
GATA3 Fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCTC AGCCACTCCTACATGG	Genotyping
GATA3 Rev	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTCT CCAGTTCGCTTTCGG	Genotyping
TBX21 Fwd	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGA GGACTACGCGCTACC	Genotyping
TBX21 Rev	TGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAAA CCGAAGTCGCATCCCT	Genotyping

3.1.8 sgRNAs

sgRNAs oligos are composed of the 5’-GGAAAGGACGAAACACCG-3’ sequence, followed by the specific target site in the table below without PAM sequence, followed by 5’- GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGG-3’. The PAM sequence is highlighted in bold.

Target Gene	**Proteospacer**
CHUCK	TAGTTTAGTAGTAGAACCCA**TGG**
IKBKB	GAAGGTATCTAAGCGCAG**AGG**
IKBKE	GCATCGCGACATCAAGCC**GGG**
IRF3	GGGGTCCCGGATCTGGGAGT**GGG**
IRF7	GCAGCCCCACGCGTGCTGTT**CGG**
MYD88	GCTGCAGGAGGTCCCGGCGC**GGG**
RELA	GCGCTTCCGCTACAAGTGCG**AGG**
TBK1	TTCAGATTCTGGTAGTCCA**TAGG**
TICAM1	GGCCCGCTTGTACCACCTGC**TGG**
GATA3	CCTACTACGGAAACTCGGTC**AGG**
TBX21	CCTGTTGTGGTCCAAGTTTA**ATC**

3.1.9 Laboratory equipment

Machine	Supplier
200/2.0 Power Supply	Bio-Rad
2100 Bioanalyzer	Agilent
4D-NucleofectorTM	Lonza
Absorbance readers	Tecan
BD FACSMelody™ Cell Sorter	BD Biosciences
Biomek® FXp liquid handler	Beckman Coulter
C1000 Touch Thermal Cycler	Bio-Rad
Centrifuge 5430	Eppendorf
Chemidoc imaging system	Bio-Rad
E-Gel Precast Agarose Electrophoresis System	
Gene Pulser Xcell	
HydroSpeed™ Microplate Washer	
MACS-Separators	
MiSeq	
SH800S Cell Sorter	
Spark® 20m multimode Reader	
TC-20™ Automated Cell Counter	

Thermo Fisher
Bio-Rad
Tecan
Miltenyi Biotec
Illumina
Sony
Tecan
Bio-Rad

3.2 Methods

3.2.1 Agarose gel electrophoresis

DNA fragments were analyzed by agarose gel electrophoresis. For a 1% gel, 1 g of agarose was dissolved in 100ml of 1x TAE buffer and heated in the microwave until the solution cleared completely. dsDNA dye SYBR™ Green was used at a 1:100 000 dilution. Loading dye from Thermo (cat. R0611) was used to prepare the samples before loading them onto the gel. Thermo 1kb plus DNA ladder (cat. 10787018) was used for sizing the samples. DNA fragments were separated under 100 V for 45 min. and imaged with Biorad Chemidoc system.

3.2.2 Cell Lines

BLaER1: these cells are a subclone of a human B cell lymphoma cell line engineered to express a construct – C/EBPαER-GFP. Treating these cells with β-estradiol converts them into functional macrophage-like cells[120].

3.2.3 Cell culture methods

Incubators used for culturing all cells in this work were kept at 37° C and increased CO_2 concentration (5%).

BLaER1 media contained RPMI 1640 supplemented with 10% FCS, 1% sodium pyruvate, 100 U/mL penicillin-strepomycin.

Transdifferentiation of BLaER1 cells: BLaER1 cells were plated into 96-well (flat-well) plates in 70 000 cells per well concentration. Transdifferentiation media contained 10 ng/mL IL-3, 10 ng/mL M-CSF, and 100 nM b-estradiol. Cells were expanded for 5 days, after which they were harvested for the single-cell experiments.

Primary human $CD4^+$ T cells were cultured in Advanced RPMI 1640 medium supplemented with 5% HS, 10 mM HEPES, and 100 U/mL penicillin-streptomycin.

T cell differentiation: 48-well culture plates were prepared one day before the experiment. Wells were coated by adding 100 µL of PBS containing (1 µg/mL) Anti-Human CD3 Antibodies overnight. Wells were washed/blocked with Advanced RPMI 1640 media once before the T cells were plated at 250 000 cells per well in 500 µL containing 0.5 µg/mL Anti-Human CD28 Antibodies and IL-2 (50 U/mL).

For T_H1 skewing condition, additional IL-12 5 ng/mL was added. For T_H2 condition, anti-IFN-γ antibody (1 µg/mL) and IL-4 50 ng/mL was added. The non-skewing condition was kept as is. Media was exchanged every 3 days by removing half of the old media and adding 250 µL fresh media containing IL-2 (50 U/mL).

3.2.4 Stimulation of immune receptors with specific agonists in BLaER1 cells

Unless otherwise stated, the BLaER1 cells were stimulated with LPS 2 ng/mL or transfected with 200 ng/well of HT-DNA in combination with 0.5 µl lipofectamine or left untreated. After 2 hours, the cells were harvested and proceeded to cell-sorting.

3.2.5 PBMCs and $CD4^+$ naive T cell isolation

PBMCs were isolated from heparinized human blood of healthy consenting volunteers (according to the Declaration of Helsinki and approved by the responsible ethical committee) using a Bicoll (Merck cat. L6115-BC) density gradient centrifugation, red blood cells were removed with RBC lysis buffer (BioLegend cat. 00-4333).

T cells were purified either directly from PBMCs or CD14-depleted flowthrough after monocyte isolation with Naive $CD4^+$ T Cell Isolation Kit II (Miltenyi).

FACS purified T cells were sorted from enriched T cells for which Pan T cell isolation kit (Miltenyi) was used on PBMCs.

3.2.6 KO cell-line generation by CRISPR-Cas9

CHUCK$^{-/-}$, IKBKB$^{-/-}$, IKBKE$^{-/-}$, IRF3$^{-/-}$, IRF7$^{-/-}$, MYD88$^{-/-}$,TBK1$^{-/-}$, and TICAM1$^{-/-}$ KO generation in BLaER1 cells was carried out by Thomas Ebert.

RELA$^{-/-}$ KO was generated by Dennis Nagl.

KOs in T cells (GATA3$^{-/-}$ and TBX21$^{-/-}$) were generated by Andreas Linder

BLaER1 KO generation:

KOs in BLaER1 cells were generated as described in[121]. Briefly: sgRNA was from our own library[122]. One day before electroporation, BLaER1 cells were plated in a 96-well at 200 000 cells per mL. The next day the cells were suspended in 250 µl Optimem medium containing 5 µg of target plasmids; both for the expression of Cas9 and gRNA. Cells were incubated at room temperature for 10 minutes, transferred to 4 mm cuvette and electroporated with a Gene Pulser Xcell™ Electroporation System from Biorad using the following settings: exponential decay protocol – 265 V, 975 µF, and 700 Ω. Cells were plated in a 6-well plate in a pre-warmed normal RPMI medium.

One day after, cells were sorted for BFP positivity. 5% highest positive cells were sorted, then plated by limiting dilution cloning into U-bottom 96-well plates. Approximately 3 weeks after living colonies were identified by absorption at 600 nm with Spark20M microplate reader. 1-2 plates per KO were picked and reformatted into a new 96-well plate.

T cell KO generation:

KOs were generated in T cells as described in [123]. Briefly: CD4^{+} T cells were kept in medium overnight following the isolation. Before nucleofaction, T cells were washed with PBS once, and per KO, 2 million cells were resuspended in 20 µl P3 buffer. CRISPR-Cas9-RNPs were prepared as follows: first chemically stabilized synthetic crRNA:tracrRNA pairs (Table 3.1.8) were denatured at 95° C for 5 minutes after which they were incubated at for 30 minutes at RT. Combined gRNA (100 pmol) was mixed with recombinant NLS-Cas9(40 pmol) protein and incubated for 10 minutes at room temperature gRNA.

Samples and RNPs were mixed and transferred into cuvettes, and the EH100 nucleofaction program was carried out in the X-unit of a 4D nucleofector. After, the cells were plated into a 24-well plate and kept in serum free Advanced RPMI 1640 media for up to an hour. To enhance the survival, T cells were activated with anti-CD3 AB + anti-CD28 AB beads in 1:5 ratio in the presence of IL-2 (50 U/mL) and 2.5% human serum. One day post-activation T cells were subjected to limiting dilution cloning in a U-bottom 96-well

plate. Monoclones were restimulated on day 7 with 2000 CD3/CD28 beads per well. Approximately 2 weeks later, living colonies were identified by Cellavista, and 1-2 plates per KO were picked and reformatted into a new 96-well plate.

3.2.7 Genotyping monoclones

Monoclonality and genotype of colonies (both BLaER1 and T cells) were confirmed with deep sequencing. For that 10 µL of sample from each well was taken and mixed with 10 µL of 2x Direct Lysis Buffer on a new plate (Lysis plate) and incubated at 65° C for 10 minutes. Proteinase K was heat-inactivated at 95° C for 15 minutes. This lysate was used as a template for the first step. To confirm the knocking out of a gene, it was necessary to first amplify the gRNA target locus (PCR 1), and then in order to multiplex the samples for deep sequencing, it is required to add barcodes to all the PCR 1 products. Primers for PCR 1 are made of target specific sequences to amplify the genomic region of interest and a universal adapter sequence where PCR 2 primers can attach. PCR 2 primers similarly contain different sequence types; complementary sequence to PCR 1 adapters, barcode sequence for multiplexing and Illumina sequencing adapter. For PCR 2 there were 16 forward primers and 24 reverse primes giving 384 unique barcode combinations.
Both PCR reactions were set up as follows:

H_20 – 3.5 µL
Phusion High GC Buffer – 1.2 µL
Fwd+Rev Primer (5 µM) – 1.2 µL
dNTPs (10 mM) – 0.12 µL
Phusion polymerase – 0.06 µL
Template (lysate or PCR 1 product) – 1.0 µL
Total 6 µL reaction.

Both PCRs were cycled under identical conditions:

$$\frac{95\,°C}{3\,min} \_ \underbrace{\frac{95\,°C}{30\,sec} \_ \frac{62\,°C}{30\,sec} \_ \frac{72\,°C}{30\,sec}}_{16\ \text{cycles}} \_ \frac{72\,°C}{3\,min}$$

All PCR products were pooled and sequenced on an Illumina MiSeq Platform (single end 300 bp read length) using the v2 chemistry. Outknocker[124] v2. was used to analyze the sequencing results. A clone was considered a KO when heterozygous out of frame mutation was identified. Clones identified to be monoclonal, but WT were kept as controls.

3.2.8 Enzyme-linked immunosorbent assay – ELISA

Cytokine concentration in the supernatant of samples was measured by ELISA following the manufacturer's instructions. Briefly: ELISA plates (high-binding) were coated with the capture antibody in recommended amounts and using the kit specific buffer. The coating was done overnight at 4° C. After coating, the plates were washed with PBS 3 times and blocked with PBS containing 10% FCS for 1 hour at RT. Samples (diluted if needed) and standards were loaded (50 µl) and incubated at RT. After 2 hours, the plate was washed with PBS again 5 times, and then the detection antibody and streptavidin-HRP were added. Plates were kept at room temperature for 1 hour in the dark. Another 5 washes followed by the addition of 50 µl of TMB solution. The reaction was stopped with 2N sulfuric acid. Gen5-Epoch microplate reader was used to get the absorbance at 450 nm and 570 nm.

3.2.9 Cell sorting

Cells were centrifuged for 6 min at 450 g and resuspended in the desired volume of FACS buffer (2% FCS in PBS + 2 mM EDTA). Before sorting, cells were strained through 40 µm filter. Cells were sorted using a SONY SH800S or a BD FACS Melody sorter. When possible NucBlue™ Live ReadyProbes™ Reagent (Hoechst 33342) staining was used to exclude dead cells.

3.2.10 Antibody staining for flow cytometry

Cells were harvested, centrifuged for 6 min at 450 g, supernatant discarded and resuspended in 50 µL of FACS buffer (2% FCS in PBS + 2 mM EDTA). 1 µL of each diluted antibody (see Table 3.1.1 for AB dilutions) was added. The sample was incubated on ice in the dark for 25 minutes, after which 500 µL of FACS buffer was added. The sample was then centrifuged for 6 minutes at 450 g, supernatant discarded and resuspended in 500 µL of FACS buffer. All samples were kept in the dark and strained through 40 µm filter prior to analyzing or sorting.

3.2.11 Intracellular cytokine staining

Intracellular cytokine staining was performed on T cells. Multiple samples (250 000 cells each) were stained in parallel in 96-well format. Prior to staining, all samples were stimulated with PMA (50 ng/ml) + Ionomycin (1 µg/ml) for 2 hours to increase the cytokine production or left unstimulated for controls. After stimulation Brefeldin A (10 µg/ml) was added for another 2 hours to inhibit protein transport.

After stimulation, cells were washed once with PBS. BD Cytofix/Cytoperm™ Kit was used as follows:

Samples were resuspended in 100 µl fixation/permeabilization solution and incubated at 4° C for 20 min in the dark.

150 µl BD Permeabilization/Wash buffer was added to each sample and centrifuged 450 g for 5 min. The supernatant was discarded, and the cells were resuspended in 250 µl BD Permeabilization/Wash buffer and centrifuge at 450 g for 5 min. The supernatant was discarded, and cells were resuspended in 50 µl BD Permeabilization/Wash containing the following antibodies: 2,5 µl IFN-γ AB, 2,5 µl IL-17 AB, and 5 µl IL-4 AB.

The samples were incubated for 30 minutes in the dark and on ice, after which they were topped up with 150 µl FACS buffer (PBS 2% FCS + 2 mM EDTA) and centrifuged at 450 g 5 min. The supernatant was discarded, and the cells were resuspended in 250 µl FACS buffer. The samples were kept in the dark and on ice until analysis.

3.2.12 Measuring DNA concentration

To assess the amount of dsDNA in a sample Quant-iT™ PicoGreen™ dsDNA Assay Kit was used. A 1:400 working dilution of dye in 1X TE (Tris EDTA, pH 7.0) was used. Measurements were made in Greiner-CELLSTAR®-96 Chimney Style Well-Plates. Assay solution was 150 µl, and 1 µl of sample per assay was used. A nine well serial dilution standard of known concentration lambda DNA was used (50 ng, 25 ng, 12.5 ng, 6.25 ng, 3.125 ng, 1.5625 ng, 0.78125 ng, 0,390625 ng and 0 ng blank) for absolute quantification. Fluorescence was measured in a Tecan reader.

3.2.13 SPRI bead preparation

Homemade SPRI beads[125] were prepared from the stock bought from GE Healthcare SpeedBeads™ magnetic carboxylate modified particles (product number: 45152105050250). An initial solution of PEG and is prepared as required for application (see table 3.2.11). NB! PEG will need around 1-4 hours to completely dissolve; warmer temperatures and the constant mixing will speed up the process.

Table 3.2.13 SPRI bead reagents table.

Reagent and initial concentration	Standard Beads (22% PEG)	Pooling beads (30% PEG)	Bead Wash Buffer (10mL)
NaCl 5M	10mL (1M)	20mL (2M)	-
Tris-Hcl 1M	500µL (10mM)	500µL (10mM)	100µL (10mM)
EDTA 0.5M	100µL (1mM)	100µL (1mM)	20µL (1mM)
PEG 8000	11g (22%)	15g (30%)	-
Igepal 10%	50µL (0.01%)	50µL (0.01%)	-
NaN3 10%	250µL (0.05%)	250µL (0.05%)	-
H2O	up to 49mL	up to 49mL	up to 10mL

The stock solution of SpeedBeads is resuspended vigorously before 1 mL is transferred into a 1.5 mL tube (300 µL if Pooling Beads are prepared instead). The tube is placed onto a magnetic rack, and the supernatant is removed. The pellet is washed 2 times with Bead Wash Buffer. After that, 1 mL of Bead Wash Buffer is used to resuspend the beads and added to the initial PEG solution. The beads binding efficiency is directly dependent on the salt and PEG concentration. To assay the beads cut-off and binding, a test clean up of low-range DNA ladders is recommended. See DeAngelis et al. 1995 for more details.

3.2.14 Solid Phase Reversible Immobilization (SPRI) bead-based DNA Clean-up

Standard SPRI beads that were prepared according to 3.2.11 were used for cDNA and PCR product clean-up and to facilitate buffer exchange. The beads were brought to room temperature and vortexed thoroughly! The sample was mixed with a required amount of SPRI beads depending on the size selection desired (Fig 3.2.12) and incubated for 5 minutes at room temperature. Samples were transferred to a magnetic stand, and the beads bound with DNA were allowed to migrate to the magnet. The supernatant was

removed, and the bead-DNA pellet was washed twice with 100-200µl 80% ethanol (depending on the size of the pellet). After the final wash, as much of ethanol as possible was removed, and the pellet was air-dried for up to 5 minutes. The DNA was eluted from the beads in the required amount of H_2O or elution buffer (TE pH 8.0).

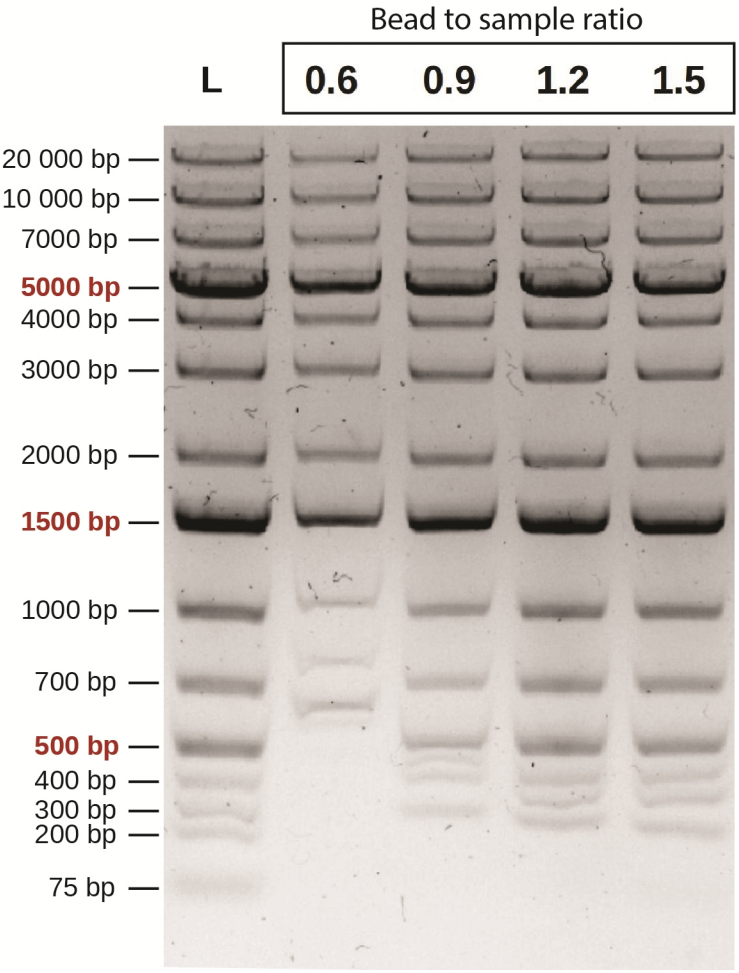

Figure 3.2.14 SPRI beads allow for size selection depending on the concentration of PEG+NaCl in the buffer. Because the beads are premade in the binding buffer (PEG+ŃaCl) in effect, the size selection comes down to the volume of beads (in buffer) used. Lower ratios allow for progressively bigger fragments to be excluded. Each lane shows the result of 10 µl of DNA ladder being "cleaned-up" at different ratios of beads. L – Ladder.

3.2.15 Transposon-based library preparation

Nextera XT DNA Library Preparation Kit was used to perform the library preparation from pre-amplified cDNA libraries (single cell and bulk). Library preparation was performed according to the manufacturer's protocol with the following modifications:

Tagmentation: cDNA for each sample was diluted to 0.2 ng/µL, and 5 µL (1 ng total) was taken per library. Single-cell libraries were run in technical duplicates and bulk samples in triplicates to avoid bottle-necks in the library complexity. Samples were mixed with 10 µl Tagment DNA buffer and with 5 µl Amplicon Tagment Mix after which they were incubated for 10 minutes at 55° C. 5 µL NT buffer was added to each reaction to stop the reaction. After 5 minutes at room temperature, index PCR was started.

Nextera Index PCR Master Mix was prepared for all samples as follows:

For 1 sample:

15 µl – PCR Mix
1 µl – P5NEXTPT5
8 µl – H_2O

24 µL of this Master Mix together with 1 µL of i7 index primer was added to each sample. Tagmented libraries were cycled as follows:

$$\underset{3\,min}{72\,^{\circ}C} \_ \underbrace{\underset{30\,sec}{95\,^{\circ}C} \_ \underset{10\,sec}{95\,^{\circ}C} \_ \underset{30\,sec}{55\,^{\circ}C}}_{13x} \_ \underset{1\,min}{72\,^{\circ}C} \_ \underset{5\,min}{72\,^{\circ}C}$$

Library clean-up: after PCR, libraries were cleaned up using the double-size selection bead clean-up method. The samples were mixed with 0.4X the volume of beads and placed on the magnet. The supernatant was removed and kept and mixed again with an additional 0.4X of the volume of beads (the same volume as in the beginning, i.e. 100 µL of sample + 40 µL of beads + additional 40 µL of beads). Samples were mixed and placed on the magnet, and the normal clean-up process followed (see 3.2.14). This double-size selection allows to remove un-tagemented long fragments and short primer-dimers in one clean-up.

Libraries were QC-ed on *Bioanalyzer* and pooled as desired for paired-end sequencing on an Illumina HiSeq1500.

3.2.16 SCRB-seq original protocol

Single cells were sorted directly into 96-well plates pre-filled with lysis buffer (5 µl of 1:500 dilution of Phusion HF buffer) as described in 3.2.10. After sorting, the plates were spun down and immediately placed on dry ice.

The next day, the cells were thawed up at room temperature for 5 minutes then placed on ice. 1 µl of Proteinase K (20 mg/ml) solution was added to each well, and the plate was incubated at 50° C for 15 minutes. This step was followed by sample desiccation – the seal was removed, and samples were heated to 95° C for 10 minutes. Following the desiccation, the plate was placed on ice and 1 µl of RT master mix, and 1 µl of 2µM E3V6NEXT barcoded primers were added to each well.

RT master mix for one 96-well plate:

40 µl 5X Reverse Transcriptase buffer

20 µl 10 mM dNTPs

2 µl 100 µM E5 TSO

20 µl H_2O

10 µl 1:300 000 ERCCs

12.5 µl Maxima H- Reverse Transcriptase

Total: 104.5

RT reaction was incubated at 42°C for 1:30 hours.

All the samples were pooled and mixed with Zymo DNA Binding Buffer 1:7 ratio, respectively. cDNA was cleaned and concentrated as by Zymo DNA Clean & Concentrator-5 (cat. D4013) protocol. Samples were eluted in 17µl of H2O.

Leftover primers from RT were removed with Exonuclease I treatment as follows: 2 µl of 10X Exo I buffer and 1 µl of Exonuclease I is added to the sample and incubated at 37° C for 30 minutes, followed by heat inactivation at 80° C for 20 minutes.

After this the samples are ready for pre-amplification. Sample DNA from previous step can be directly used and a PCR reaction is set up as follows:

10 µl 5X Kapa HiFi buffer

1.5 µl 10 mM dNTPs

1.5 µl 10 µM SINGV6 primer

16.5 µl H2O

1 µl Kapa HiFi polymerase

The reaction is amplified according to the following program:

$$\underset{3\,min}{98\,^{\circ}C} - \underbrace{\left[\underset{15\,sec}{98\,^{\circ}C} - \underset{30\,sec}{65\,^{\circ}C} - \underset{6\,min}{72\,^{\circ}C}\right]}_{20x} - \underset{10\,min}{72\,^{\circ}C}$$

First safe stopping point. Samples can be kept at 4° C for overnight or frozen for longer periods. Pre-amplified cDNA is cleaned up following the standard SPRI clean-up (see 3.2.14) and quantified (see 3.2.12).

Sequencing ready libraries are produced following the Transposon-based library preparation (see 3.2.15).

3.2.17 SCRB-seq v.3 protocol

Single cells were sorted directly into 96-well plates as described in 3.2.10. Wells were pre-filled with 4 µl lysis buffer.

Lysis buffer for one plate (96-wells):

380 µl H_2O
44 µl 25 mM dNTPs
14 µl 40 U/µl RNAsine
3.4 µl 10% TritonX-100
Total: 440 µl

Additionally, each well also contained 1 µl of 2 µM Barcoded oligo(dT) primers. After sorting, the plates were spun down and immediately placed on dry ice.

The next day the cells were thawed up at room temperature for 1 minute followed by RNA denaturing: samples were incubated at 70° C for 3 minutes and cooled down to 4° C then immediately placed on ice.

RT master mix for one 96-well plate:

283.5 µl H_2O
210 µl 5X Reverse Transcriptase buffer
21 µl 100 µM TSO
10 µl 1:200 000 ERCCs
10.5 µl Maxima H- Reverse Transcriptase
Total: 535 µl

5 µl of this master mix was added to each well. RT reaction was incubated at 42° C for 1:30 hours.

All the samples were pooled and cleaned up using pooling beads (Table 3.2.13). Samples were mixed in 1:1 ratio with the beads. After clean-up, samples were eluted in 17 µl of H2O. Leftover primers from RT were removed with Exonuclease I treatment as follows: 2 µl of 10X Exo I buffer and 1 µl of Exonuclease I is added to the sample and incubated at 37° C for 20 minutes, followed by heat inactivation at 80° C for 10 minutes.

Following the digest, the samples are ready for pre-amplification. PCR mix can be added directly to the samples from the previous step.

Pre-amplification PCR reaction is set up as follows:

25 µl 2X Terra buffer
2.5 µl H_2O
1.5 µl 10 µM SINGV6 primer
1 µl Terra polymerase

The reaction is amplified according to the following program:

$$\underset{3\,min}{98\,^{\circ}C} \_ \underbrace{\underset{15\,sec}{98\,^{\circ}C} \_ \underset{30\,sec}{65\,^{\circ}C} \_ \underset{4\,min}{68\,^{\circ}C}}_{20\text{x}} \_ \underset{10\,min}{72\,^{\circ}C}$$

First safe stopping point. Samples can be kept at 4° C for overnight or frozen for longer periods. Pre-amplified cDNA is cleaned up following the standard SPRI clean-up (see 3.2.14) and quantified (see 3.2.12). Sequencing ready libraries are produced following Transposon-based library preparation (see 3.2.15).

3.2.18 Low-input RNA-barcoding and sequencing protocol.

50 000 cells were harvested and washed per sample. Cells were transferred into RLT lysis buffer containing 0.04M DDT and frozen for at least one day in -80° C. Addition of 1% of Triton X100 was used for T cell low-input bulk sequencing.

Samples were thawed and digested with 1 µL of 20 mg/mL Proteinase K per 50 µL of sample and incubated for 10 minutes at 50° C followed by heat inactivation for 10 minutes at 80° C.

After the digest, RNA was extracted with SPRI beads. Pooling beads (Table 3.2.13) were mixed with 1:1 ratio, and a standard clean-up process was followed. Samples were eluted in 17 µL of H_2O.

For T cell low-input bulk sequencing Beckman Coulter RNAdvance Viral Reagent Kit (cat. C63510) was used, and the accompanying protocol was followed.

DNAse I digest: Since SPRI beads bind both DNA and RNA, a DNAse I digest was required. Samples (17 µL) were mixed with 2 µL of DNAse I buffer (10x) and 1 µL of DNAse I enzyme. The reaction was incubated at 37°C for 30 minutes, after which 1 µL EDTA (100 mM) was added to each sample. Heat inactivation followed at 70°C for 10 minutes.

From this point on 5 µL of the sample was used as an input for RT, and the standard SCRB-seq v.3 protocol (see 3.2.17) was followed with the following modifications:

I. Water amount in RT master mix was reduced by 1 µL per sample to facilitate the higher template volume.

II. Barcoded oligo(dT) primers were added together with RT master mix.
III. Pre-amplification cycles were reduced to 18 as bulk-seq has significantly more input material.

3.3 Data analysis

All single-cell data was sequenced in the Laboratory for Functional Genome Analysis (LAFUGA, Gene Center, LMU Munich) using a HiSeq1500 machine. Sequencing was always paired-end, and the length varied from 50 to 100 bp. The low-input bulk RNA-seq of T cells was sequenced in Max Planck Institute of Biochemistry's NGS Core Facility (Martinsried) using a NextSeq 500.
Raw cytometry data was analyzed using the Flowjo software.

3.3.1 Demultiplexing, Mapping, and Gene Counting

Whenever possible FASTQ demultiplexing, QC, mapping and gene counting were done with zUMIs[126]. If needed, the FASTQ files were also demultiplexed using JE Demultiplexer[127]. FASTQC[128] was used for quality control. STAR[129] aligner was used for mapping and gene counting.

3.3.2 Data Analysis and Visualization

Statistical analysis and visualization were done in R[130] using the RStudio[131] app for Windows and Linux environment:
For single-cell analysis Seurat[132] v4.0 package was followed whenever possible.
For bulk sequencing data analysis, DEseq2[133] package was used.
Most of the visualization was done using ggplot2[134], and its accompanying suites.
Bioconductor[135] open source project was used to manage the host of packages and programs used in R.

1. All R scripts used are available at: https://github.com/kgunnar-lmu/PhD-Thesis

4 Establishing RNA sequencing and barcoding based on SCRB-seq

4.1 Introduction

Single-cell sequencing technology has progressed rapidly since its inception. Currently, there are dozens of different methods and hundreds of protocols about capturing, lysing, and preparing single-cell libraries. In more practical terms, the main pipelines fall primarily into two major categories: droplet-based cell-capture and preparation methods use microfluidic chips and oligo-coated beads like Drop-Seq[28] and Chromium system from 10x Genomics[136] and microwell plate-based (or just plate-based) platforms that usually require flow cytometry for cell sorting and capture. Plate-based methods offer higher flexibility and have a larger variety of protocols for each step, from the initial cell capture to the final sequencing ready library.

4.1.1 Development of plate-based single-cell sequencing methods

In 2012, a few years after the first single-cell RNA-seq paper was published, Smart-seq[33] laid the groundwork for many plate-based methods. It was subsequently updated in 2014 with the addition of Smart-seq 2[34] and in 2020 with Smart-seq 3[35]. The SMART-seq protocol is an unofficial standard in the single-cell sequencing world, and it has been shown to have the highest sensitivity and power (the probability of detecting an effect)[137,138]. However, one major flaw is the requirement of significant hands-on time, and due to the handling of each cell individually until the last step also rather costly. The solution to simplify this process was to “tag” or attach a known sequence (a barcode) to each cDNA molecule during RT or amplification, after which one could pool all the cells together and proceed with far fewer samples[32]. One major drawback of this kind of pooling is that it is no longer possible to get full-length transcript sequences as the barcode will be in either or both ends (3’-5’) of the molecule, and after the fragmentation, only the cDNA fragments with barcode are useful and selected for sequencing. Thus, we end up with counts of genes or digital gene expression.

Another significant change, especially for single-cell sequencing, was the introduction of counting of absolute numbers of molecules with the aid of Unique Molecular Identifiers or UMIs for short[139]. UMI, like a barcode, is a short (usually between 10-20) sequence of randomized bases and is attached to the oligo(dT) primer. After RT, each cDNA fragment

will contain a UMI sequence and makes a one-of-a-kind combination of UMI and sequence. After the PCR, if some sequences are amplified more readily than others, it is possible to identify them. During the analysis and mapping, these sequences are collapsed, so every UMI-sequence pair is only represented once. UMIs allow accurate identification of PCR duplicates, which is especially relevant for single-cell libraries where the capture efficiency of mRNA molecules and amplification bias can vary considerably even within the same experiment.

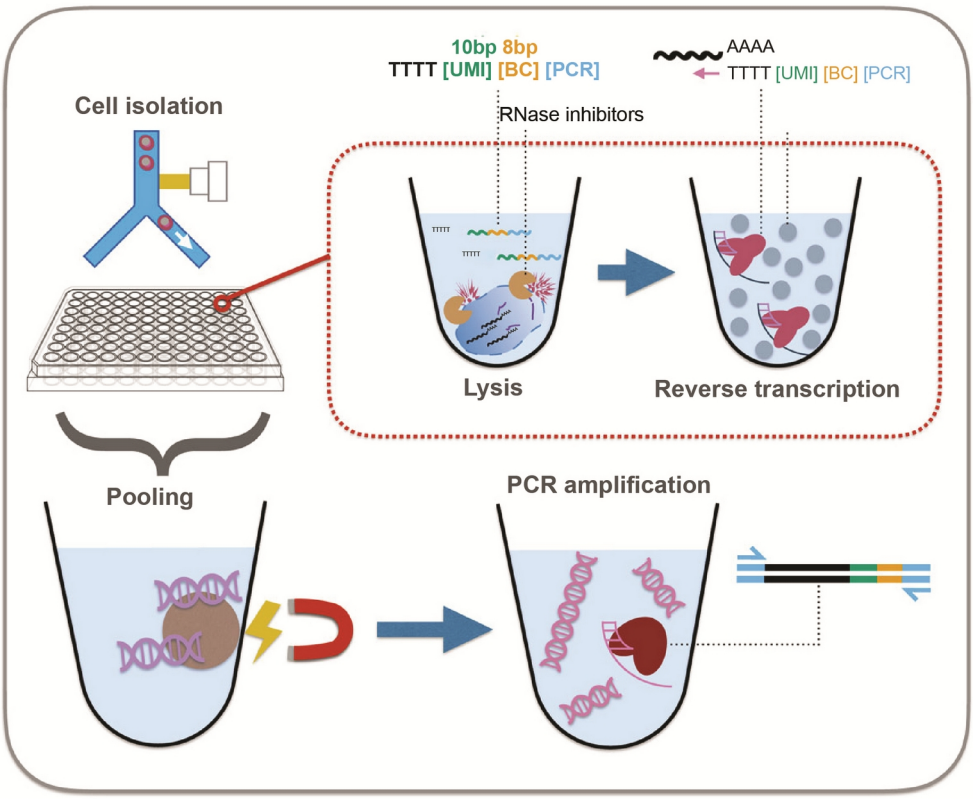

Figure 4.1.1 SCRB-seq protocol overview. 1. Cells are sorted into microwell plates using flow cytometry. 2. cells are lysed in a hypotonic solution in the presence of barcoded oligos, dNTPs, and RNAse inhibitors. 3. after lysis, an RT master mix is added to each well. 4. all samples are pooled, cleaned up, and concentrated using SPRI beads. 5. single-primer PCR amplification produces full-length, barcoded cDNA. Figure adapted from[48].

One protocol that incorporates all those advances is the Single-Cell RNA Barcoding and sequencing or SCRB-seq. SCRB-seq is an affordable and easily scalable 3' digital gene expression RNA sequencing protocol that allows simultaneous sequencing of thousands of cells. It is sensitive and simple to set up. Cells are captured into a microwell plate using flow cytometry. This allows for the use of extracellular markers for additional information about the cells, also called index sorting. This information is saved and can be used later

in conjunction with the sequencing data. After sorting, cells are lysed, and mRNA is captured and reverse transcribed using barcoded oligo(dT) primers. Once barcoded, cDNA can be pooled, cleaned up, and amplified, after which the samples are ready for library preparation and sequencing (Fig. 4.1.1).

Another valuable property of SCRB-seq is the possibility to use the same protocol with low-input bulk samples (~50 000 cells). With low-input bulk input, the protocol has much increased sensitivity. It is a lot cheaper than regular bulk-seq sample preparation, and because of the multiplexing, it is possible to prepare hundreds of samples at the same time. The single-cell resolution is lost, and only a basic gene expression profile can be acquired (no alternative splicing, SNPs, or RNA editing can be detected), but it can be helpful for screening purposes or for samples that are difficult or expensive to acquire.

4.1.2 Innate immune system and PRR signaling

Our immune system is comprised of two main arms: innate and acquired immunity. The innate immune system acts as the first line of defense against exogenous and endogenous threats, such as infection, tissue damage, and cancer. It also functions to instruct and orchestrate adaptive immune system responses. Innate immune responses have been found in all kingdoms of life, and in many cases, the general principles that regulate innate immunity are conserved between species. Innate immune responses are not specific to a particular pathogen; instead, they rely on receptors that have evolved to detect unchanging structures in microbial pathogens, commonly referred to as microbe-associated molecular patterns (MAMPs). For those, there are dedicated receptors in the host called pattern recognition receptors (PRRs)[6].

In mammals, the PRRs can be broadly separated into extracellular and intracellular sensors. One major group of extracellular PRRs are the proteins belonging to the Toll-like receptor family (TLR). TLRs are transmembrane receptors, and as such, they sample MAMPs in the extracellular space. A prominent member of this family is TLR4 that recognizes lipopolysaccharide (LPS), a common cell wall component of gram-negative bacteria[6].

LPS signaling occurs through many interactions. First, it is sensed by TLR4 and its co-receptor MD2. This triggers downstream signaling cascades. TLR4 is unique in the TLR family as it can use all the known Toll-interleukin-1 receptor (TIR) domain-containing adaptor proteins. This TLR4 activation results in two separate signaling pathways

(Fig.4.1.2 A): myeloid differentiation primary response gene 88 (MYD88) dependent and independent pathway. MYD88 dependent pathway leads to a fast cytokine signaling to alert and activate lymphocytes, also known as the NF-κB pathway[140,141].

The MYD88 independent pathway leads to the type I interferon response through interferon regulatory factor 3 (IRF3). This response tends to be slower and helps to trigger antimicrobial programs to limit the spread of the infection and modulate the immune responses and restrain cytokine production[141,142].

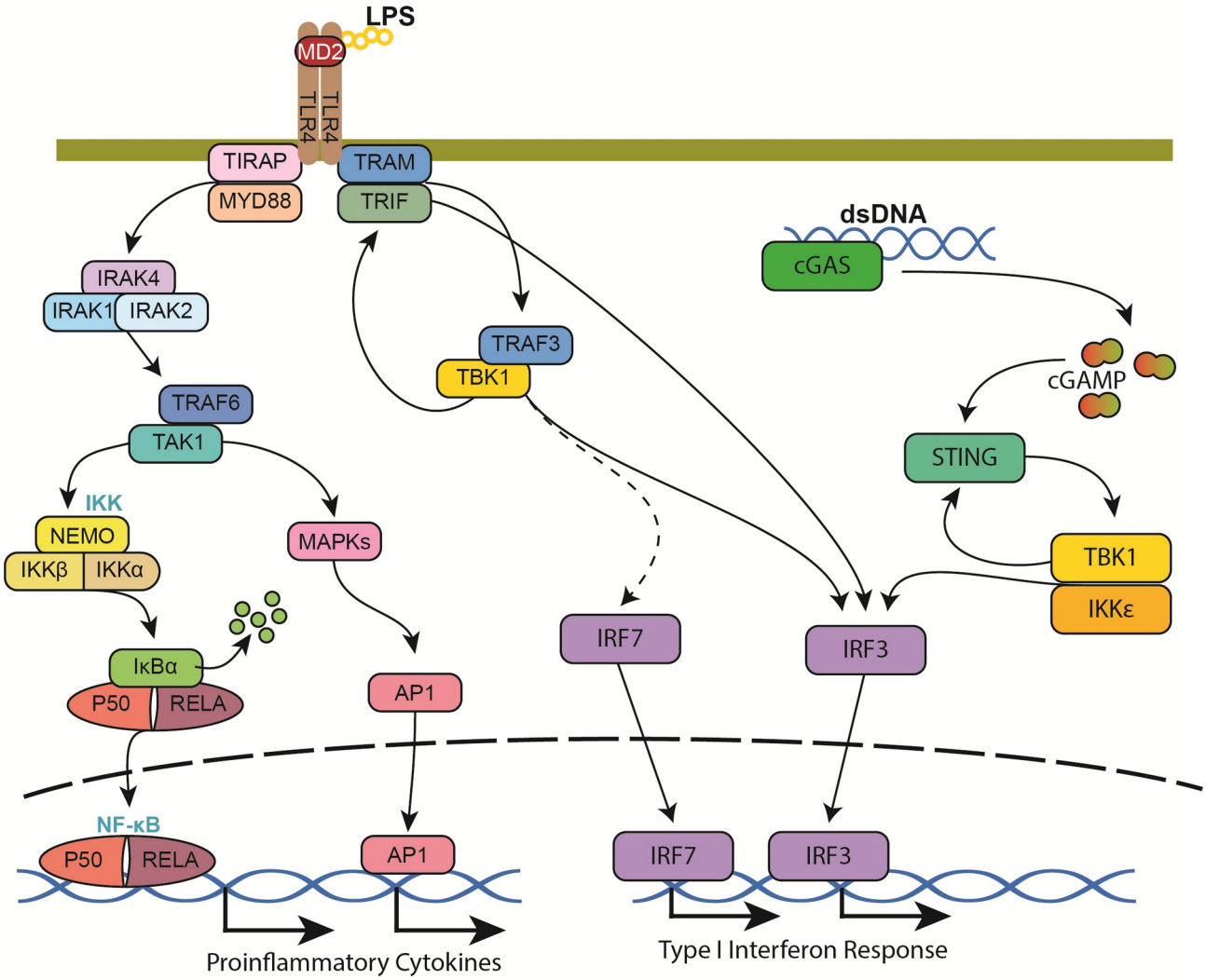

Figure 4.1.2 The TLR4 and cGAS-STING signaling pathways. TLR4 receptor senses LPS and triggers MYD88 dependent and independent signaling leading to proinflammatory cytokine signaling and Type I interferon response. cGAS senses dsDNA in the cytosol and triggers the cGAS-STING signaling pathway resulting in Type I interferon response.

There are many genes that get upregulated directly by NF-κB. Genes such as *Ccl2, Ccl3, Ccl4, Ccl5, Cxcl1, Cxcl2, Cxcl5, Cxcl10, Ccrl2* induce leukocyte recruitment while Hdc,

Nos2, Ptges, Ptgs2 are inflammatory mediators. Also, some inflammatory cytokines like *Il-1α, Il-1β, Il-6, Il-18, Tnf,* and anti-viral *Ifn-B* have increased transcription[143]. NF-κB also limits the inflammatory response by upregulating genes, such as *Tnfaip3,* and *Nfkbia* that inhibit TLR4 signaling in order to protect the host from tissue damage by excessive inflammation.

Intracellular PRRs can also recognize pathogens via MAMP molecules. In addition, some intracellular PRRs can also sense endogenous molecules when they have translocated into a wrong compartment in the context of cell damage. One such sensor is Cyclic GMP-AMP Synthase (cGAS) that detects cytosolic double-stranded DNA (dsDNA). Activation of cGAS leads to the production of cyclic GMP–AMP (cGAMP), which in turn triggers the Stimulator of INterferon Genes (STING) and leads to the type I interferon response through IRF3 (Fig.4.1.2 B)[144,145]. This cGAS–STING pathway induces several interferon signaling genes (ISGs) that help prevent the replication, assembly, and release of viruses[142].

4.2 Overview

As discussed above, the SCRB-seq protocol was already established, but it was optimized for 384-well plates. To avoid the massive upfront cost of 384 barcoded primers, we initially opted to set up SCRB-seq in our lab for 96-well format. Also, during the setup, a few improvements to the method were published[48], and we decided to test those in our setup.

To show the capabilities of the new and improved SCRB-seq, we prepared libraries and sequenced more than 300 macrophage-like differentiated BLaER1 cells. We also used the SCRB-seq protocol to prepare and sequence 168 low-input bulk samples of different knockouts (KOs) of innate immune system regulators and IκB kinases.

4.3 Results

4.3.1 Original SCRB-seq protocol yields low amounts of cDNA and has poor sequencing results

The SCRB-seq protocol was initially designed to work with a 384-well plate format with 2 µl total RT reaction volume per well. While trying to establish the protocol, we decided to opt for a 96-well format to save on the initial investment cost by ordering only a quarter (96) of high purity barcoded oligo(dT) primers (TruGrade® DNA Oligo from IDT). Yet, the initial results from SCRB-seq were not promising. We prepared four plates of transdifferentiated BLaER1 macrophages stimulated with LPS for two hours or left unstimulated as a control. As determined by Quant-iT PicoGreen dsDNA Assay, the final concentration of cDNA libraries, after pre-amplification, was less than 1 ng/µl, and the *Bioanalyzer* (Agilent 1000 DNA chip) detected only low amounts of cDNA or none at all (Fig.4.3.1 A). The most promising sample was subjected to sequencing, which resulted in a very low sequencing quality. Out of more than 100 million reads, only 3.6% were uniquely mapped to the human genome, with a median of 535 genes detected per cell, most of which were mitochondrial reads (Fig.4.3.1 B).

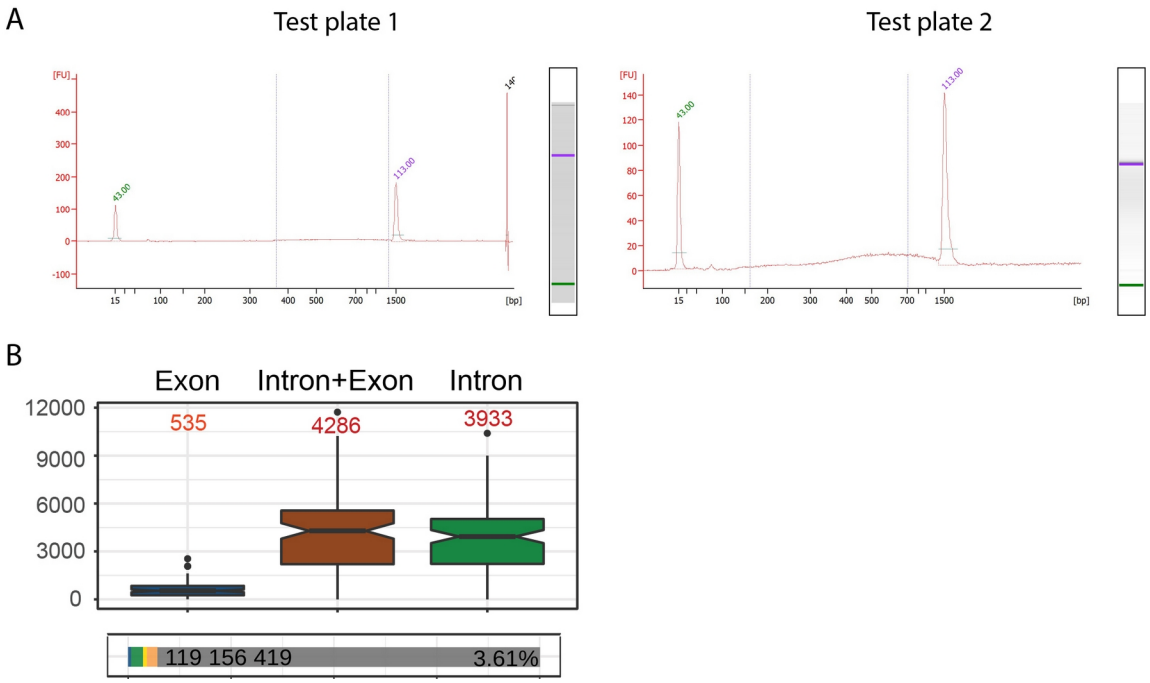

Figure 4.3.1 (**A**) Two plates of scRNA-seq libraries prepared using SCRB-seq protocol. Bioanalyzer traces show the libraries after Nextera tagmentation and PCR. The amount of DNA translates to intensities depicted by Fluorescent Units (FU) on the y-axis. The size of DNA fragments is given in base pairs on the x-axis. The expected peak should be around 400-600 bp, but only on the second panel, a small accumulation of rather long DNA fragments are detectable. (**B**) zUMIs pipeline's statistical output of one plate (94) of single cells. In the top panel, the average number of detected genes in the whole library, based on where in the gene body they mapped to. In the bottom panel, the total number of reads in the library on the left and a percentage of uniquely mapped reads on the right. A gray area indicates unmapped reads.

4.3.2 Changes to SCRB-seq lysis buffer and increased volume of RT reaction increase cDNA yield in 96-well plate format

To optimize the SCRB-seq protocol to suit our needs better, we tried several modifications that have been described to increase the cDNA yield[34,48]. To make up for the larger wells, we increased the RT reaction volume to 10 µl. This lowered the concentration of enzymes in a given reaction, but the total amount used remained the same. Indeed, increasing the reaction volume had a significant, positive effect on cDNA synthesis, as indicated by the PicoGreen assay (Fig. 4.3.2 A).

After establishing a reliably working protocol, we introduced additional modifications. Oligo(dT) barcoded primers were redesigned to extend the barcode from six to eight nucleotides to assist with correct barcode assignment and increase the hamming distance to 1, which means that the new barcodes can now handle a single mutation/error and still be correctly assigned. Before ordering new barcoded primers, a single primer was ordered from IDT (HPLC quality) to produce test libraries. No difference was found in cDNA synthesis efficiency between the original SCRB-seq primer and the extended barcode primer (Fig.4.3.2 B).

As Smart-seq2[34] was the most sensitive[137] single-cell sequencing protocol, we tried to modify our protocol by following its lysis and RT steps. For RT, both Smart-seq2 and another popular protocol, MARS-seq[31], use the SuperScript Reverse Transcriptase (Thermo Fisher Scientific). We compared it to the original SCRB-seq Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific). While both enzymes yielded sufficient cDNA, it was clear that Maxima H- was more efficient, especially for lower amounts of input RNA (Fig.4.3.2 C).

The final change to the original protocol was to modify the lysis buffer. We decided to use a hypotonic lysis buffer containing 0.2% Triton X-100, thermostable ribonuclease inhibitors, dNTPs, and barcoded oligo(dT) primers. We opted to skip the Proteinase K digest, heat inactivation, and RNA desiccation (post lysis). This increased average fragment length, as high temperatures, even for a short time, have been shown to degrade RNA, especially in the presence of metal ions (found in many buffers)[146,147]. Also, by using thermostable ribonuclease inhibitors, the mRNA was protected from the cell-capture stage until the pooling. Having the barcoded primers already in the lysis buffer streamlined the workflow by reducing the handling time necessary for pipetting. Including dNTPs following

lysis has shown to be beneficial for cDNA synthesis[148]. The new lysis buffer did not increase the yield, but it produced more intact cDNA fragments, as seen from the Bioanalyzer traces (Fig.4.3.3 A and C).

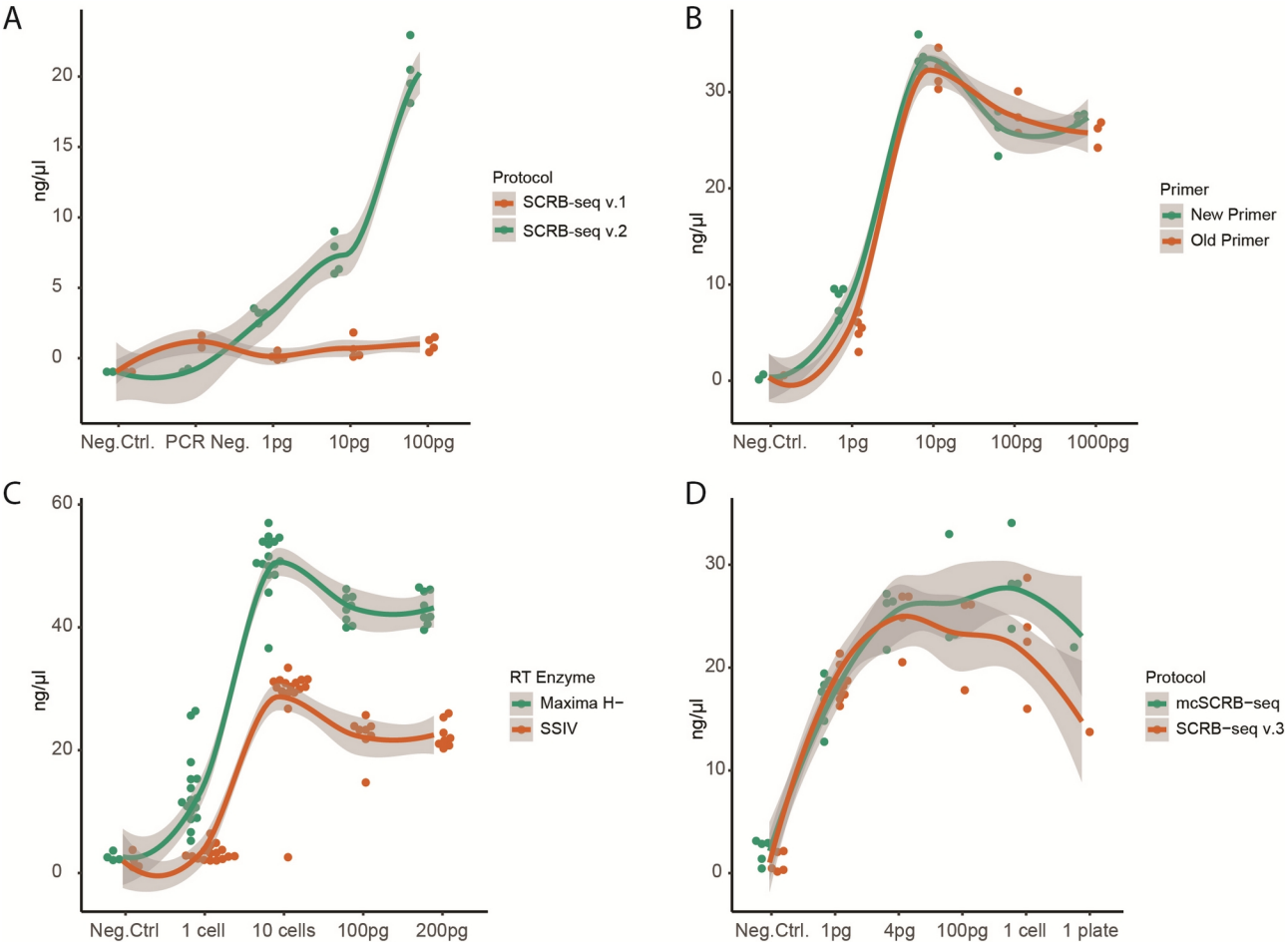

Figure 4.3.2 Quant-iT™ PicoGreen™ dsDNA Assay. All libraries were quantified after pre-amplification. Shown as dots are the cDNA yields of independent experiments. The input used in an experiment is given on the x-axis, either UHRR, BLaER1 cells, or no input – **Neg.Ctrl**. Gray area indicates the 95% confidence interval, solid line shows the locally estimated scatterplot smoothing or LOESS curve. (**A**) Original SCRB-seq (v.1) compared to increased lysis and RT reaction volume protocol (SCRB-seq v.2). **PCR Neg.** – a negative control from pre-amplification step onward. (**B**) Original (old) primer compared to extended barcode sequence primer (new). (**C**) Comparison of two different RT enzymes. SSIV – SuperScript IV. (**D**) Comparison of improved SCRB-seq (v.3) and mcSCRB-seq protocols.

4.3.3 mcSCRB-seq does not improve cDNA quantity nor quality

Around the time we were setting up SCRB-seq, a study was published showing that adding PEG (polyethylene glycol) 8000 to the RT reaction increases cDNA synthesis and

increasing the overall sensitivity (RNA capture efficiency) of the sequencing. This new protocol was also based on SCRB-seq and is called mcSCRB-seq[48]. We followed the instructions described in the paper but the benefits of adding PEG (7.5% end concentration) to an RT mixture were not immediately obvious. At certain concentrations, mcSCRB-seq seemed to increase cDNA yield, but not in a significant way (Fig.4.3.2 D). When analyzing the mcSCRB-seq libraries with the Bioanalyzer, we discovered that, on average, the fragment lengths were shorter than SCRB-seq v.3 libraries (Fig.4.3.3).

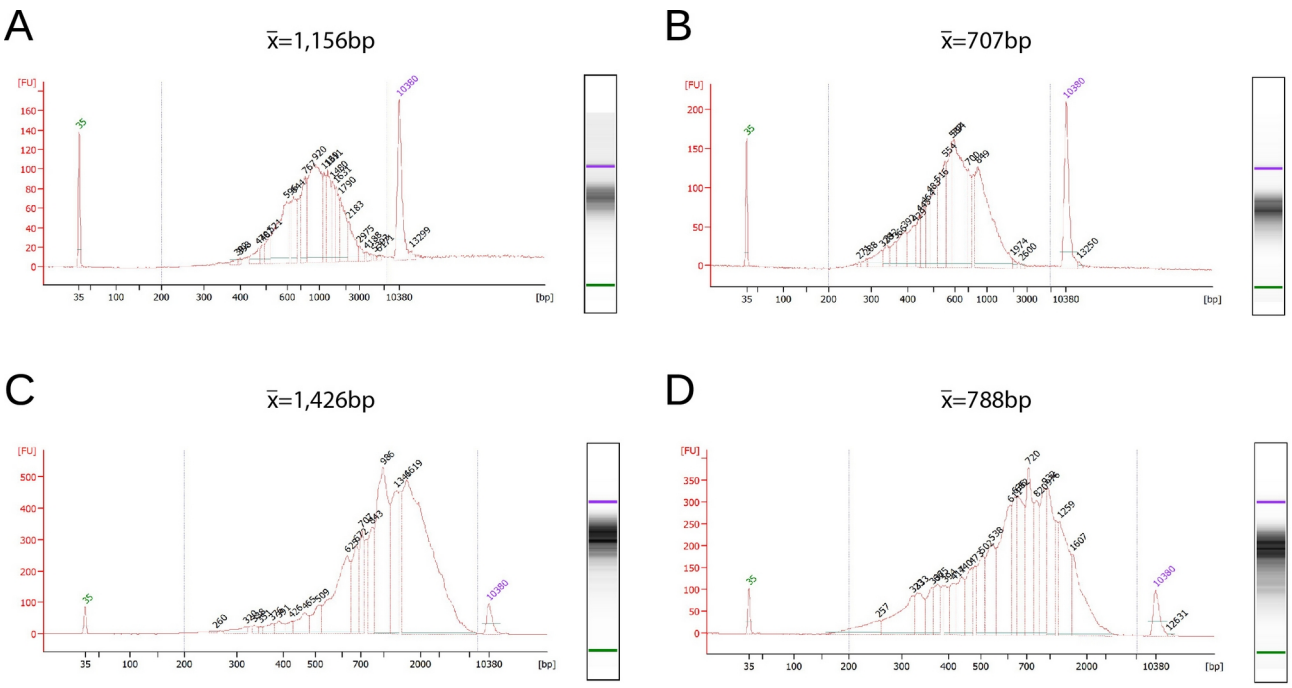

Figure 4.3.3 Bioanalyzer traces. (**A**) A library produced from 100pg UHRR input using SCRB-seq v.3. (**B**) A library produced from 100pg UHRR input using the mcSCRB-seq protocol. (**C**) A library produced from 32 single cells, pooled before pre-amplification, SCRB-seq v.3. (D) A library produced from 32 single cells, pooled before pre-amplification, SCRB-seq v.2. $\bar{X}$ - average fragment length of samples.

4.3.4 Improved SCRB-seq outperforms the original and mcSCRB-seq protocols

To measure if the alterations to the SCRB-seq protocol translated into improved sequencing in a biologically relevant experiment, we stimulated BlaER1 macrophages with LPS followed by sequencing and analysis. We transdifferentiated BLaER1 cells into macrophages, stimulated them with LPS, and harvested the cells for sorting and library preparation. We sorted four 96-well plates: one plate to be prepared using SCRB-seq v.2,

two plates to be prepared using SCRB-seq v.3, and one plate to be prepared using mcSCRB-seq.

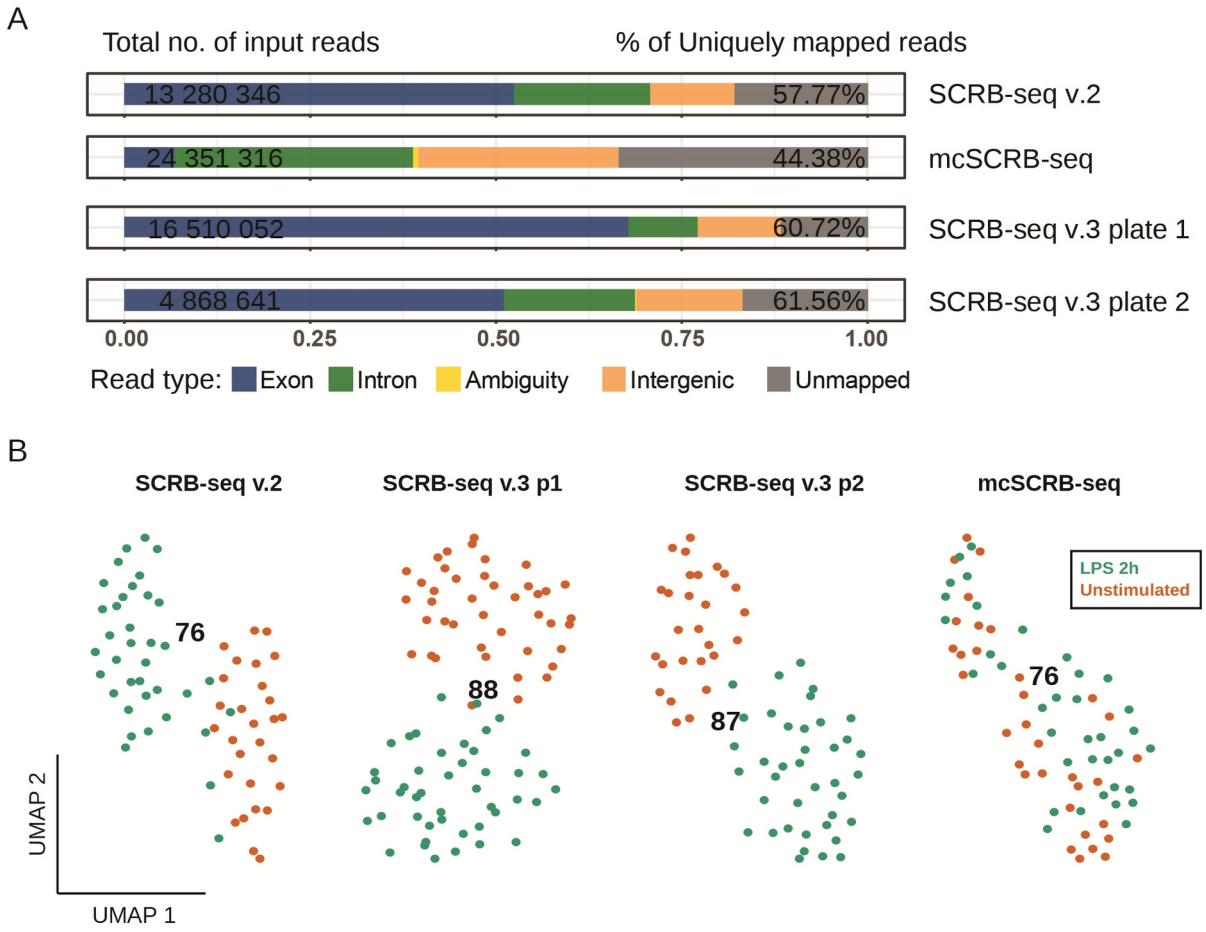

Figure 4.3.4. (**A**) zUMIs pipeline statistical output. The total number reads in each library of 94 cells on the left side. Percentage of uniquely mapped reads on the right. (**B**) scRNA-seq data of four different libraries and three different methods. Index sorting information was used for color overlay in the UMAP plots (based on top 100 HVGs) showing, how well sequencing data captures the effects of LPS stimulation. Numbers on each plot show how many cells out of the 94 were detected and passed the QC and filtering.

After sequencing, we used zUMIs pipline[126] to perform demultiplexing, mapping, feature counting, and quality control (QC) statistics. The two plates prepared using SCRB-seq v.3 had the highest uniquely mapped read percentage, which we took as a sign that those libraries were of good quality. It has been shown that uniquely mapped reads or reads that map to only one location within the reference genome indicate higher quality RNA-seq data[25]. Another indicator of library quality is the percentage of exonic reads, where a low percentage of exonic reads can indicate RNA degradation or DNA contamination[149].

Library prepared using mcSCRB-seq performed the worst in regards to uniquely mapped reads and the percentage of exonic reads. The library derived from the original SCRB-seq protocol with higher reaction volumes and no evaporation step performed marginally worse than SCRB-seq v.3 method (Fig. 4.3.4 A).

Since LPS is a well-studied stimulus, and we had the index sorting information (additional information to verify the accuracy of clustering), we could use the simple clustering of cells based on the stimulation as an indicator of sequencing quality. As expected, we saw a clear separation of cells based on the LPS treatment in all libraries prepared using SCRB-seq improved protocol (Fig. 4.3.4 B). For the mcSCRB-seq library, no clear treatment effect was evident. When we looked for differentially expressed (DE) genes, no hits were found, below the false discovery rate (FDR) of 0.05 or lower (data not shown).

All of the changes introduced into the SCRB-seq protocol helped produce cDNA with higher quality and quantity, and it works equally well for control RNA input (UHRR) as well as for living cells. Our improved SCRB-seq protocol has the highest uniquely mapped reads and exonic reads percentages and the lowest cell dropout level.

4.3.5 Improved SCRB-seq reveals a robust inflammatory response after LPS stimulation in BLaER1 cells

To better analyze our data and to increase the statistical power, we integrated the remaining three successful datasets. As the differences between them are expected to be caused mainly by technical noise, we used the scTransform function in the Seurat package to normalize, scale, and remove confounding sources of variation. As confounding variables, we chose library size (total UMIs per cell) and plates (SCRB-seq v.2,SCRB-seq v.3 p1 and SCRB-seq v.3 p2). The last two datasets, plate one and plate two, mixed together well even without batch correction (data not shown). This was expected as they were sequenced on the same flow cell. The first dataset stood apart from the rest initially, but when regressing out confounding variables, cells from all three datasets made two distinct clusters. The upper population almost exclusively contained LPS stimulated cells, while the lower one was made up of all the unstimulated cells plus some stimulated cells. It is conceivable that these cells either failed to respond to a stimulus or that there was a carry-over during cell sorting, as the unstimulated cells were always sorted first (Fig. 4.3.5 A).

We can again overlay the UMAP plot with the cell sorter index information, and it becomes clear that biological effect dictates the clustering of cells. We also noticed that a few cells from the stimulated group locate to the unstimulated group. For this reason and because the index sort information may not always be available, we ran Louvain clustering to find subsets in our data in an unbiased way. Resulting clustering largely agrees with indexed sort information, where Louvain group 1 is composed of LPS stimulated cells and group 2 is unstimulated cells and cells that did not react to the stimulus (Fig. 4.3.5 A).

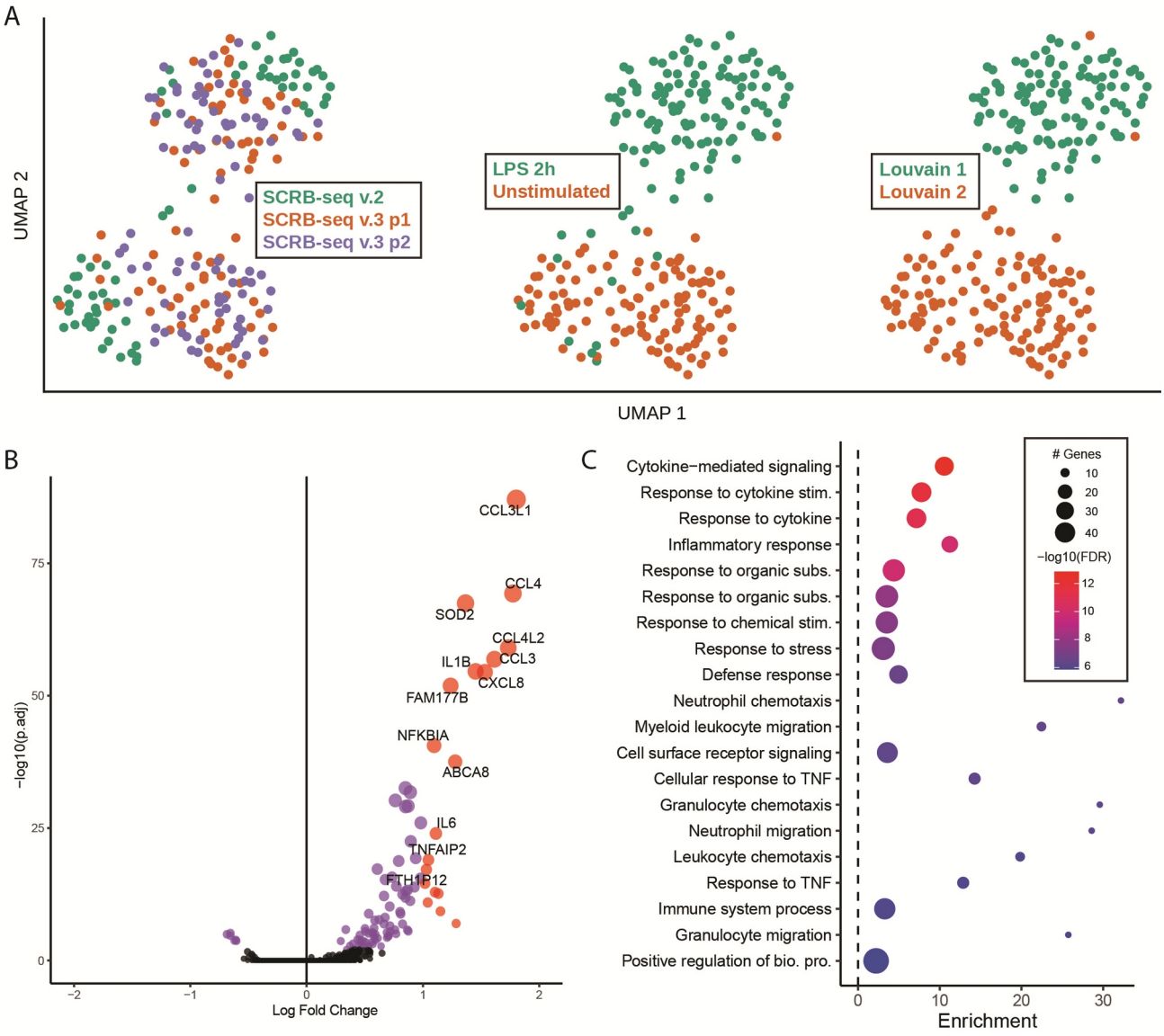

Figure 4.3.5 (**A**) UMAP plots (based on top 3000 HVGs) on three integrated datasets. From left to right: different datasets as seen in Figure 4.3.4.A, cell sorter index information, unsupervised Louvain clustering. (**B**) Volcano plot showing DE results between Louvain cluster 1 and 2. (**C**) Pathway enrichment analysis based on all significantly upregulated genes (FDR<1%).

To find out what effect LPS stimulation has on BLaER1 cells and to further show the capabilities of improved SCRB-seq protocol, we conducted DE analysis between these clusters. Doing so revealed a robust upregulation in several proinflammatory cytokines like *IL-1B*, *CXCL8 (IL-8)*, *IL-6*, and chemokines like *CCL4* and *CCL3* (Fig.4.3.5 B).

As a final approach to validate our findings thus far, we performed pathway enrichment analysis using the PANTHER Pathway tool[150] on all genes significantly upregulated in the 1st cluster with FDR less than 1%. The analysis revealed a significant enrichment of pathways related to cytokine signaling and response, inflammatory response, and defense response (Fig.4.3.5 C).

4.3.6 Low-input bulk sequencing using SCRB-seq protocol enables low-cost transcriptome profiling of hundreds of samples

Another application for SCRB-seq is low-input bulk sequencing. In regular sequencing, the library preparation is performed on millions to tens of millions of cells, resulting in micrograms on input RNA, whereas low-input bulk sequencing using SCRB-seq protocol would allow anything from a single-cell to tens of thousands of cells to be used as input (picograms to nanograms of RNA).

The main difference between low-input bulk and single-cell SCRB-seq is the need for RNA isolation. For single cells, the lysis reaction volume is many orders of magnitude greater than any given cell. Thus, there is no need for RNA isolation and purification (also no need to remove genomic DNA), but with many thousands of cells, the cellular debris, as well as leftover culture media, can interfere with the RT reaction. For this reason, the SCRB-seq v.3 protocol was supplemented with SPRI bead RNA clean-up and with DNAse I digest to remove genomic DNA (gDNA).

After initial testing, we found the optimal input of cells to be around 50,000. This number of cells was required to overcome small variations caused by cell counting and pipetting errors while not inhibiting downstream sample preparations, including bead-based clean-up in a 96-well format. Higher numbers of cells would also require the RNA clean-up to be performed in a separate vessel. We also observed that 50,000 cells combined with the single-cell reaction mix (matched enzyme concentrations, RT, and pre-amplification reaction volumes) to be most optimal for cDNA synthesis and amplification (Fig. 4.3.6 A). To verify this, we prepared four libraries, two with 50,000 and two with 2 million cells, and

proceeded with the library preparation until the end of the pre-amplification step. We then used the Bioanalyzer read-out to assess the quality of libraries. As seen in Figure 4.3.6 B, increasing the cell number had no observable benefit, but instead, there was a clear increase in primer-dimers around 500 bp, as shown by the strong peak in the Bioanalyzer trace. We, therefore, decided to proceed with 50,000 cells per sample for low-input bulk sequencing.

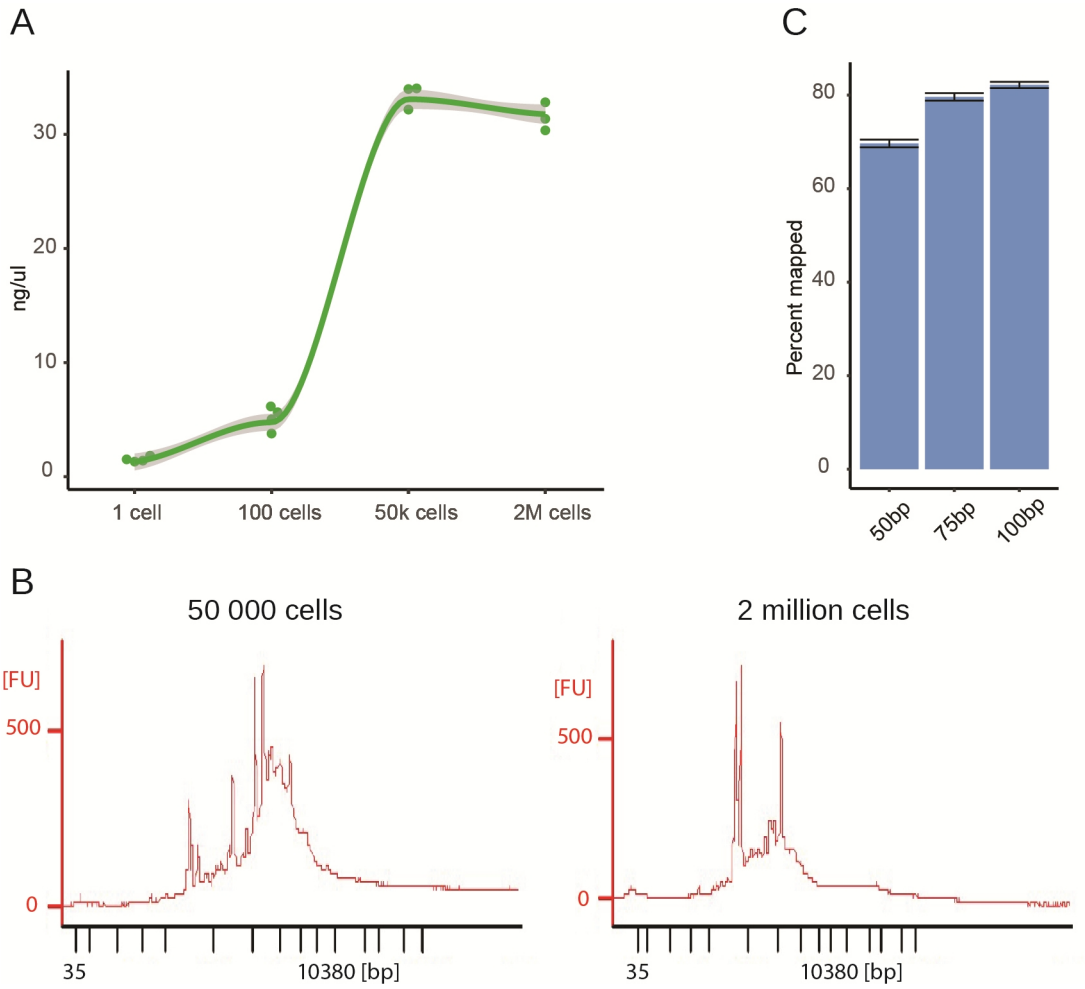

Figure 4.3.6 (**A**) Quant-iT™ PicoGreen™ dsDNA Assay. cDNA concentration measured after 20 cycles of pre-amplification for differing numbers of input cells (BLaER1). Each dot represents an independent experiment (library). Gray area indicates the 95% confidence interval, solid line shows the locally estimated scatterplot smoothing or LOESS curve. (**B**) Bioanalyzer traces of 2 representative samples. All libraries were treated the same, only differing in the input amount. (**C**) Uniquely mapping read percentage depending on the read length depicted as mean +/- SEM of three independent libraries.

As sequencing costs can be prohibitively high, we attempted to find the most cost-effective sequencing length for the cDNA reads. We subjected our samples to 18 bp + 100 bp paired-end sequencing on the Illumina 1500 HiSeq machine. The reads were trimmed to 3

sets of differing read-lengths: 50 bp, 75 bp, and 100 bp. Data from these three sets were processed with zUMIs[126]. As SCRB-seq is effectively a single-end sequencing method, longer reads will increase mapping efficiency. We saw that 75 bp read length was the most cost-effective approach resulting in 10% more or 79,6% uniquely mapped reads compared to the 50 bp reads, and 100 bp long reads showed diminishing returns by only improving a few percentage points (82.1%) (Fig.4.3.6 C). With the protocol optimized to maximize output while limiting cost and handling, we applied the protocol as a proof of principle using a heavily studied pathway in our laboratory.

4.3.7 Low-input bulk-seq reveals Type-I interferon and proinflammatory cytokine signaling pathways in BLaER1 macrophages after PRR stimulation

We employed the optimized low-input bulk SCRB-seq to elucidate the roles of kinases of the IKK family (IκB kinase) in the cGAS-STING and TLR4 signaling pathway. Our goal was to identify transcripts that show a clear NF-κB or Interferon regulatory factor (IRF) signal in a knockout-dependent manner. To this end, we prepared 168 samples comprising of WT cells, as well as single, double, and quadruple knockouts with two different stimulation conditions (see table 4.3.7). The LPS stimulation was chosen for its strong NF-κB induction via the TLR4 signaling pathway, while dsDNA stimulation was chosen for its ability to induce interferon signaling through the cGAS/STING pathway. Cells were stimulated with LPS (2 ng/mL) or transfected with HT-DNA (200 ng/well) for 2 hours, and subsequently, the cells were harvested for RNA-Seq analysis.

Table 4.3.7 Bulk RNA-sequencing sample table. Listed are all the genotypes used for this experiment. In the cells under each stimulation condition, the first number indicates how many samples per genotype were prepared, separated by backslash the second number indicates how many remained after QC and subsequently used in the analysis.

		Stimulation			
	Genotype	dsDNA 2h	LPS 2h	Unstimulated	Total
1	$CHUCK^{-/-}$	4/4	4/3	4/4	11
2	$CHUCK^{-/-}$ x $IKBKB^{-/-}$	4/3	4/4	4/4	11
3	$IKBKB^{-/-/-}$	4/4	4/4	4/4	12
4	$IKBKE^{-/-}$	4/3	4/4	4/4	11
5	$IRF3^{-/-}$ x $IRF7^{-/-}$	4/4	4/4	4/3	11
6	$MYD88^{-/-}$	4/4	4/4	4/2	10
7	$RELA^{-/-}$	4/4	4/4	4/4	12
8	$TBK1^{-/-}$	4/4	4/4	4/3	11
9	$TBK1^{-/-}$ x $IKBKE^{-/-}$	4/4	4/4	4/4	12
10	$TBK1^{-/-}$ x $IKBKE^{-/-}$ x $CHUCK^{-/-}$ x $IKBKB^{-/-}$	4/2	4/2	4/2	6
11	$TICAM1^{-/-}$	4/4	4/4	4/4	12
12	$TICAM1^{-/-}$ x $MYD88^{-/-}$	4/3	4/2	4/2	7
13	WT	8/8	8/8	8/8	24
Total number of samples prepared: 168				**No. of samples left after QC: 150**	

After sequencing, we used zUMIs to demultiplex, map, and count features. We continued the analysis by following the standard Seurat workflow to effectively handle the large number of samples. We first analyzed the clustering within our dataset. Both the graph-based clustering (Louvain Fig.4.3.7. A) and hierarchical clustering (based on top 500 HVGs Fig.4.3.7. B) revealed three main clusters, reflecting the effects of different stimuli and genetic knockouts on gene expression. The clusters are distinguished based on their stimulations: dsDNA in red, LPS in blue, and mixed samples in green. The latter group is mainly composed of unstimulated cells, or cells, in which key signaling pathways are disrupted due to genetic perturbations (Fig. 4.3.7 A).

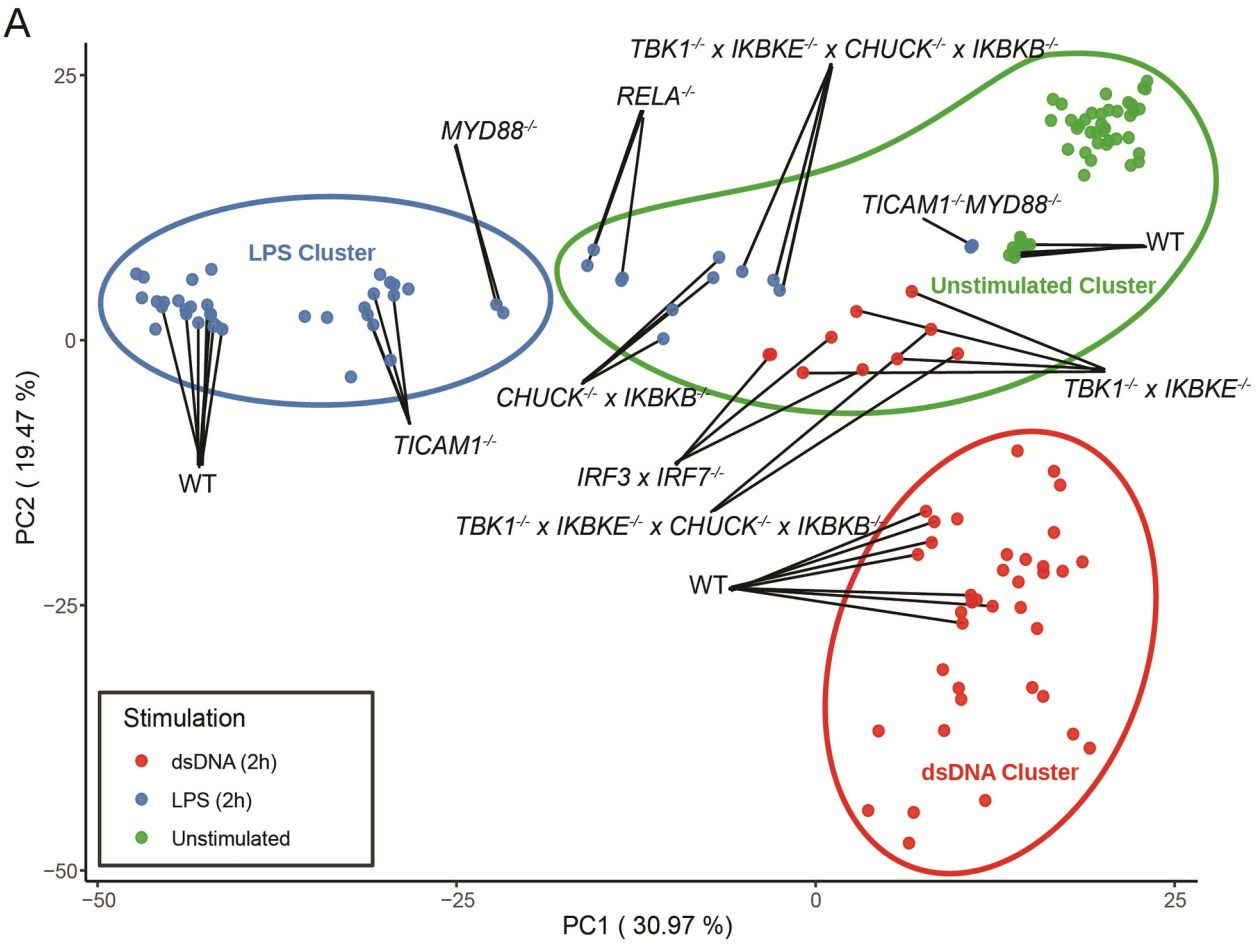

Figure 4.3.7 (**A**) PCA (based on top 1000 HVGs) plot showing 168 samples. Colors of samples show stimulation: LPS, dsDNA, or unstimulated. The colors of lines indicate three clusters found by the Louvain algorithm, which we conditionally called LPS cluster (blue), dsDNA cluster (red), and unstimulated cluster (green). Samples that are stimulated but have their essential pathways perturbed cluster with unstimulated samples.

Samples in the red cluster (dsDNA stimulated) exhibit a robust interferon signature that is abolished in *IRF3*$^{-/-}$ x *IRF7*$^{-/-}$ and *TBK1*$^{-/-}$ x *IKBKE*$^{-/-}$ double and *TBK1*$^{-/-}$ x *IKBKE*$^{-/-}$ x *CHUCK*$^{-/-}$ x *IKBKB*$^{-/-}$ quadruple knockouts (Fig. 4.3.7 A and B). cGAS-STING signaling mostly triggers the interferon response and shows only a relatively weak NF-κB dependent cytokine response.

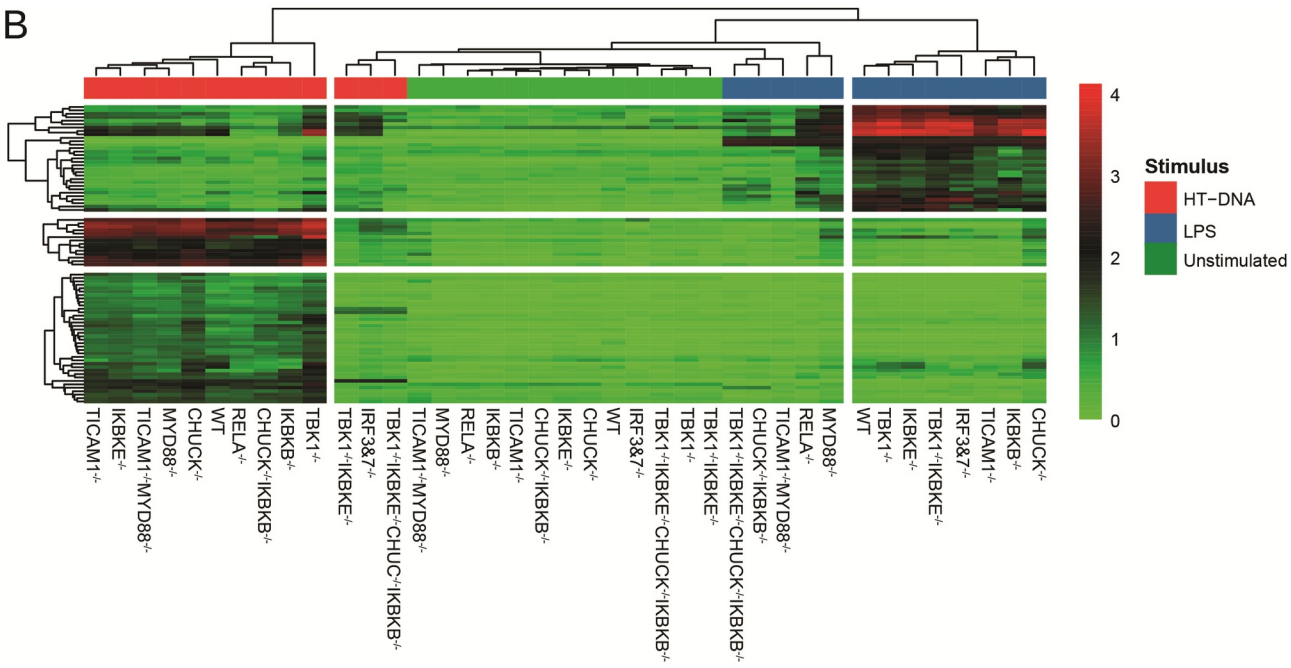

Figure 4.3.7. (**B**) Clustered heatmap based on top 100 HVGs. Both samples (columns) and genes (rows) are hierarchically clustered based on euclidean distance. Gene expression values averaged across replicates were used to build this heatmap. Based on the dendrogram, three main groups of samples are indicated by splitting the heatmap. Colored bars indicate treatment: dsDNA (red), LPS (blue), and unstimulated (green).

As expected, the blue cluster (LPS stimulation) exhibits a strong NF-κB dependent gene expression profile due to activation of the TLR4 receptor by LPS, however, in contrast to previously published work, we observed an unusually low interferon response. Despite this, we found several ISGs that showed a moderate up-regulation after two hours of LPS stimulation in NF-κB dependent manner, such as *CXCL10* (IP10), MX2, and CCL8. Conversely, other well-known ISGs like *IFIT1, IFIT2*, MX1, and *ISG15* showed a clear NF-κB independent expression profile. In addition, we also found several less well-known ISGs that were upregulated in a dsDNA-specific manner only, such as *OASL, EPSTI1, and HERC6* (Fig. 4.3.7 C). The TLR4 dependent induction of transcripts was interrupted in

RELA$^{-/-}$ and *MYD88*$^{-/-}$ x *TICAM1*$^{-/-}$ double knockouts while remaining functional in both *MYD88*$^{-/-}$ and *TICAM1*$^{-/-}$ single knockouts (Fig. 4.3.7 A and B).

The NF-κB pathway is also activated, albeit to a lesser degree, following dsDNA stimulation. Many highly expressed genes in the NF-κB regulated cluster also show a slight upregulation following dsDNA stimulation, for example, *CCL4L2*, *CCL4*, and *IL-1B* (Fig. 4.3.7 C).

In addition to observing the well-known elements of these two pathways, our data indicated that two main kinases are required for each signaling pathway. *TBK1* and *IKBKE* for the interferon pathway and *CHUCK* and *IKBKB for the* NF-κB pathway. When either pair was knocked out, it caused samples to cluster with unstimulated samples. Additionally, we observed that both of these pairs of kinases are redundant in their function for the two signaling pathways, meaning that a single knockout of either will not be enough to halt the signaling cascade. In other words, only when *TBK1* and *IKBKE* were depleted, a complete termination of antiviral gene expression was seen in the cGAS-STING signaling pathway. On the other hand, samples in which both kinases of the canonical IKK complex were knocked out (*CHUCK* (IKKα) and *IKBKB (*IKKβ)) had a completely interrupted TLR4 response and thus clustered together with the unstimulated cells. As expected, *TBK1* and *IKBKE* are only required for interferon signaling and have no role in the NF-κB pathway.

Similarly, *CHUCK* and *IKBKB* are redundant in their function for the NF-κB signaling pathway and do not affect interferon signaling (Fig. 4.3.7 C). We also noticed that NF-κB signaling was reduced in both *MYD88* and *TICAM1* knockouts but not completely abolished. Only in double-knockout samples of *MYD88* and *TICAM1* did we observe complete signaling ablation in gene expression (Fig.4.3.7. C) and sample clustering (Fig.4.3.7. A).

Another interesting notion from these studies is that the low but consistent NF-kB response observed downstream the cGAS-STING signaling pathway was also dependent on the canonical IKKs; *CHUCK* and *IKBKB*. Also, here, both kinases appeared to be redundant since only the deletion of both kinases displayed a full effect. However, in light of the rather weak NF-kB response observed under these conditions, additional studies are required to confirm these results.

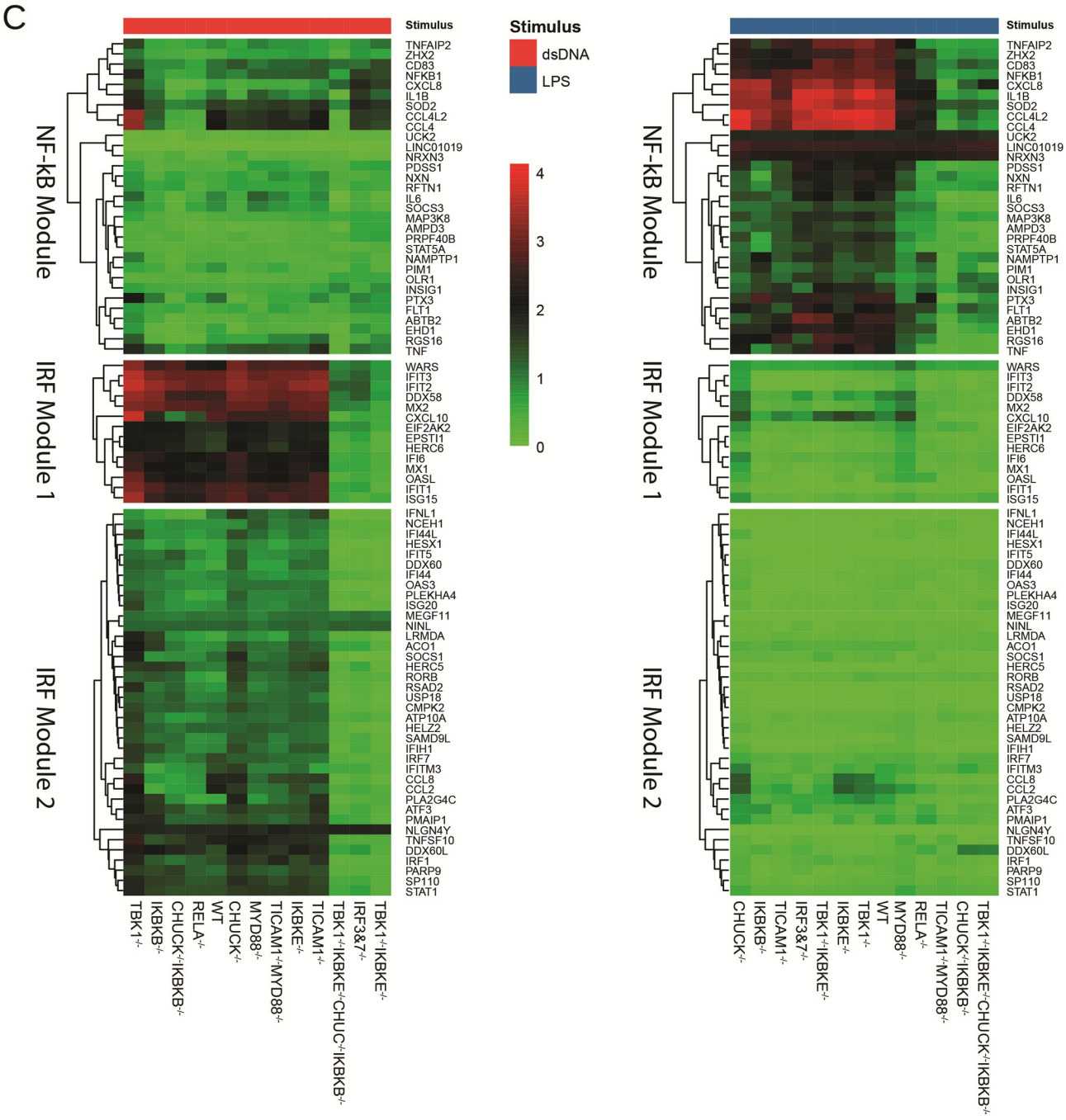

Figure 4.3.7. (**C**) Zoomed-in version of Figure 4.3.7.B heatmap. Shown are the three major gene expression modules, an NF-κB dependent module, and two IRF dependent modules. On the left: dsDNA stimulated samples. On the right: LPS stimulated samples.

4.4 Discussion

4.4.1 Single-cell RNA barcoding and sequencing

Single-cell RNA sequencing (scRNA-seq) has become increasingly more popular in recent years[26]. It has rapidly changed since the first single cells were sequenced by Tang et al. in 2009. A multitude of varying methods is now available that significantly differ in sensitivity and throughput[137,151]. To increase flexibility, reduce cost and throughput, we decided to set up a plate-based method compatible with flow cytometry. We chose SCRB-seq, a method with minimal setup requirements and low upfront cost that is easily expandable if higher throughput is required[30].

The original SCRB-seq protocol was described using 384-well microwell plates and included an RNA desiccation step after Proteinase K digest to remove the lysis buffer and heat-inactivate the enzymes. This was done to keep the RT reaction volumes as low as possible (2 µl per sample). These steps were not optimal for our setup as the low reaction volumes in increased well size caused unwanted evaporation, and through that, variation in RT efficiency. The original protocol also included a desiccation step right after lysis (which we omitted) as it seemed to be too damaging to the RNA.

Following the lysis buffer's optimization, omitting RNA desiccation, and increasing the RT volume, we observed significant improvements in our libraries' quality, both before and after sequencing (Fig.4.3.4 and 4.3.3). We attempted to improve our protocol even further using a newer reverse transcriptase, SuperScript IV (Life Technologies), that is used in many protocols[34,136,148]; unexpectedly, it did not provide any marked increase in the cDNA yield. This could be due to SuperScript IV being designed for speed and resistance against inhibitors and RNAse activity–all beneficial traits –but unnecessary in our situation. Also, Zucha et al. and Bagnoli et al. have shown that the Maxima H- reverse transcriptase's (used in our protocol) sensitivity is as good or better than SuperScrpit IV[48,152].

While we were setting up and testing our improved SCRB-seq protocol, another group published their own improvements to the same original method, which they called mcSCRB-seq. Our improved SCRB-seq and mcSCRB-seq protocols only differ in the lysis and RT steps. mcSCRB-seq still uses Proteinase K digest while omitting RNA desiccation in favor of heat inactivation. It incorporates polyethylene glycol (PEG) 8000 in the RT

reaction to achieve the molecular crowding effect. It has been shown that adding macromolecules increases enzymatic activity, possibly due to a reduction in effective reaction volume[153]. Bagnoli et al. displayed that this effect also increases the efficiency of scRNA-seq library preparation, but they used standardized input (UHRR) and mouse embryonic stem cell lines in their assessment while not considering very small cells with very low RNA concentration. In our hands, using mcSCRB-seq to sequence BLaER1 macrophages, failed to produce good quality libraries; we saw much more inferior sequencing results than with SCRB-seq v.3. The results contained more unmapped reads, a lower percentage of exonic reads, and cells failed to segregate based on the stimulus (Fig. 4.3.4 B). This could be because BLaER1 cells in their B cell and macrophage stage are rather small and contain relatively small amounts of RNA[154]. Hagemann-Jensen et al. found that B cells, along with T cells, are among the most challenging cells to capture in scRNA-seq[35].

Similar to the authors of the mcSCRB-seq paper, we also noted that increasing the PEG 8000 concentration above a certain threshold leads to unspecific amplification[48], which in our case, interfered with sequencing results (Fig. 4.3.4 A). Although we did not increase the PEG concentration over the recommended threshold, the amount of available template per reaction was much lower due to the cell type used.

We also noted a shorter average fragment length of mcSCRB-seq libraries, partly because of the increased unspecific amplification. It could also have been due to differences in the lysis step as mcSCRB-seq still incorporates Proteinase K digestion and heat inactivation. We noticed similarly shortened fragments in our SCRB-seq test where we had not yet optimized the lysis buffer. Proteinase K digest seems like a sound option to rid the samples of RNases, but the same can be achieved with adding RNase inhibitors. By skipping the digest, we save time and protect the samples from high temperatures (heat inactivation of Proteinase K).

Our improved SCRB-seq outperformed the original protocol. The libraries had high-quality cDNA as assessed by the Bioanalyzer, and the sequencing results displayed a high percentage of exonic reads while maintaining a low dropout rate. This success was also observed when applying our protocol to a relevant biological setting in BlaER1 macrophages stimulated with LPS. As already mentioned, we did not manage to produce the same high-quality libraries using the mcSCRB-seq protocol. Cells prepared using mcSCRB-seq did not show gene expression changes in an LPS dependent manner. From

the rest of the libraries, we found a robust inflammatory signal (Fig. 4.3.5). We were also able to integrate all the remaining libraries into a single experiment, thus showing that experiments from different sequencing runs can be combined[155]. Combining experiments increases the statistical power[156] and allows us to detect lowly expressed transcripts.

4.4.2 Low-input bulk RNA sequencing and barcoding

As the SCRB-seq protocol is entirely compatible with larger input of cells, we thought to test it as a low-input bulk RNA-seq method. This type of bulk RNA sequencing has many benefits. It allows for much higher throughput than regular bulk sequencing. Preparing dozens or even hundreds of samples in this fashion is no more difficult than preparing a single sample. This protocol controls PCR amplification bias due to early barcoding (UMIs on oligo(dT) primers)[157]. Because of pooling, it is also much more affordable than regular bulk RNA-seq. The wet-lab cost of a single low-input bulk sequencing sample is only marginally higher than for one cell in SCRB-seq protocol (the difference comes from RNA isolation/clean-up). The only real difference compared to scRNA-seq comes from the sequencing depth. For single cells, the most cost-effective sequencing depth seems to be between 20 000 – 200 000 reads[29,35,156], but for bulk samples, we could increase the sequencing depth up to a few million before saturation is reached. It is, of course, highly cell type-dependent, and test sequencing runs should be performed. Recommended sequencing depth would also depend on the aim of the project. For screening purposes, fewer reads would suffice.

The drawbacks of this bulk sequencing are the same as they are for SCRB-seq. Because of the protocol's setup, we end up with only 3' ends of transcripts. This is excellent for gene counting[158] and multiplexing of samples but cannot detect alternative splicing, single-nucleotide polymorphisms (SNPs), or other RNA changes that require a whole transcript.

The best uses for this method would be experiments that require a high number of samples (e.g., screens) or where cells are difficult to source and are expensive to cultures such as induced pluripotent stem cells (iPSCs) or small model organism cells. This approach could also be used in conjunction with scRNA-seq when higher sensitivity is required.

We found 50,000 to be the most optimal number of cells per sample. The handling volume was not too high, and the clean-up efficiency remained satisfactory. Increasing cell numbers did not yield higher cDNA amounts. RNA extraction using these cell numbers

yields enough cDNA to be measured on a Bioanalyzer system as an added, optional QC step. We would strongly suggest using robotic liquid handling systems and automated cell counting to speed up the RNA extraction, minimize the sample dropout and pipetting errors for smaller amounts of cells. Using the robotic liquid handlers would also dramatically speed up the whole process and increase throughput.

4.4.3 The role of IκB kinases in the TLR4 and cGAS-STING signaling pathways

As proof of concept, we prepared 160 low-input samples of 12 different knockout cell lines. After sequencing and QC, several samples dropped out, and the sequencing depth was quite uneven between samples. This was most likely caused by unequal numbers of input cells in the samples, as we lacked the capabilities to count cells in each sample separately. Instead, cell counts were estimated based on the number of cells plated at the beginning of the experiment. We noted that proliferation differs among clonal knockouts, most likely leading to the variation in our data. As such, this experiment would benefit from more accurate cell counting.

The selection of these KOs could allow us to reconstruct the cGAS-STING and TLR4 signaling networks in an unbiased way. While this dataset was rather limited by the number of perturbations (only 12), it was still considerably high throughput for bulk sequencing (160 samples), and in this regard, it could serve as a blueprint for even larger studies. Eventually, this kind of setup could be considered for large-scale perturbation studies.

The biological question was to find transcripts that show a clear NF-κB (*RELA*) and IRF3/7-dependent signal and to identify the relevant kinases upstream of these transcription factors. Despite the initial problems, the samples that passed QC clustered quite well based on the different stimuli used. We observed that samples with essential genes from their respective pathways deleted, clustered with unstimulated samples. For example, when the canonical IKKs (*CHUCK* and *IKBKB*) or a subunit of NF-κB (*RELA*), were knocked out in the TLR4 signaling pathway, the samples clustered with unstimulated clones. On the other hand, if either one of the TLR4 adaptors (*MYD88* or *TICAM1*) was perturbed, signaling was still intact, even though to a lower extent. Only when both adapter molecules were knocked out was the signaling completely terminated. Similarly, the cGAS-STING pathway was interrupted when both essential kinases (*IKBKE* and *TBK1*) or transcription factors (*IRF3* and *IRF7*) were knocked out.

We also found that *CHUCK* and *IKBKB* are redundant kinases in the TLR4 associated NF-

κB pathway and that the kinases *TBK1* and *IKBKE* are redundant in their function for the cGAS-STING pathway. The finding that *CHUCK* and *IKBKB* are redundant was a novel one since, so far, it has been shown in mouse studies that *IKBKB* alone is essential for the activation of the canonical NF-κB pathway[159,160]. Even more surprising was the finding that *TBK1* and *IKBKE* act redundantly in the cGAS-STING signaling pathway (interferon response), especially when considering that *TBK1* is required for interferon response induced by TLR4 signaling[161]. In this dataset, we can clearly see that Type I interferon response is not perturbed in either of the single knockouts (*TBK1* and *IKBKE*), but it is abolished in the double knockout samples (Fig. 4.3.7 C).

However, we could not confirm the requirement of *TBK1* for Type I interferon response after LPS stimulation (TICAM1 dependent interferon signal). This might be because the interferon response after LPS stimulation is delayed – the samples were only stimulated for two hours before harvesting and library preparation. We could still detect an interferon response after LPS stimulation, albeit a rather weak one (Fig. 4.3.7 C). Despite this, we identified several well-known ISGs[162,163] that were upregulated upon LPS stimulation, such as CXCL10, MX2, DDX58, PLA2G4C, CCL2, and CCL8 in most of the samples. Additionally, we also found ISGs that showed a clear NF-κB independent induction: IFIT1, IFIT2, IFIT3, IRF1, and OASL. These transcripts can be particularly useful as they allow studying interferon induction in an NF-κB independent manner. Knowledge about these transcripts is especially beneficial when studying pathways that simultaneously impact NF-κB and interferon signaling, e.g., in the context of genetic screens. Therefore, transcripts that can distinctly be assigned to either IRF or NF-κB activation are important for distinguishing between these pathways using gene expression as a read-out.

SCRB-seq can be used to conduct extensive knockout experiments involving multiple genotypes and multiple stimuli. The current difficulties lie with sample preparation and RNA extraction. Using low-input bulk SCRB-seq to detect major changes in gene expression is already practical. As shown by Svensson et al. and Heimberg et al., major changes in gene expression are resistant to noise, and the trade-off between sequencing depth and cost does not always justify deeper sequencing[117,164]. However, to find more subtle changes in a small number of genes, much deeper sequencing would be required. We would recommend performing calculations before sequencing to determine the exact number of samples vs. throughput or perform a small-scale test sequencing.

5 Human T helper cell differentiation on a single-cell and population level

5.1 Introduction

Human CD4$^+$ T helper cell differentiation has been investigated for several decades, and in many aspects, it is considered to be well understood. While basic principles such as distinct cytokine profiles and master transcription factors have been established, recent advances in methods have opened new avenues of inquiry. Some of the remaining questions include: how and when are cell fates decided following the activation, how do T helper subsets differ from each other on the single-cell level, and how does the transcriptional profile of differentiating T helper cells change over time[115,165].

Under physiological conditions, the T helper cell subsets arise based on location and pathogen exposure[64]. In vitro, we can recreate conditions to skew or polarize the development of naive T helper cells by using different cytokines and antibodies, together with the initial activation and proliferation signal[166] (Fig. 5.1). It is also possible to activate the CD4$^+$ T cells without any additional cytokine skewing and let them proliferate and acquire a phenotype spontaneously. These cells are called T_H0, and are sometimes considered to still retain their multipotentiality (they can still acquire T_H1 or T_H2 characteristics)[167] (Fig. 5.1).

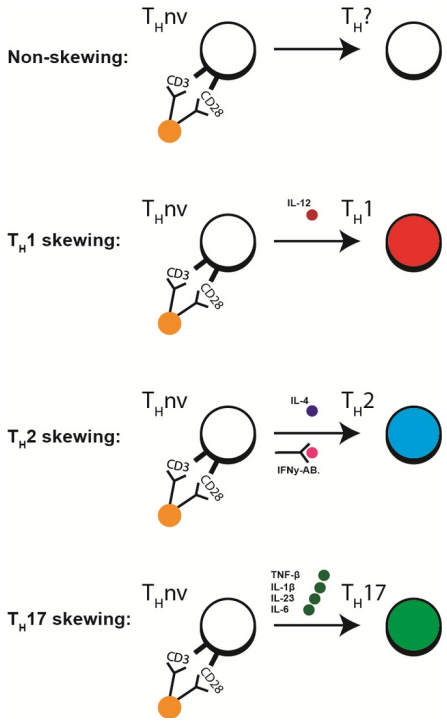

Figure 5.1 In-vitro CD4$^+$ T helper cell differentiation. Adding specific cytokines during the activation process makes it possible to skew T helper cell development towards the desired lineage. Omitting cytokines under the non-skewing condition causes T helper cells to commit towards T_H1 or T_H2 fate spontaneously.

T helper cell lineage is usually assayed by measuring cytokine production and release either by ELISA or flow cytometry. High levels of IFN-γ characterizes T_H1 and T_H17 T helper cells, whereas T_H2 cells produce less predominantly IL-4. As mentioned previously, activated naive T helper cells spontaneously acquire either T_H1 or T_H2 (rarely T_H17) characteristics. Although T helper cells can be influenced through both an autocrine and paracrine fashion, it is still possible that a population of non-skewed differentiated T helper cells contain both T_H1 and T_H2 like cells.

5.2 Overview

To find out more about cell fate decisions on the single-cell level, we employed SCRB-seq to capture and sequence non-polarized but activated $CD4^+$ T helper cells at several different time points. While preparing for the experiment, we also tried to optimize our experimental design to mitigate known technical variations that can affect plate-based methods and give rise to batch effects. In analyzing this first set of single-cell data, we were unable to detect discrete subsets of differentiated T helper cells on a single-cell level.
We reasoned that one issue hindering us from detecting T helper subsets was the inherently low expression levels of many lineage-specific transcripts, such as cytokines and transcription factors. To compensate for that we repeated this experiment with restimulating the cells before cell capture as well as with increased differentiation time. Because of these modifications, we noticed a marked increase in captured transcripts. However, we could still not verify any distinct T helper subsets.
In order to validate our scRNA-seq findings and to rule out low sensitivity as a factor, we conducted a third experiment. For that we set up a low-input bulk sequencing pipeline to generate, assay, and sequence T helper cells. As positive controls, we also prepared knockout naive T helper cells deficient for TBX21 or GATA3 to force the lineage commitment towards T_H2 or T_H1, respectively. Thanks to bulk sequencing's increased sensitivity, we were able to separate control samples into distinct T helper types, but only based on their cytokine expression. However, this distinction vanished when comparing the whole transcriptome of those samples.

5.3 Results

5.3.1 Establishing in vitro differentiation of T helper cells

To determine whether T helper differentiation under non-skewing conditions in vitro indeed leads to spontaneously differentiated T helper cells, we cultured naive $CD4^+$ T cells under non-skewing conditions (see methods 3.2.3). As positive controls, we added IL-12 to induce T_H1 skewing, IL-4 and IFN-γ neutralizing antibodies to induce T_H2 skewing, and IL-1β to induce T_H17 skewing. After 14 days of differentiation, we confirmed the successful differentiation of T helper cells on a population level by ELISA. As expected, IFN-γ production was mainly seen for T_H1 and T_H17 skewing conditions, while considerable amounts of IL-4 were only seen for T_H2 skewing conditions. Non-skewing conditions led to comparable levels of IFN-γ and IL-4 production as observed in T_H2 or T_H1 skewing conditions, respectively (Fig. 5.3.1. A).

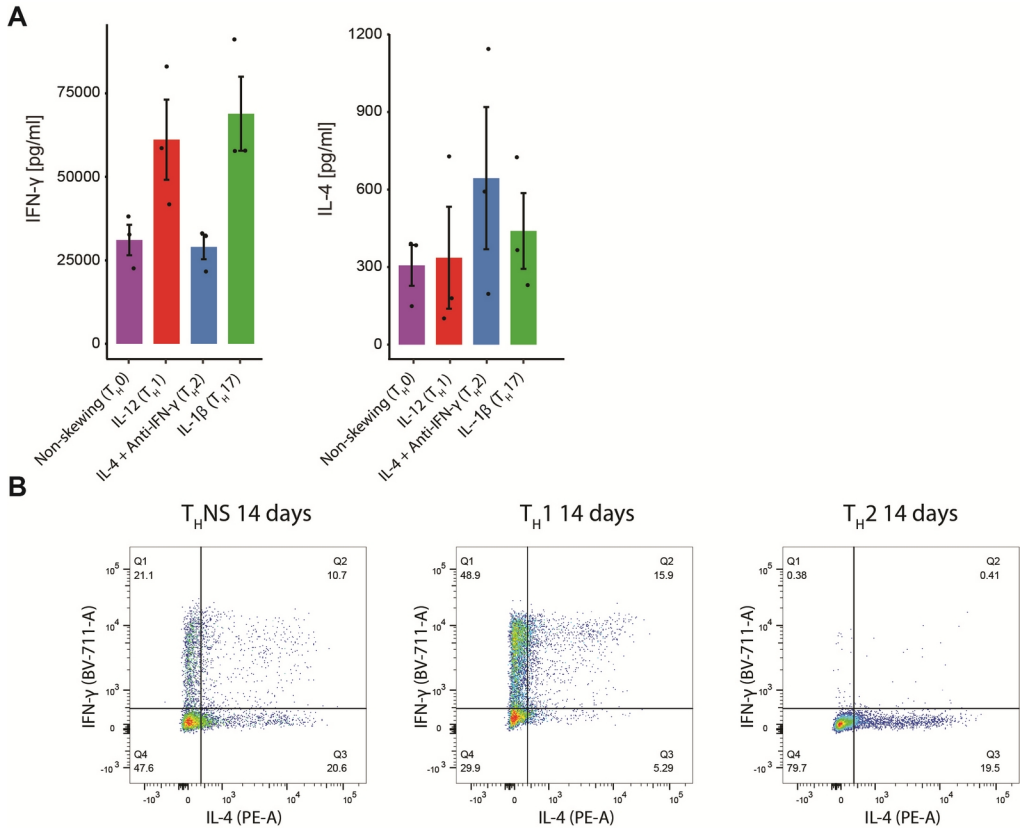

Figure 5.3.1 (**A**) 200 000 naive T helper cells were plated per sample, activated, and polarized as in Figure 5.1. After 14 days, the cells were stimulated for 24h with PMA (50 ng/ml) + Ionomycin (1µg/ml), after which the supernatant was collected, and the release of IFN-γ and IL-4 was measured by ELISA. Graph shows mean +/- SEM of three donors. (**B**) 200 000 naive T helper cells were plated per sample, activated, and polarized towards TH0, TH1, and TH2 lineage. After 14 days, the cells were stimulated for two hours with PMA (50 ng/ml) + Ionomycin (1 µg/ml), after which the intracellular cytokine staining was performed as

described in methods 3.2.11. One of three biological replicates per condition is shown.

The disadvantage of bulk measurement techniques such as ELISA is that one cannot draw any conclusions with regards to the proportion of cells that commit to a particular T_H phenotype of interest. However, this problem can be overcome by assessing cytokine production on a single-cell level using flow cytometry. Conducting such analyses provided a far more dichotomous picture.

We repeated the previous experiment, omitting the T_H17 condition this time, as it did not produce measurable levels of IL-17 (data not shown). T_H1 skewing conditions induced a sizable proportion of IFN-γ single-positive cells, a fraction of IFN-γ/IL-4 double-positive cells, and only a small proportion of IL-4 positive cells. On the other hand, T_H2 skewing conditions led to a single population of IL-4 positive cells. Interestingly, non-skewing conditions resulted in an equal proportion of IFN-γ single positive and IL-4 single-positive cells (both around 20%) and an additional population of IFN-γ/IL4 double-positive cells (around 10%). So in total, approximately half of the T cells population displayed signs of lineage commitment under non-skewing conditions, with a roughly equal share of cells of a T_H1 and T_H2 profile (Fig.5.3.1 B). We, therefore, opted to use this non-skewing setup to perform a scRNA-seq experiment to further characterize the spontaneous differentiation of T cells under non-skewing conditions.

5.3.2 SCRB-seq causes strong batch effects that need to be corrected in downstream analysis

To identify novel transcripts responsible for T helper cell differentiation, we obtained PBMCs from two different donors and purified naive $CD4^+$ T helper cells in two different ways. We used magnetic assisted cell sorting or MACS kit from Miltenyi Biotec that yields $CD4^+$ T cells with roughly 90% purity (as stated by the manufacturer) or FACS sorting for $CD3^+$, $CD4^+$, $CCR7^+$, $CD45RA^+$ positive T cells. FACS sorting should result in much higher purity (well over 95%[168]) of naive $CD4^+$ T cells than MACS sorting. We reasoned that this increased purity ensures that the $CD4^+$ T cell differentiation would not be affected by potential antigen-presenting cells or influenced by contaminating memory T cells leftover from PBMCs isolation.

Both samples were plated and activated, then cultured as described in methods, chapter 3.2.3. Cells were harvested at zero, four, eight hours, and one, three, and seven days after stimulation, then sorted into a 96-well PCR plate pre-filled with lysis buffer. These plates

were then flash-frozen on dry ice. At each time point, two plates were sorted for sequencing (a total of 2256 single cells). Library preparation started one day after the final time-point and followed the SCRB-seq v.3 protocol. As batch effects are a known problem for plate-based single-cell sequencing protocols, we randomized the order of plates to minimize the batch effect between plates[169,170]. Plates were prepared two at a time (Table 5.3.1) until Nextera XT library preparation step. Starting from this point, all samples were handled simultaneously. Samples were sequenced on Illumina HiSeq 1500 platform, with each experiment on a separate lane.

zUMIs software was used for demultiplexing, mapping, and feature counting. A total of 2109 out of 2256 cells passed the QC. Standard Seurat workflow was followed until dimensionality reduction and clustering. In Figure 5.3.2 A, batches corresponding to capture plates are visible on a UMAP plot. Also, earlier time points seem to be more prone to technical variation. The later time points (three days and seven days) already grouped with each other even without any batch effect correction but could still be separated based on which lane the cells were sequenced. It is, therefore, vital to correct for batch effects before continuing with the analysis. We used Seurat's SCTransform[155] function to regress out the unwanted variation.

This specific experimental setup resulted in three distinct sources of technical variation: The first being MACS purified cells in lane-1 compared to the FACS sorted cells in lane-2 of the Illumina HiSeq flow cell (batch 1). The second was originating from the pairs of plates prepared together (batch 2). Lastly, each plate corresponds to a time-point and a donor (batch 3)(Table 5.3.2).

Table 5.3.2 Experimental design. Overview and order of plates from both experiments. Time shows how long after the activation the cells were harvested. Randomized order of library preparation indicated by "Prep.order". Batch 1 indicates pairs of plates that were handled together during library preparation. Batch 2 indicates each plate. Batch 3 indicates different donors and sequencing lanes.

Experiment 1 MACS purified naive CD4⁺ T-cells							
Time	**Method**	**Plate #**		**Prep.Order**	**Batch 1**	**Batch 2**	**Batch 3**
0h	MACS	1		4h-1	a	a	a
0h	MACS	2		7d-2	a	b	a
4h	MACS	3		8h-2	b	c	a
4h	MACS	4		24h-1	b	d	a
8h	MACS	5		0h-1	c	e	a
8h	MACS	6		4h-2	c	f	a
24h	MACS	7		7d-1	d	g	a
24h	MACS	8		3d-1	d	h	a
3 days	MACS	9		24h-2	e	I	a
3 days	MACS	10		8h-1	e	j	a
7 days	MACS	11		3d-2	f	k	a
7 days	MACS	12		0h-2	f	I	a

Experiment 2 FACS sorted naive CD4⁺ T-cells							
Time	**Method**	**Plate #**		**Prep.Order**	**Batch 1**	**Batch 2**	**Batch 3**
0h	FACS	13		0h-1	g	m	b
0h	FACS	14		4h-2	g	n	b
4h	FACS	15		8h-2	h	o	b
4h	FACS	16		24h-1	h	p	b
8h	FACS	17		4h-1	I	q	b
8h	FACS	18		7d-2	I	r	b
24h	FACS	19		7d-1	j	s	b
24h	FACS	20		8h-1	j	t	b
3 days	FACS	21		24h-2	k	u	b
2 days	FACS	22		3d-1	k	v	b
7 days	FACS	23		3d-2	I	w	b
7 days	FACS	24		0h-2	I	x	b

A clear indication of the batch effect in our data was the distinct clusters that cells formed in UMAP plots. We can see that batches are dictated mainly by the capture and RT plates

of origin when we overlay the plot with index sort information (Fig. 5.3.2 A). Each small grouping of cells is one of the two plates per time point and per donor. Also visible is the separation by a donor that would not be expected to such a degree. We can correct for batch effects by trying to regress out the confounding variables listed above.

When controlling for batch 1 (donors and lanes), we did not notice any major improvements in the cells' positioning (Fig. 5.3.2 B). When using batch 2 as the confounding variable, we noted quite an improvement in cells' positioning. However, some time points (especially the earlier ones) of the same donor remained separated (Fig. 5.3.2 C). Only when we corrected for each capture plate (batch 3) did we finally lose the distinct clustering seen in previous plots (Fig. 5.3.2 D). This somewhat surprising finding implies that our SCRB-seq protocol causes the most technical variation in the earliest steps of the library preparation; cell capture and lysis.

For further analysis, batch corrected data where individual plates are considered confounding variables were used unless stated otherwise.

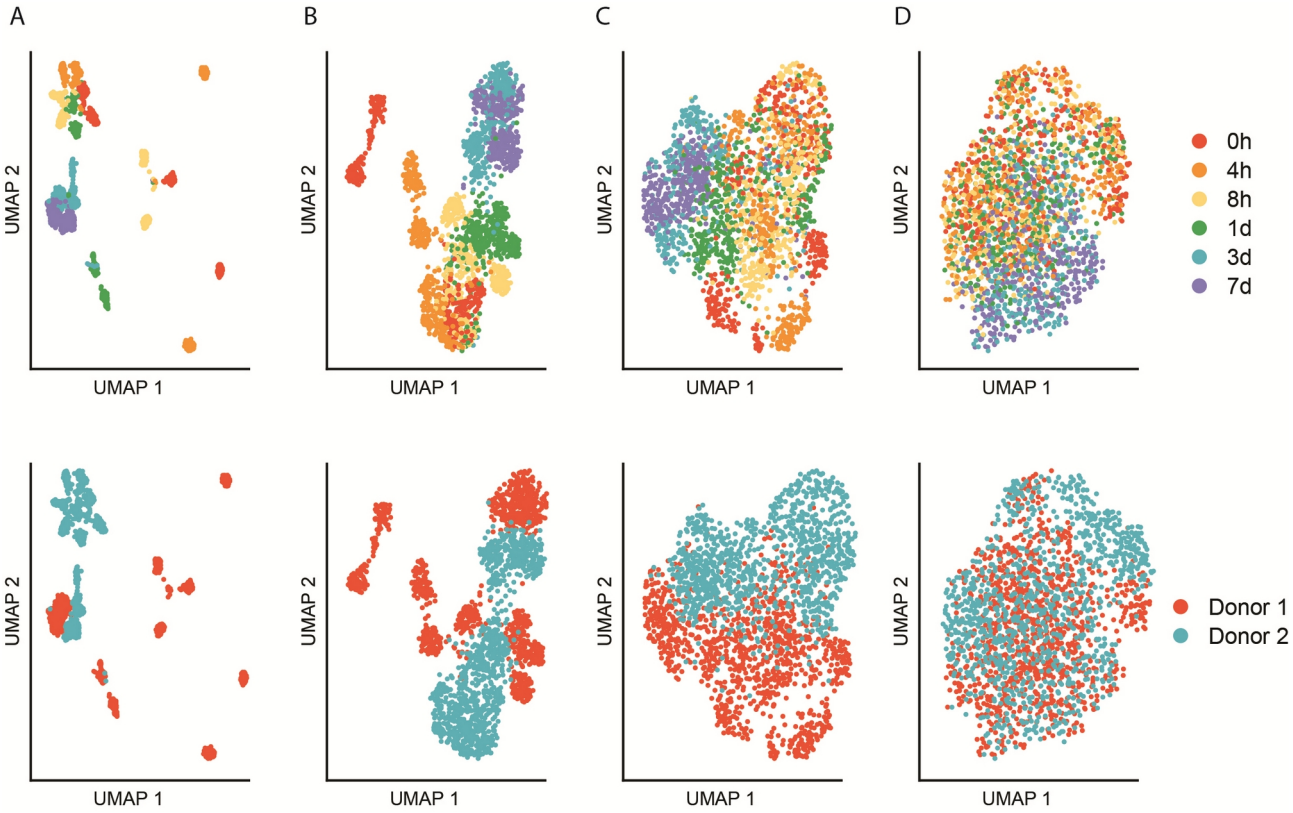

Figure 5.3.2 UMAP plots of scRNA-seq data. SCRB-seq has a strong batch effect caused by differences in cell capture, lysis, RT, and library preparation. (**A**) uncorrected data, (**B**) sequencing lane/donor corrected data, (**C**) handling batch corrected data (**D**) individual plates corrected data. Upper panels are colored by time after activation, and lower panels are colored by the donor. All UMAPs are based on the top 3000 HVGs.

5.3.3 CD4$^+$ T helper cells differentiated in-vitro and under non-skewing condition yield a heterogeneous and continuous population

Our data revealed that CD4$^+$ T helper cells form one heterogeneous and continuous population with no distinct sub-populations. When overlaying the UMAP plot with capture time information, we can see that the primary source of variation is the differentiation time (Fig. 5.3.2 D, upper panel). Differentiated cells do not group into smaller, distinct clusters. If anything, the cells from earlier time points exhibit more variation between themselves and between donors and tend to form small sub-populations because of that. One possible explanation could be that naive CD4$^+$ T cells are physically much smaller than activated and differentiating cells. Larger cells usually contain more mRNA[154]. We could find indications of that in the uncorrected data, where earlier time points formed distinct clusters based on their plate of origin, the later time points mixed better. Also, cells from later time points had more reads associated with them (data not shown).

UMAP projection of differentiated cells does not reveal groupings that could be attributed to any of the CD4$^+$ T helper cell subsets. Unbiased clustering (Louvain algorithm based on SNN as implemented by Seurat's workflow) failed to identify any meaningful subsets in our data. Depending on the parameters used, we see either two main clusters, one early group (mainly comprising of cells from zero, four, eight hour, and one day post-activation) and a late group of cells (mostly one, three, and seven days post activation) or many, smaller but biologically meaningless clusters (data not shown).

When looking at the expression of lineage-specific cytokines, we found that many of the same cells that express high levels of the T_H1 defining cytokine *IFN-γ* also co-express T_H2 specific cytokines such as *IL-4* and *IL-13*. We did notice several cells that only expressed *IL-4* or *IL-13* and no *IFN-γ*, but they were randomly distributed and formed no clusters with other similar expression pattern cells. This could be due to the very low expression levels of those cytokines, where only a single read, or two, would qualify those cells as highly expressing either *IL-4,* or *IL-13*, or both. Expression of lineage-specific transcription factors reflected the same image as did cytokine expression. There was a time-dependent expression gradient in *GATA3,* and *TBX21* levels; other than that, no region or a cluster of exclusive expression were detected (Fig. 5.3.3 A).

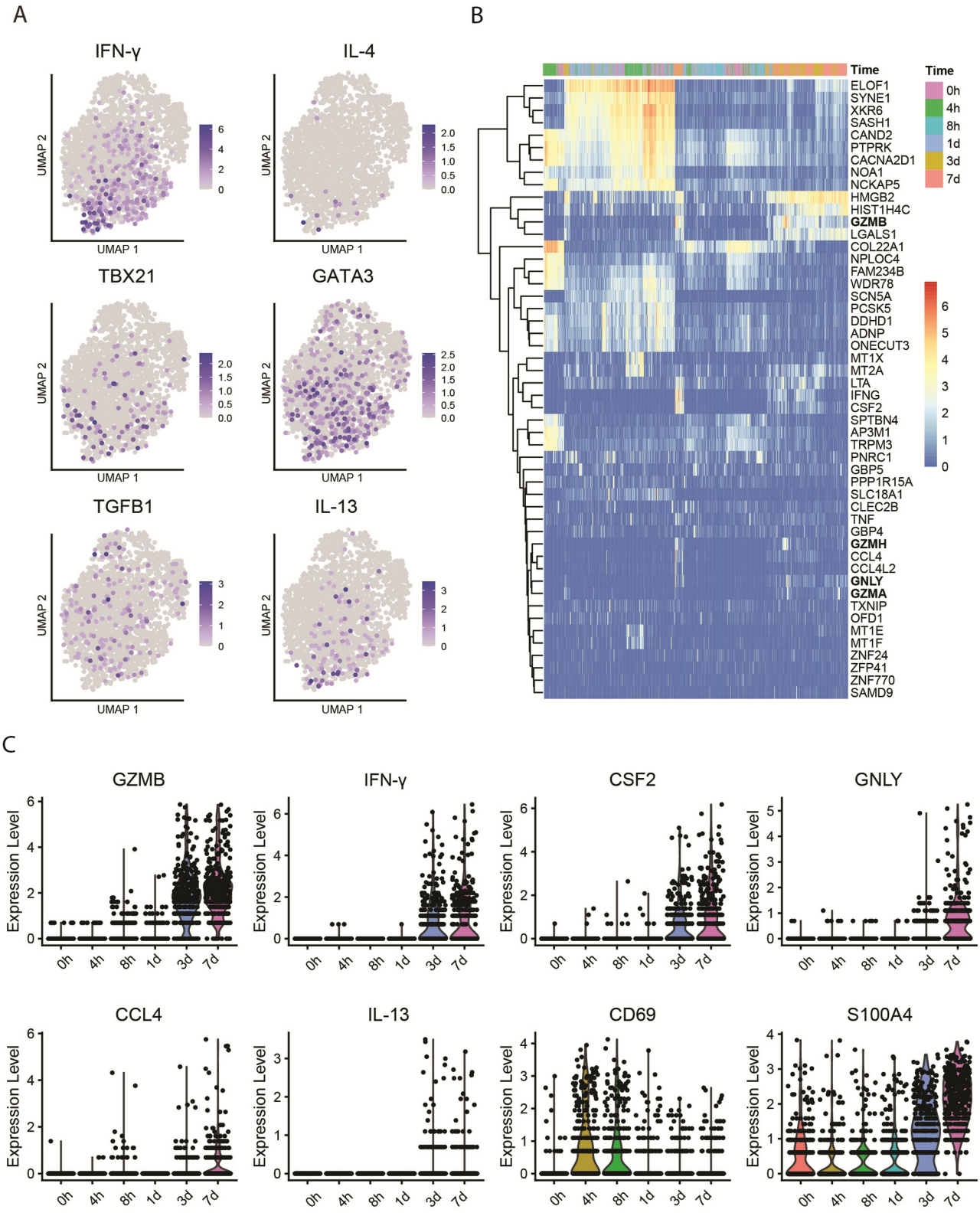

Figure 5.3.3 (**A**) Gene expression plots in UMAP projection. The positioning of cells is identical to Figure 5.3.2.D UMAPS are based on top 3000 HVGs. (**B**) Heatmap based on top 50 HVGs. Based on these genes' expression, cells from the earlier time points cluster together, as do cells from later time points. Cytotoxic marker genes are highlighted. (**C**) Violin plots of highly expressed cytokines, chemokines, and activation markers.

Highly variable genes (HVGs) (Fig. 5.3.3 B) contained many cytotoxicity markers like granzyme A, B, and H (*GZMA*, *GZMB*, *GZMH*) as well as granulysin (*GNLY*). These genes are typically associated with $CD8^+$ or cytotoxic T cells[171]. Still, in our experimental conditions, almost all of the differentiated $CD4^+$ T cells from day three onward express high levels of *GZMB*, and many of them also co-express *GNLY* and *GZMH* as well as *GZMA*.
If we consider these HVGs as markers for differentiation, we observe that the cells began differentiating around the third day. The expression of those markers was still increasing on day seven. Additionally, *CD69*, a classical T cell activation marker, started showing down-regulation around day one, but it did not quite reach the expression levels it had in naive T cells by day seven (Fig. 5.3.3 C). Taken together, this suggests that T helper cells activated and expanded under non-skewing conditions need around three days to start showing differentiation effects but probably need longer than seven days for final lineage commitment.

5.3.4 T helper cells differ in their early and late response to activation

As we saw in the previous chapter, there seems to be a break in our data where T helper cells go from reacting to an activation signal to increasing the expression of cytotoxicity markers and certain lineage-specific cytokines. In short, it looks like there are two different processes going on. To determine what transcriptional changes occur at those points, we decided to use our SCRB-seq protocol's full capabilities. Since we had information of when and from which plate the cells originated, we could directly compare selected time points.
Looking at genes differentially expressed (DE) between four hours after activation vs. baseline (zero hours), we detected many immune response and, more specifically, T cell activation markers like *CD69, TNF, LTA, TRAC, CD3D, NFKBID* (Fig. 5.3.4 A). To validate our findings, we conducted pathway enrichment analysis using the PANTHER Pathway tool. Doing so, we found that many gene ontology (GO) terms related to immune response, gene expression, and response to stimulus were significantly enriched (Fig. 5.3.4 B).

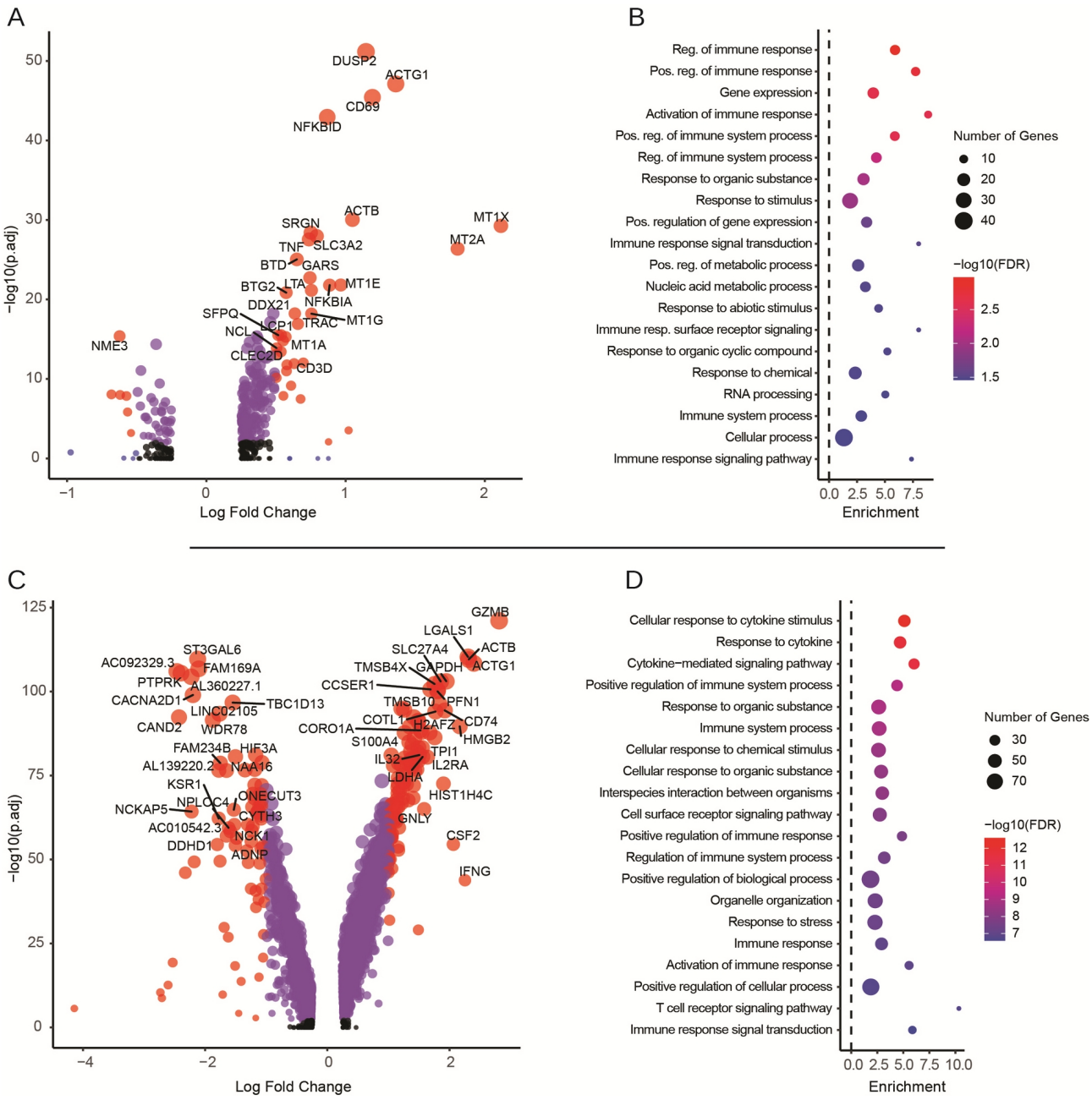

Figure 5.3.4 (**A**) A volcano plot showing DE results between cells from four hours vs. zero hours post-activation. (**B**) Pathway enrichment analysis based on all upregulated genes with FDR < 1% from (A). (**C**) A volcano plot showing DE results between cells from seven days vs. zero hours post-activation. (**D**) Pathway enrichment analysis based on up-regulated (LFC > 1) genes with FDR < 1% from (C).

We also conducted the same analysis between seven days and baseline (zero hours). This time, we discovered that many upregulated transcripts were part of the top HVGs, thus confirming that most of the variation in our data comes from later time points. The differentiating T helper cells show strong up-regulation of many cytokines and granzymes

such as *TNF*, *GZMB*, *CSF2*, *IFN-γ*, as well as well-known differentiation markers such as the IL-2 receptor (*CD25*) and *S1004A*. Also, the GO analysis revealed different enriched pathways compared to earlier results. GO terms like cytokine stimulus and chemical signaling dominate the top of the list. Unlike the activated T cells, differentiated T cells show a substantial down-regulation in the expression of many transcripts, including transcription factors such as *ONECUT3* and *HIF3A*. We also performed the pathway enrichment analysis on this set, but no pathways with an FDR of 10% or less were retrieved as a hit. This is not an unexpected result insofar as many down-regulated transcripts in that list are pseudogenes (*AC092329.3, AL360227.1, AL139220.2*).

5.3.5 Restimulating $CD4^+$ T helper cells

When trying to assay $CD4^+$ effector T cells with conventional methods like ELISA and flow cytometry, it is necessary to stimulate the cells to increase cytokine production to a level where it becomes detectable (Fig. 5.3.5 A). This stimulation, also called restimulation, can be achieved by adding anti-CD28/anti-CD3 ABs or by using chemical compounds that mimic TCR engagement. One well-established stimulation procedure is the use of PMA and Ionomycin. PMA (phorbol-12-myristate-13-acetate) can activate protein kinase C (an essential part of TCR signaling[172]), and Ionomycin functions as a calcium (Ca^{2+}) ionophore that leads to an increased intracellular Ca^{2+} concentration. Together they mimic the TCR activation signal and lead to increased cytokine expression[173]. Although short-term restimulation with PMA/Ionomycin does not seem to affect T cell viability[174], over time, cells will succumb to activation-induced cell death (AICD)[175], thus impacting their gene expression profile. Since RNA-seq is much more sensitive than any other conventional assays, we initially omitted restimulation before cell capture from our first scRNA-seq experiment. However, since we could not detect different T helper types, we reasoned that maybe PMA/Ionomycin stimulation might help us capture more lowly expressed genes, which might help us in assigning cell types. FACS purified naive T cells from two donors were plated and activated. During the harvesting of cells at four different time points, we restimulated half of them with PMA and Ionomycin (see methods 3.2.3) (Table 5.3.5).

Table 5.3.5 Experimental setup. Overview and order of plates from both experiments. Time shows how long after the activation, the cells were harvested. Method – naive CD4$^+$ T cell isolation method. “Prep.order” - randomized order of library preparation to reduce the batch effect. Batch 1 indicates each plate. Batch 2 indicates pairs of plates that were handled together during library preparation.

Time	Method	Donor	Plate #	Stimulus		Prep.Order	Batch 1	Batch 2
0h	FACS	Donor 3	1	na		D3_14d_7	a	a
0h	FACS	Donor 3	2	na		D4_0h_1	b	a
0h	FACS	Donor 4	3	PMA/iono 2h		D4_7d_6	c	b
0h	FACS	Donor 4	4	PMA/iono 2h		D3_7d_5	d	b
3 days	FACS	Donor 3	5	na		D4_14d_8	e	c
3 days	FACS	Donor 3	6	na		D3_0h_1	f	c
3 days	FACS	Donor 4	7	PMA/iono 2h		D3_7d_6	g	d
3 days	FACS	Donor 4	8	PMA/iono 2h		D4_3d_4	h	d
7 days	FACS	Donor 3	9	na		D4_7d_5	I	e
7 days	FACS	Donor 3	10	na		D3_14d_8	j	e
7 days	FACS	Donor 4	11	PMA/iono 2h		D4_0h_2	k	f
7 days	FACS	Donor 4	12	PMA/iono 2h		D3_3d_3	l	f
14 days	FACS	Donor 3	13	na		D3_3d_4	m	g
14 days	FACS	Donor 3	14	na		D4_3d_3	n	g
14 days	FACS	Donor 4	15	PMA/iono 2h		D3_0h_2	o	h
14 days	FACS	Donor 4	16	PMA/iono 2h		D4_14d_7	p	h

We subjected 94 cells from each donor from each time point and stimulation condition to single-cell sequencing (total of 1504 cells). 1397 single-cells passed the QC and filtering. We also noticed some batch effect in this dataset, but it was less pronounced, primarily because the cells from later time points are less affected by technical noise; differentiated cells are physically bigger and transcriptionally more active (Fig 5.3.5 C). After correcting for the capture plate again, the two main populations of cells remained: activated but resting T cells and restimulated T cells (Fig. 5.3.5 B). Restimulated cells and cells from later time points have more reads associated with them, translating into more UMIs and transcripts (Fig. 5.3.5 C).

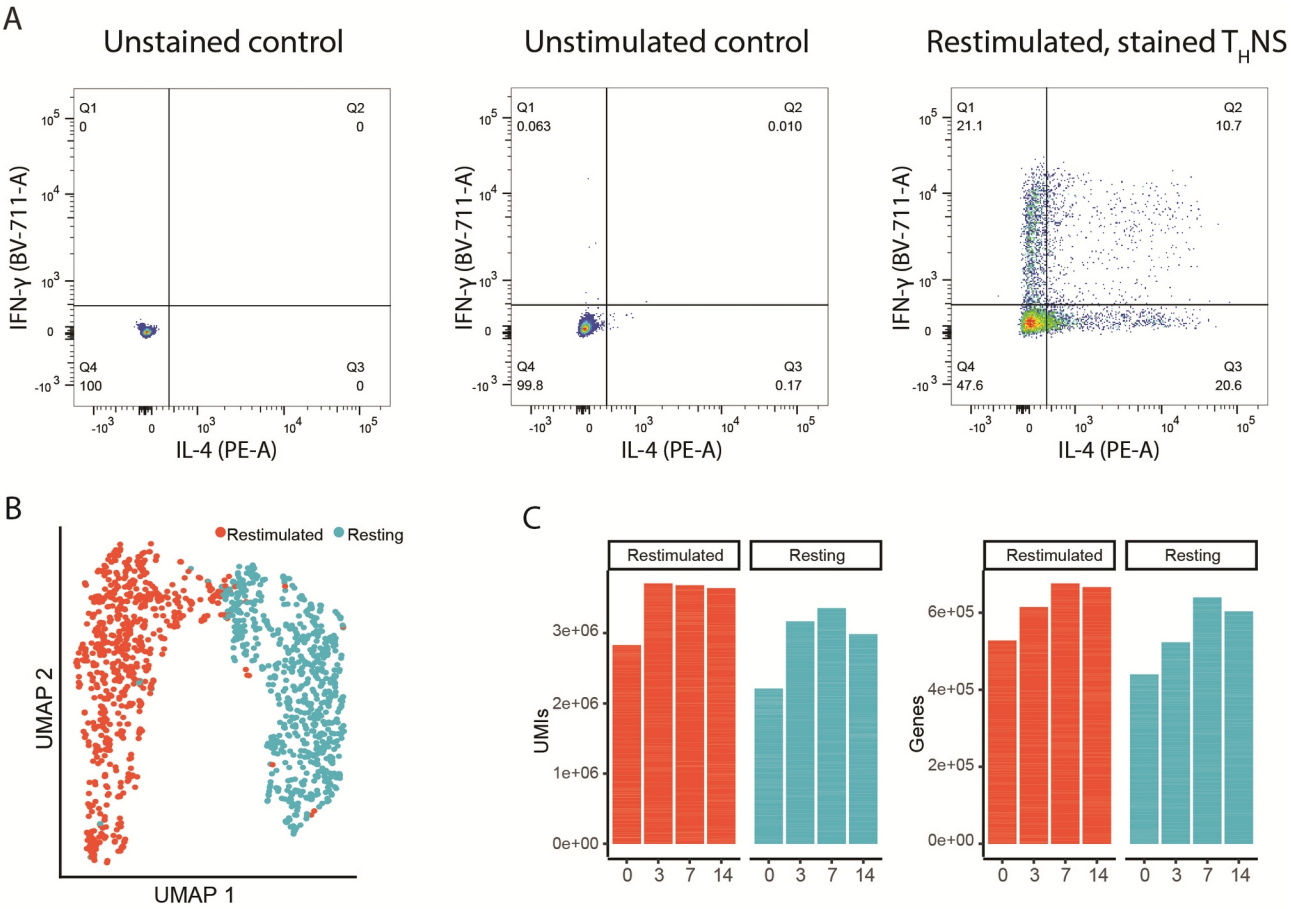

Figure 5.3.5 (**A**) FACS-based detection of hallmark T helper cell cytokines requires restimulation.

200 000 naive T helper cells were plated per sample and activated. After 14 days, the cells were either left unstained, unstimulated, but stained, or restimulated and stained. One of three biological replicates per condition is shown. (**B**) UMAP plot of 1397 single cells from two different donors and four time points. Cells are color-coded based on the PMA/Ionomycin restimulation. (**C**) Barplots showing the mean of UMIs and genes per cell detected at each given time point in restimulated vs. resting cells. Each time point had the same number of input cells, and libraries were pooled in an equimolar fashion for sequencing.

5.3.6 Restimulation increases CD4$^+$ T helper cell transcriptional activity and enhances the expression survival and activation markers

To control for PMA/Ionomycin restimulation effects, we conducted DE analysis comparing all the restimulated cells to all resting cells from each time point. As expected, we saw up-regulation in T cell activation marker genes like *CD69* and *IL-2* and pro-survival markers such as *MYC*, *PIM3*, and *TNF* (Fig. 5.3.6 A). Pathway enrichment analysis based on all the upregulated (log fold change – LFC>1 and FDR<1%) genes also revealed cell death-related terms like regulation of apoptotic process and regulation of cell death (Fig. 5.3.6

B).

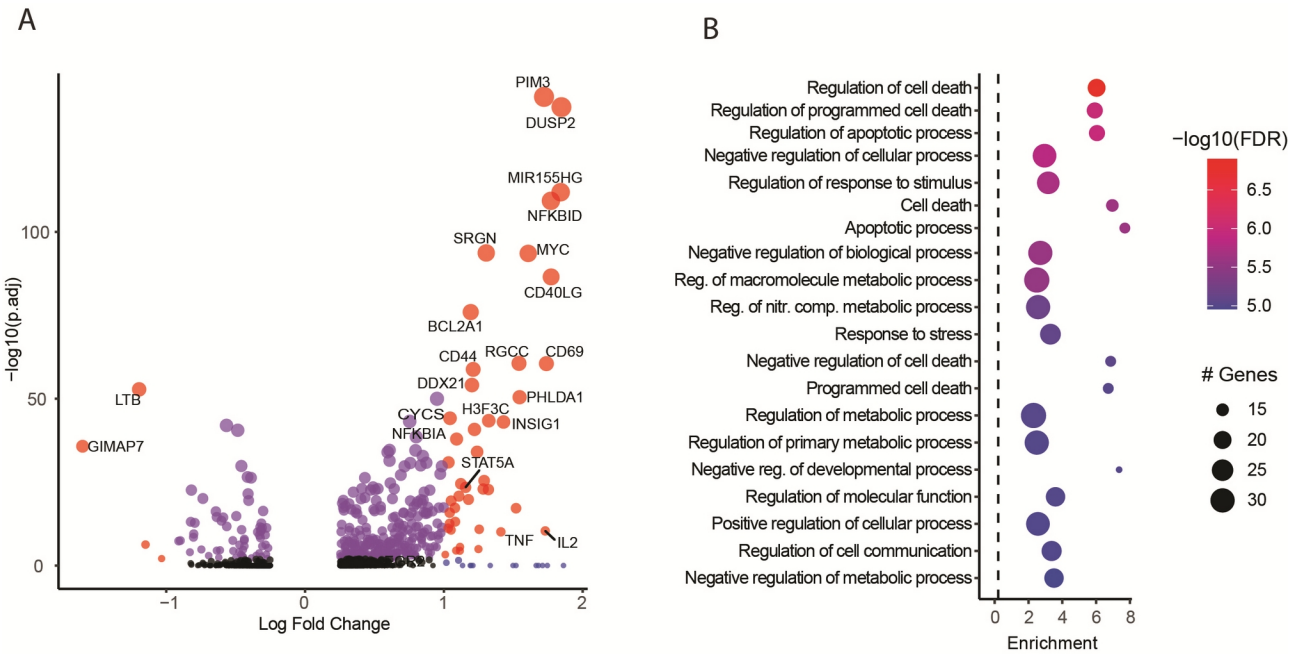

Figure 5.3.6 (**A**) Volcano plot showing DE results between restimulated and resting cells. (**B**) Pathway enrichment analysis based on up-regulated genes (LFC > 1, FDR < 1%) from (A).

5.3.7 Cytokine response is increased in restimulated CD4⁺ T helper cells

As mentioned above, PMA/Ionomycin stimulation increased transcription and caused up-regulation of pro-survival markers and activation markers. In general, many genes were more highly expressed in the restimulated group. Against our initial expectation, increased transcription still did not reveal any distinct T helper cell subsets, as we had previously observed them in our FACS-based analysis. We still detected a clear time-dependent expression of cytokines like *IL-4* and *INF-γ* and cytotoxicity markers (*GZMB*, *GZMA*), indicating T helper cell differentiation. Similar to the previous dataset, we again noticed that the same cells (or group of cells) that express high levels of *INF-γ* (T_H1 specific cytokine) also expressed high levels of *IL-13* (T_H2 specific cytokine) (Fig. 5.3.7. A).

We could not detect any additional meaningful clusters in our data other than the two already mentioned, the clustering based on PMA/Ionomycin stimulation and the gradient of differentiation time. Louvain clustering based on the SNN algorithm showed two main clusters in restimulated cells and three in resting cells, roughly corresponding to an early-

to-mid and a late cluster (Fig.5.3.7.B). Hierarchical clustering based on top 50 HVGs revealed that cells grouped mostly based on time after activation and less based on restimulation (Fig.5.3.7 C).

Using our scRNA-seq data, we confirmed that PMA/Ionomycin restimulation increased cytokine expression. Cells show restimulation effects, mostly indicated by the expression of cell survival and activation markers such as *MYC*, *PIM3*, *IL-2*, and increased overall transcriptional activity. The expression of T helper cell hallmark cytokines is also increased compared to resting cells. However, like the resting cell population, the restimulated cells also fail to segregate into distinct T helper cells' subsets (Fig. 5.3.7 B).

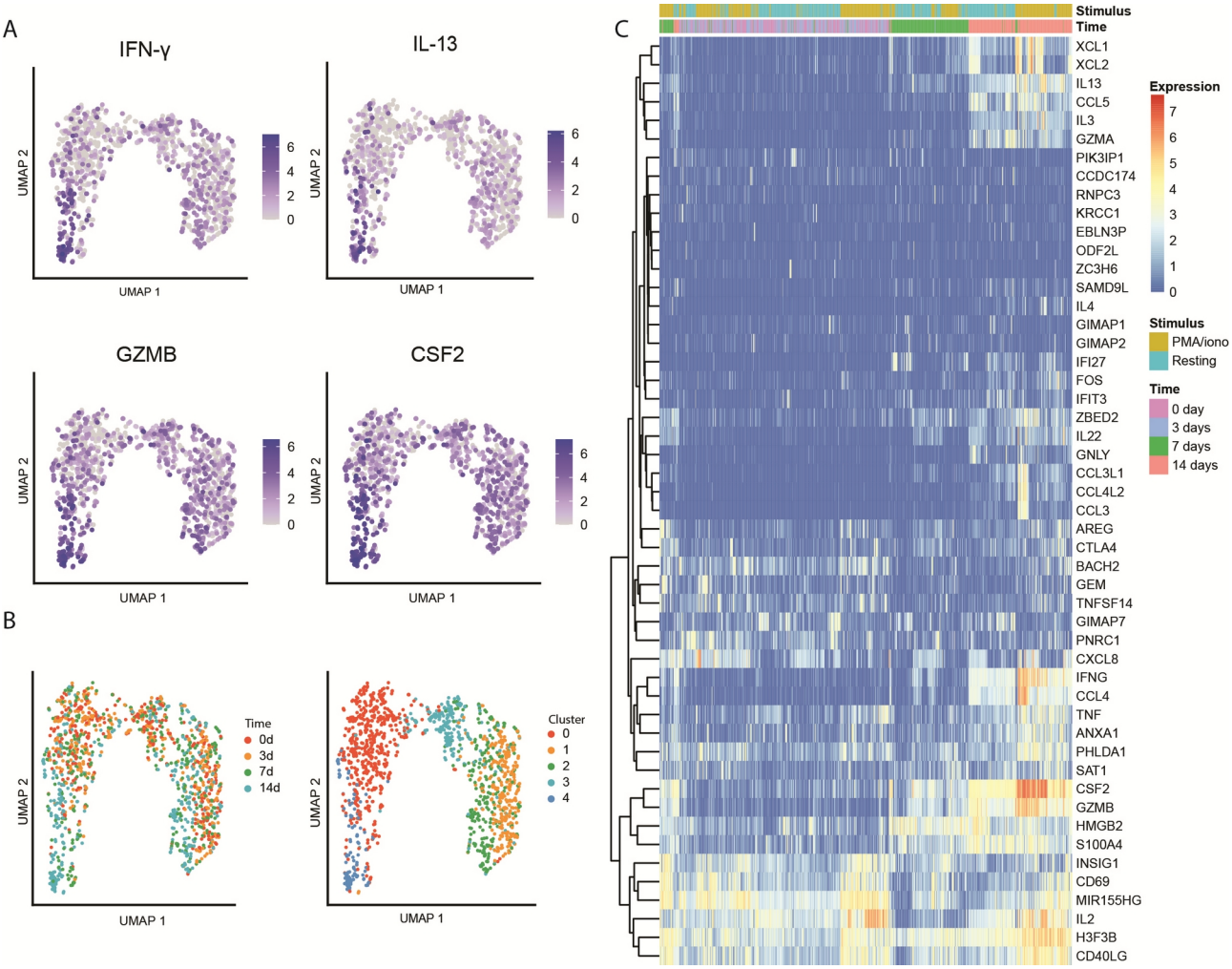

Figure 5.3.7. (**A**) Gene expression plots in UMAP projection. The positioning of cells is identical to (B). (**B**) UMAP plots of restimulated and unstimulated T cells. Cells are color-coded based on the differentiation time (left) and Louvain clustering (right). All UMAPs are based on the top 3000 HVGs. (**C**) Heatmap based on the top 50 HVGs. Cells and genes are hierarchically clustered based on euclidean distance.

5.3.8 Trajectory inference by Slingshot roughly follows the time gradient

Lastly, we wanted to leverage the power of trajectory inference (TI) to detect time-dependent regulators of human T helper cell differentiation fates in our single-cell data. Because the SCRB-seq data exhibits an inherently strong batch effect, we could not use some algorithms like Monocle or Destiny. We, therefore, decided to use Slingshot, a tree-based TI algorithm[176]. We split the data into two sets: restimulated and resting, because both populations exhibited similar expression profiles, and we were not interested in the differences between stimulation conditions. We analyzed them separately, and TI was based on UMAP reduced data. Slingshot identified only one trajectory in either dataset, and it roughly corresponds to the time gradient (Fig. 5.3.8 A). Since the signal was much stronger in restimulated population, we continued our analysis with only that dataset. Genes that changed their expression most significantly over the identified pseudotime trajectory were mostly cytotoxicity markers such as *GZMB, GZMH*, and cytokines *IFN-γ, CSF2, CCL4L2, IL-13, CCL5* (Table.5.3.8). The same genes were also identified by DE analysis of restimulated cells 14 days post activation vs. 0 hours (Fig.5.3.8 B and Table 5.3.8).

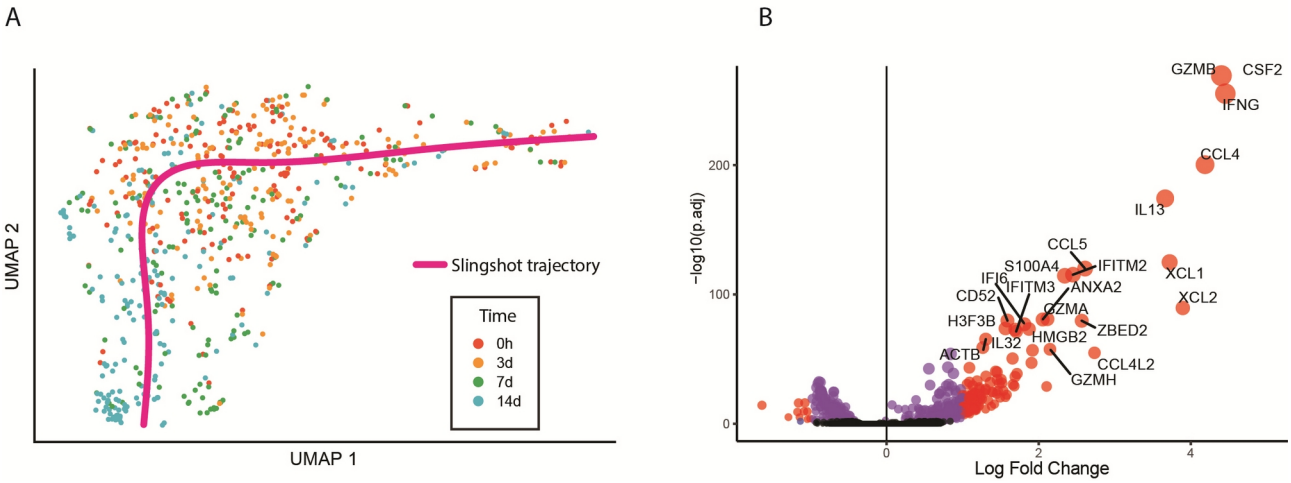

Figure 5.3.8 (**A**) UMAP projection of restimulated T cells overlaid with Slingshot's inferred trajectory in pink. Cells are colored by differentiation time. (**B**) Volcano plot showing DE results of restimulated cells between 14 days and 0 hours.

Next, we looked for transcription factors among pseudo-temporally expressed genes and found seven with FDR less than 1% (Table 5.3.8). All of them are well known in the context of T helper cell differentiation or cellular activation and proliferation. *ZBED2* is a zinc finger

protein recently described as *IRF1* antagonist[177]. *JUND* is part of the *AP1* transcription factor complex and protects cells from apoptosis and senescence[178]. *GATA3* is a well-known transcription factor and a master regulator of T_H2 lineage. *RUNX3* is implicated in cytotoxic T cell development and a tumor supressor[179,180]. *BHLHE40* has been shown to regulate *CSF2* production in mice and to be essential for pathogenicity in neuroinflammation[181]. *FOXP1*, similar to *FOXP3*, is a known regulator of regulatory T cell (Treg) development[182]. *HDGF* has been tied to cellular growth, proliferation, and differentiation[183]. *NR4A3* (together with *NR4A1*) is known to be differentially expressed in activated T cells[184] and implicated in $CD8^+$ T cell development[185].

Table 5.3.8 Slingshot TI and DE results. On the left, the top 10 most significantly upregulated genes between zero hours and 14 days. In the middle, the top 10 genes identified by Slingshot as having the highest significant associations between expression and pseudotime. On the right, all transcription factors differentially expressed along the Slingshot lineage with FDR<1%

Top 10 DE genes 0 h vs. 14 days			
#	**Genes**	LFC	**FDR**
1	CSF2	4.9	3.15E-297
2	GZMB	4.4	7.36E-272
3	IFNG	4.5	2.70E-258
4	CCL4	4.2	2.60E-201
5	IL13	3.7	4.36E-175
6	XCL1	3.7	1.31E-126
7	CCL5	2.6	5.59E-122
8	S100A4	2.3	4.89E-116
9	IFITM2	2.4	6.45E-116
10	XCL2	3.9	2.09E-90

Top 10 genes identified by Slingshot		
#	**Genes**	**FDR**
1	CSF2	3.29E-44
2	GZMB	2.19E-43
3	GZMH	2.22E-42
4	LGALS1	1.19E-37
5	IFNG	7.36E-33
6	XCL1	2.08E-32
7	CCL4L2	1.18E-31
8	IL13	2.52E-31
9	CCL5	1.47E-30
10	CCL4	1.9E-29

Top TFs identified by Slingshot			
#	**Genes**	**p.val**	**FDR**
1	ZBED2	2.44E-29	7.32E-26
2	JUND	1.76E-19	5.29E-16
3	GATA3	1.83E-13	5.49E-10
4	RUNX3	1.16E-09	3.47E-06
5	BHLHE40	2.74E-09	8.23E-06
6	FOXP1	1E-06	0.003
7	HDGF	1.12E-06	0.00335
8	NR4A3	2.04E-06	0.00612

5.4 Analysis of T helper cell transcriptome on a population level

5.4.1 Low-input bulk sequencing for increased sensitivity

Although SCRB-seq is one of the most sensitive single-cell RNA-seq methods available[137], it may still lack sensitivity to detect T helper cell lineages' subtle differences. To address that potential issue, we prepared libraries of 56 bulk samples in restimulated and resting conditions. We also included in those samples knockout T cells deficient for either master transcription factor *GATA3* or *TBX21* (*GATA3*$^{-/-}$ and *TBX21*$^{-/-}$ respectively). In order to achieve homogeneous populations of cells that were reliably deficient for the targeted gene of interest, we prepared samples from monoclonal colonies by limiting dilution

cloning using activated $CD4^+$ T cells that were expanded under non-polarizing conditions (see methods 3.2.6 and 3.2.3). Genotype and clonality were assessed using deep sequencing (see methods 3.2.7). After three weeks, the samples were split for library preparation and analysis of cytokine concentration in the supernatant by ELISA. Unfortunately, we had issues with the RNA clean-up procedure, which led to a loss of half of all the samples (Table 5.4.1).

Table 5.4.1 Low-input bulk RNA-seq T helper cells sample table. The numbers indicate how many samples of each genotype were prepared and how many passed QC.

Low-input bulk RNA-seq				
#	**Genotype**	**PMA/ionomycin**	**Resting**	**Total**
1	TBX21$^{-/-}$	6/3	6/3	6
2	GATA3$^{-/-}$	6/2	6/3	5
3	WT	32/15	32/15	30

For positive controls, we chose the transcription factors *TBX21* and *GATA3* since they are well-characterized regulators of T helper cell fates. *TBX21* is required for T_H1 cell fate, whereas *GATA3* inhibits *IFN-γ* expression and is considered to be the main regulator of T_H2 cell fate[186,187]. By knocking out a necessary component in the differentiation program of a specific T cell lineage, we hoped to skew those samples towards the opposing cell fate. All the samples were expanded under non-skewing conditions, and all samples were monoclonal, ensuring that each colony would spontaneously acquire only one distinct T helper cell phenotype. We prepared 32 WT samples to would spontaneously give rise to different T helper subsets to ensure enough variability in our dataset that both T helper lineage would be represented.

5.4.2 T helper cell transcriptome on a population level

At first glance, the results of low-input bulk data look similar to single-cell data. The relatively strong effect of PMA/Ionomycin restimulation dictated the clustering into two main groups, regardless of the effects that the genotypes might have. *GATA3*$^{-/-}$ and *TBX21*$^{-/-}$ clones stay very near to each other, and there is much variation in wild-type samples (Fig.5.4.2 A).

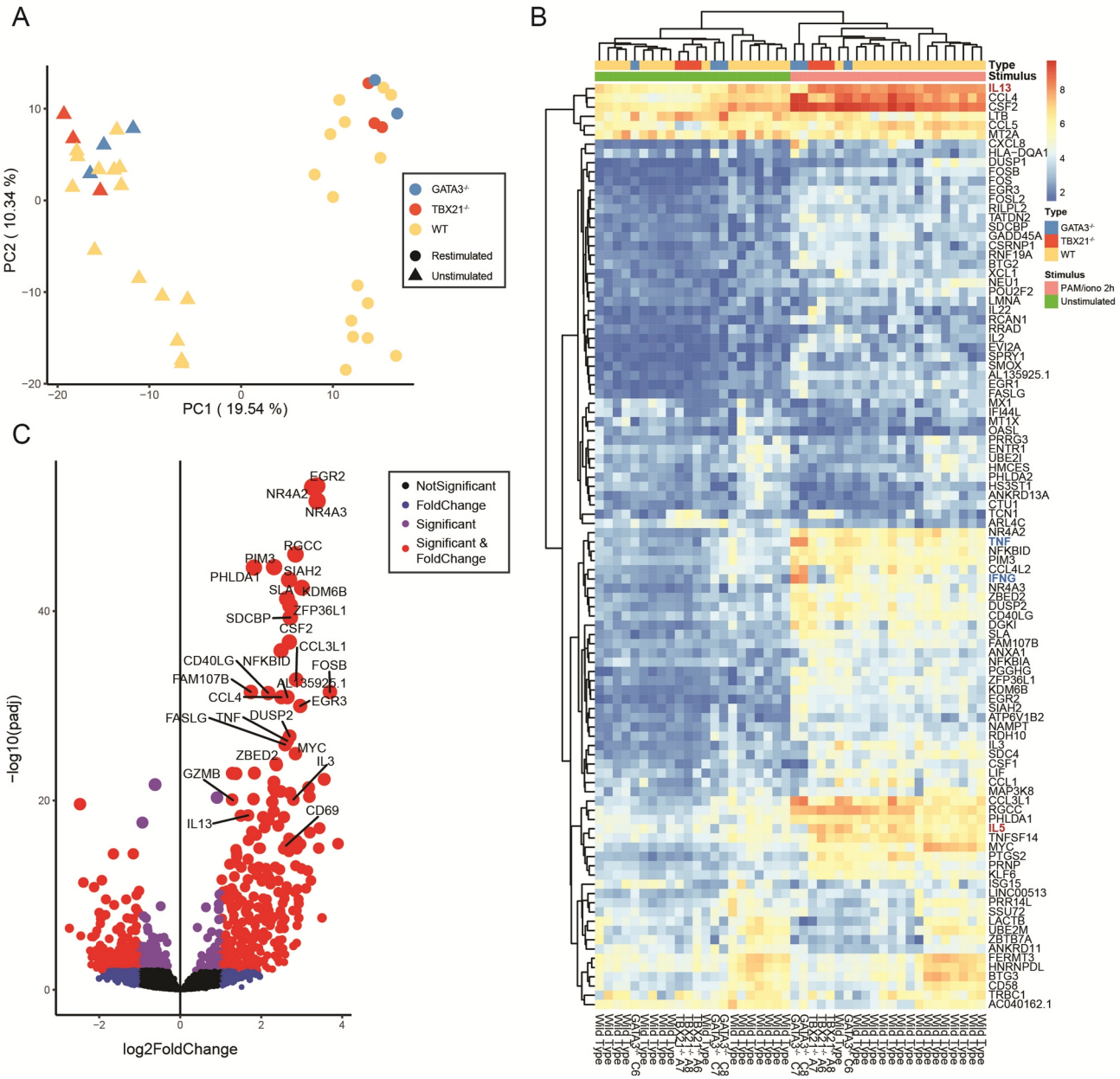

Figure 5.4.2 (**A**) PCA of low-input bulk RNA-seq data, based on top 1000 HVGs. Colors indicate genotype, shape PMA/Ionomycin stimulation. (**B**) Heatmap based on top 100 HVGs. Hierarchical clustering based on euclidian distance was used for both samples and genes. (**C**) Volcano plot showing the results of DE analysis between restimulated and unstimulated samples.

As with single-cell data, restimulation with PMA/Ionomycin increases the expression of cytokines and chemokines, but the effect on total reads per sample was now absent, possibly due to the nature of the bulk sequencing. We noticed the relatively high expression of some cytokines like *CSF2, CCL4,* and *IL-13* present in the unstimulated

samples (Fig.5.4.2 B), which was still significantly increased in the restimulated samples (Fig.5.4.2 C). Conversely, many of the T helper hallmark cytokines were only expressed at background levels in unstimulated samples (*IFN-γ*, IL-5). Interestingly none of the granzymes or other cytotoxicity markers (*GNLY*) showed up as HVGs, as they did in both of the previous single-cell RNA-seq datasets (Fig.5.4.2 B). These genes are still expressed at high levels (data not shown) but more uniformly in all samples. Unlike our two previous single-cell experiments, this time, we did not have samples from different time points along the differentiation process. This could explain why the cytotoxicity markers were absent in HVGs as they showed a strong time-dependent expression profile. This confirms our suspicion that granzymes, especially *GZMB* and *GNLY*, are expressed in all differentiated T helper cells.

Next, we looked at what effects PMA/Ionomycin stimulation had on the T helper cells on a population level. Similar to the single-cell data, the T cell activation (*CD69*, *IL-2*) and cell survival markers (*MYC, TNF, NFKBID*) were upregulated (Fig.5.4.2 C). Yet, the effect was much more pronounced than in single-cell data, resulting in many more genes DE between restimulated and unstimulated samples. Additionally, many of the T helper lineage-specific cytokines like *IFN-γ, IL-4, IL-13, TNF* were present in the dataset (Fig.5.4.2 C). As in the single-cell data, when we looked for enriched pathways in the upregulated transcripts, we observed cell-death-related pathways as well as cytokine signaling (data not shown).

When trying to discern T helper subtypes in the data, we utilized Louvain and hierarchical clustering methods. Neither method managed to isolate our two control genotypes, T_H1 (*GATA3*$^{-/-}$) and T_H2 (*TBX21*$^{-/-}$), into separate clusters. The clustering was mainly defined by restimulation with additional clusters within the WT samples (Fig.5.4.2 B).

When measuring the cytokine concentrations in the supernatant, we detected a clear difference between the two controls (Fig.5.4.2 E). However, when looking at those samples at a whole transcriptome level, the difference was entirely lost. Only when limiting our RNA-seq analysis to T helper hallmark cytokines, *GATA3*$^{-/-}$ and *TBX21*$^{-/-}$ samples separated into distinct groups (Fig.5.4.2 F).

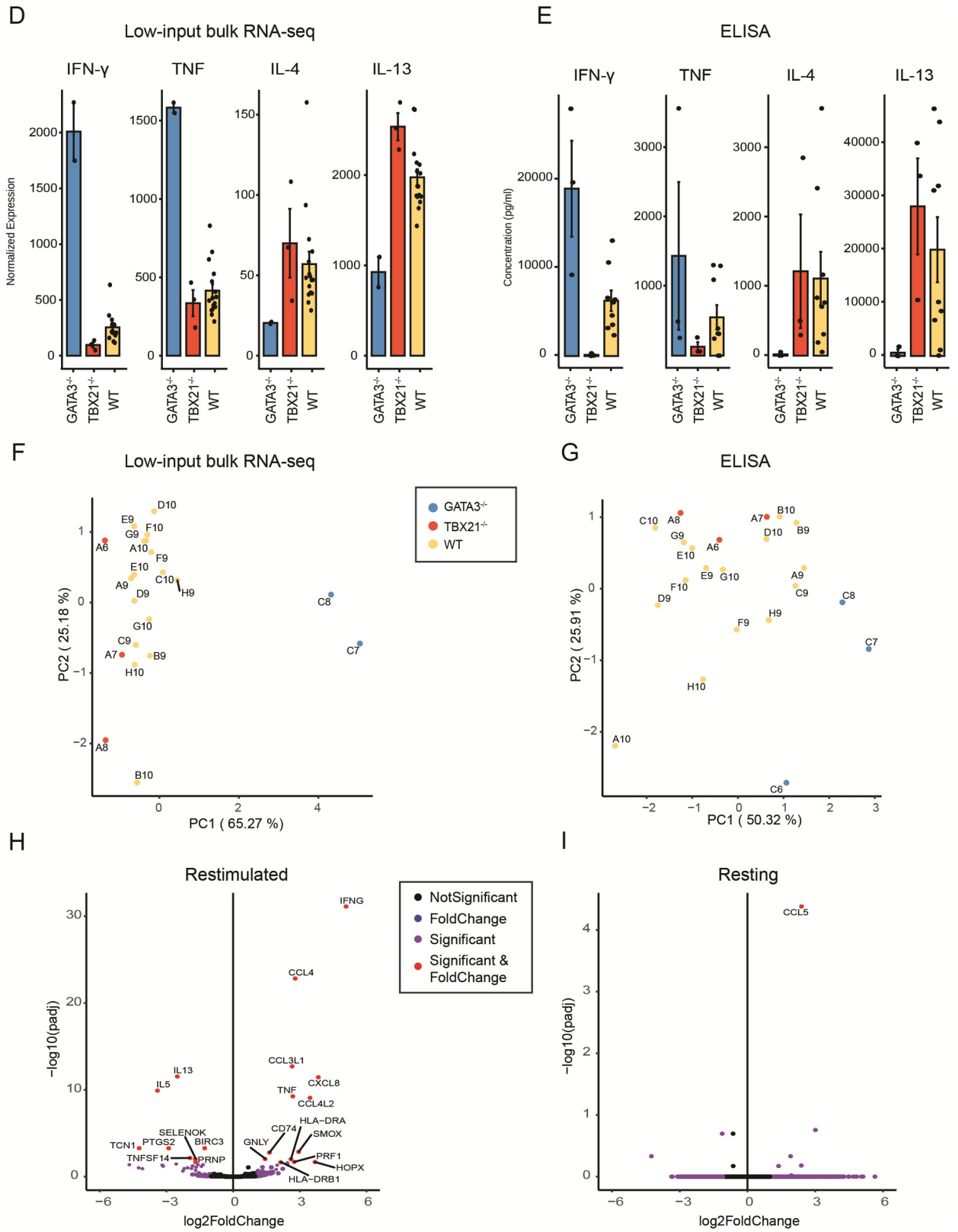

Figure 5.4.2 (**D**) Barplots of 4 different T helper lineage defining cytokines. Expression levels measured by RNA-seq are represented as mean +/- SEM of three independent experiments due to sample dropout, only two samples of *GATA3*$^{-/-}$ were used. (**E**) Barplots of 4 different T helper lineage defining cytokines. Cytokine concentration as measured by ELISA is depicted as mean +/- SEM of at least three independent

experiments. (**F**) and (**G**) PCA built using four hallmark cytokines (IFN-γ, TNF, IL-4, IL-13) based on their expression levels in RNA-seq data (F) and the measured concentrations in the supernatant (G). (**H**) Volcano plot showing DE analysis results between restimulated GATA3$^{-/-}$(n=2) and TBX21$^{-/-}$(n=3) samples. (**I**) Volcano plot showing DE analysis results between unstimulated GATA3$^{-/-}$(n=3) and TBX21$^{-/-}$(n=3) samples.

The cytokine expression as determined by ELISA is reflected on mRNA level (Fig.5.4.2 D). *GATA3* KOs express only baseline levels of T_H2 specific cytokines (*IL-4, IL-13*) but have the highest levels of *IFN-γ*, both on protein and mRNA level. Whereas T_H1 specific cytokine expression was reduced in *TBX21*$^{-/-}$. Again, even more so on the protein level. *TBX21* KOs produced higher T_H2 specific cytokines, but the difference was less pronounced for *GATA3*$^{-/-}$. However, when looking at the expression levels (both mRNA and protein) of these critical cytokines in the WT samples, we saw that they were mostly scattered somewhere in the middle compared to KOs (Fig.5.4.2 F and G). If it were not for the extremely high *IFN-γ* expression in *GATA3*$^{-/-}$, all the samples would fall into a continuous population with *TBX21*$^{-/-}$ on one side and *GATA3*$^{-/-}$ on the other. The extremely high IFN-γ expression in *GATA3*$^{-/-}$ is most likely due to the absence of *GATA3* protein since it has been shown that *GATA3* functions as a negative regulator of IFN-γ expression[188].

As the last part of the analysis, we looked at the differentially expressed genes between the two genotypes *GATA3*$^{-/-}$ and *TBX21*$^{-/-}$. We only found DE in a few transcripts. In *GATA3*$^{-/-}$ samples, most of the upregulated genes are known to be *IFN-γ* dependent such as *CCL4, CSXL8, TNF, HLA-DRA, HLA-DRB1, CCL3L1, CCL4L2* (Fig.5.4.2 H)[189,190]. *TBX21*$^{-/-}$ express increased T_H2 specific cytokines *IL-5* and *IL-13*, but interestingly, the original T_H2 hallmark cytokine *IL-4* is not significantly upregulated. This might be entirely due to the mRNA's low expression level. Indeed, in the first single-cell experiment, we captured almost no *IL-4* mRNA. In general, IL-13 seems to be the most easily detected of all the T_H2 hallmark cytokines on protein and mRNA level.

These small differences between *GATA3*$^{-/-}$ and *TBX21*$^{-/-}$ are almost absent when we compare the unstimulated samples. The sole upregulated gene in *GATA3*$^{-/-}$ is *CCL5*, which again is known to be upregulated in an *IFN-γ* dependent manner[191] (Fig.5.4.2 I). That suggests that even in the resting state, the *GATA3*$^{-/-}$ produces *IFN-γ* over the background level.

5.5 Discussion

Many recent single-cell sequencing studies of T cells, especially $CD4^+$ T cells, have failed to show discrete, lineage-committed T helper cell sub-populations[112,115,192–196]. To determine discrete T helper subset signatures, we set out to take a more controlled approach using a more sensitive sequencing method[48,137] and in vitro T cell differentiation. After purifying human $CD4^+$ T cells and activating them with bead-bound anti-CD3 and soluble anti-CD28 antibodies, we let them expand for up to seven days in the absence of polarizing cytokines (non-skewing condition). From prior experience, we knew that naive T cells grown under such non-skewing conditions gave rise to both T_H1 and T_H2 like cells when analyzing them with flow cytometry. With that in mind, we acquired naive $CD4^+$ T cells from two different donors and set up a time-course experiment to find time-dependent regulators of T helper cell fates and distinguishable T_H1 and T_H2 subsets. Although our single-cell sequencing method is considerably more sensitive than some commercial solutions used in many published studies[151], it has drawbacks. After the initial analysis and data visualization, we noticed a strong batch effect. It has been noted before that plate-based methods have more issues with batch effect[170], but an additional complicating factor in our case seems to have been the use of T cells, especially in the resting state. When not activated (naive or memory), T cells are quiescent, and even though memory T cells are slightly larger, they are still one of the smallest cells in the human body[197]. As transcript abundance correlates with cell size and activity (Fig 5.3.4 C), we noticed a lot more batch effect in earlier time points compared to later ones. Having less available mRNA to capture seems to make library preparation more prone to technical variation. The plates themselves were the most significant source of variation as correcting for the donor, sequencing lane, or handling of pairs of plates had little impact on results (Fig 5.3.1). This implies that most of the technical variation is caused by the cell capture and lysis and not by handling during RT and pre-amplification.

After regressing out unwanted variation, we were left with a heterogeneous and continuous population of T cells. On one side of the population were the cells from earlier time points, while the last two time points (day three and seven) positioned to the opposite side. We could not define any specific regions in that population that could be ascribed to any of the known T helper types. As expected, there was an absence of T helper lineage

defining cytokines at earlier time points, but the expression level of these transcripts did not increase much, even at later time points. *IFN-γ* was the only hallmark cytokine in the top 100 HVGs (Fig. 5.3.2 B). This discrepancy between protein level read-out (FACS, ELISA) and RNA-seq data is likely explained by the restimulation procedures before starting the protein level assays.

Even though the lineage-specific cytokines were lowly expressed, we are not limited to predefined genes of interest in scRNA-seq data. Instead, we can look at the entire captured transcriptome of early and late differentiating $CD4^+$ T cells. One interesting gene that showed up in our first experiment and in all the following ones was *CSF2. CSF2* or Granulocyte-Macrophage Colony-Stimulating Factor *GM-CSF* is usually described as a hematopoietic-cell growth factor[198], but it has also been associated with T helper cells since Mosmann et al. first described different T helper types. In their study, they assigned the expression of *CSF2* to T_H1 cells[165]. More recently, however, there have been multiple groups trying to associate the production of *CSF2* with a novel T helper cell type [107,108,195,199]. We noticed a relatively high expression of *CSF2* in late differentiating T helper cells. Considering all three experiments together, we see a relatively uniform and high CSF2 expression dependent on differentiation time and restimulation. Additionally, we found no DE of *CSF2* between *GATA3*$^{-/-}$ and *TBX21*$^{-/-}$ even though *CSF2* is often associated with T_H1 fate (Fig.5.3.2. C, Fig.5.3.6 A, Fig.5.4.2 C and B).

Another unexpected result was a high granzyme expression (*GZMB, GZMA, GZMH*) in mid to late differentiating T helper cells. Granzymes are usually associated with cytotoxic T cells. $CD8^+$ T cells use granzymes to trigger apoptosis in target cells[200]. However, there have been reports of cytotoxic $CD4^+$ T cells over the past two decades[201,202]. More recently, the discoveries facilitated by scRNA-seq have revealed that granzymes, especially *GZMB,* are present in activated and memory $CD4^+$ T cells[112,113,193,195,203]. It seems that granzyme production is a part of the normal functioning of T helper cells, and cytotoxicity might not only be a function of $CD8^+$ T cells. $CD8^+$ T cells kill cells by releasing perforins which polymerase into pores in the lipid bilayer of target cells. Although the pores in the target cell are already fatal, the killing efficiency is increased further by the release of granzymes[64]. It is conceivable that in this, the $CD8^+$ and $CD4^+$ T cells can cooperate in killing target cells, where $CD8^+$ T cells provide perforin and some granzymes, and $CD4^+$ T cells release additional granzymes. Moreover, there could be as of yet unknown role for

granzymes in helper T cells.

As we already mentioned, there is a marked difference in the cells between early and late time points. To explore this further, we looked for DE genes in each of these time points. In the four-hour post-activation cells, we notice upregulation in many classical activation markers and very few cytokines. Whereas in the late time point cells (three and seven days), there is a significant increase in cytokines, chemokines, and granzymes. Our first experiment indicates that it takes about three days from the initial TCR stimulation to detect differentiation-related transcriptional changes that indicate a memory T cell program. Somewhat faster than 7-15 days as described in the literature[204]. However, the difference might be in the lag between transcriptional changes and protein expression[205]. We also notice a continued increase in the numbers of cells that produce T helper-associated cytokines along the differentiation time. Not all cells are *CSF2* or *GZMB* positive by day 7 (Fig.3.2 C), which indicates that further increasing the differentiation time might yield more informative results.

Because of the generally low expression levels of T helper cell-associated cytokines, we decided to repeat the experiment with restimulated T cells, and we also increased the differentiation time. Since we did not notice any significant changes between the two ways of purifying naive $CD4^+$ T cells, we decided only to use FACS purified T cells in our next experiment. We tried to reduce the known batch effect even further by sorting both restimulated and resting cells onto the same plate and pooling the libraries to run on the same lanes. Despite these precautions, we still noticed some batch effect between capture plates, but it was less pronounced, mostly because more cells in this dataset were from later time points when they are less affected by technical noise as they are physically bigger and transcriptionally more active. Still, we needed to consider this in the analysis.

PMA and Ionomycin stimulation had the expected effect of increasing the cytokine expression[174], and the T cells fell into two clusters based on the stimulation (Fig. 5.3.5. B). Earlier time points were less affected by restimulation as it is effectively just a more potent TCR stimulus. Due to the restimulation, we detected more T helper cytokines from both T_H1 and T_H2 lineage, but no clear sub-populations of T helper cells were visible. The PMA and Ionomycin stimulation does not seem to affect the T cells negatively[174]. In our hands, the two-hour restimulation not only helped to induce more cytokines, but restimulated cells also showed increased transcription. For each day, the average number of reads, UMIs, and

genes was higher in the restimulated group (Fig.5.3.4 C). As with the previous experiment, we found time-dependent changes in gene expression. Both restimulated and resting cells behaved similarly, but the cytokine and granzyme expression was much higher in the restimulated cells. We detected an increased expression of granzymes and *CSF2* in both populations, especially in later time points of restimulated cells (Fig.5.3.6 A).

Although no distinct sub-populations were immediately detectable, we still used Seurat's built-in Louvain clustering algorithm based on SNN to find possible memory T helper cell regions in our data. Since there is no well-defined approach to selecting the parameters for neighbor calling and clustering, one should interpret these results carefully[115]. We found that four to five clusters describe our data the best; two clusters of early differentiating cells and two clusters in late ones, separated by restimulation (Fig.5.3.6 B). Allowing for more clusters to be assigned, the algorithm started finding smaller groups in the early instead of the late differentiating cells, especially in the resting population. This was most likely due to higher variability caused by technical noise, the same reason we saw a more pronounce batch effect in the earlier population of cells. No smaller clusters were detected in the late time points regardless of the restimulation. Even though clustering did not identify T helper subsets, we still used trajectory inference analysis. Because of the strong batch effect in our data, we could not use some TI methods such as Monocle[206] and Destiny[207] as they rely on their own dimensionality reduction methods, and the batch effect skews the results (data not shown). We followed Saelens et al. recommendations and chose one of the top-performing TI tools called Slingshot[176]. In comparison with other methods, Slingshot is especially good at ordering the cells and for branching lineage assignment[60]. We also chose only the restimulated population of T cells for this analysis for two reasons. First, the restimulation increased the cytokine expression and so increased the biological signal strength, and second, having both resting and restimulated cells in the same analysis confused the algorithm and only produced bifurcating lineages between the two stimulation conditions. The latter case was not useful as it resulted in transcripts differentially expressed between stimulation conditions which we already knew. However, as with clustering, even when running the TI algorithm on restimulated samples only, we found no diverging paths nor managed to identify different memory T helper cells. Slingshot inferred a single trajectory that followed the main variation in the dataset – the differentiation time. The same transcripts were identified by Slingshot and differential

expression analysis between zero hours and 14 days (Fig 5.3.8 A and Table 5.3.8). Because Slingshot utilizes reduced data and dimensionality reduction can introduce distortions into RNA-seq data[116], we also tested Slingshot's performance in higher dimensional space using PCA (and up to 50 PCs), but the algorithm never found more than one lineage (data not shown).

When looking at our single-cell data, we do see the expression of T helper defining cytokines. However, instead of an exclusive expression of T_H1 or T_H2 markers, we notice that cells express high levels of (all detected) T helper cytokines (*IFN-γ, IL-4, IL-13*) and granzymes as well as other chemokines or they have much reduced (baseline) levels of these genes. There is even a small positive correlation between *IFN-γ* and *IL-13* (data not shown). This kind of coexpression of exclusive T helper hallmark cytokines has been noted on the population-level since the early days of human T cell research[208]. Here we also discovered that this holds true on the single-cell level. This effect becomes even more pronounced when considering the restimulated population of cells.

Restimulating in vitro differentiated T cells with PMA and Ionomycin helped capture more lowly expressed and relevant transcripts without impacting on cell viability, yet it did not help us to distinguish T helper types. One recently published single-cell dataset of in vivo differentiated but in vitro restimulated (with anti-CD3, anti-CD28 ABs) T cells from multiple human tissues found similarly significant differences between stimulation conditions but could not detect discrete or even clear but overlapping T helper types[112].

We wanted to ensure that the lack of distinct T helper subsets in our single-cell data was not caused by low sensitivity or other issues, such as batch effect (removal) and cell culture methods. To that end, we decided to investigate the matter further with low-input bulk sequencing. We can easily increase our SCRB-seq protocol's sensitivity by increasing the number of cells per well during capture. This means that we will lose the single-cell resolution, but we gain the possibility to study the same sample with many different methods and detect lowly expressed transcripts that might not have been captured in scRNA-seq.

For bulk sequencing, we prepared knockouts of two of the primary T helper master regulators, *GATA3* and *TBX21*, and additional 32 WT samples. All the samples were expanded in a monoclonal fashion and should spontaneously differentiate into T helper subsets (Fig.5.1 B and C). The knockout samples of master transcription factors were

included to skew cells towards specific T helper lineages. When normal *GATA3* functioning is perturbed, the T cells should spontaneously acquire the T_H1 phenotype and vice versa for *TBX21* and T_H2[209,210]. We confirmed the T helper phenotypes of our KOs with ELISA before proceeding with the sequencing.

The bulk sequencing results reflected our previous finding on a single-cell level. The majority of variation in the data was caused by the PMA and Ionomycin restimulation. Interestingly there was a lot more variation within all the WT clones than there was between the knockouts. *GATA3*$^{-/-}$ showed increased *IFN-γ* and *TNF* expression, and *TBX21*$^{-/-}$ had increased *IL-4* and *IL-13* expression. Although the trends in the expression of these hallmark cytokines in the RNA-seq data and ELISA data agreed, there were still some noticeable differences. On the protein level, the cytokine expression pattern is more clear-cut. The T_H2 specific cytokine expression is abolished in *GATA3*$^{-/-}$, and T_H1 specific cytokine expression is severely reduced in *TBX21*$^{-/-}$. However, on the transcriptional level, we can still detect residual levels of those cytokines. It is hard to discern whether these differences are due to measurement sensitivities or post-transcriptional regulation[211]. In line with the latter hypothesis, we also noticed similarly contrasting expression profiles in flow cytometry data, which is also a protein-based assay.

Our expectation of using the KOs to guide our analysis by defining T helper subsets based on their gene expression profile (*GATA3*$^{-/-}$ are T_H1-like and *TBX21*$^{-/-}$ are T_H2-like) was not met. As already noted before, the WT samples showed more variation among each other than the KO clones, and no clustering dictated by the genotype was noticeable in the PCA plot. Also, clustering algorithms (Louvain, k-means, hierarchical clustering) did not find clusters in a way that would separate our two control genotypes from each other.

The master regulators do not seem to govern the whole T helper cell transcriptional program instead, they only regulate a very limited number of transcripts, such as *IL-4, IL-5, IL-13,* and *IFN-γ*. To pursue this further, we conducted a DE analysis between those two genotypes. Since the samples do not separate from each other in the PCA, the difference in their gene expression was also minor. Most of the upregulated genes in *GATA3*$^{-/-}$ are known *IFN-γ* targets. Even this slight difference all but vanished if we compared these two genotypes in the resting state. The sole DE transcript was *CCL5*, a gene that is also known to be induced by I*FN-γ*[212–214]. Considering that *GATA3* is a known *IFN-γ* repressor[188,215], it is not surprising that *GATA3*$^{-/-}$ samples had the highest expression of

IFN-γ, even in the resting state as indicated by *CCL5* upregulation. Taken together, these results indicate that *IFN-γ* differential expression is the primary driver of variation between *GATA3*$^{-/-}$ and *TBX21*$^{-/-}$.

One way to explain these results, which is also in line with other published data, is to use the T cell plasticity model, where already differentiated memory T helper cells undergo additional changes (chromatin remodeling and or DNA methylation) in response to environmental cues to switch between predefined cell types. Typically T helper cells are considered to be stable due to epigenetic changes and histone acetylations. Additionally to that, the master regulators should not only induce their T_H type defining cytokines but also suppress the characteristics of the opposing T_H type[208]. However, over the last three decades, each of these regulatory checkpoints has been found to be bypassed both in vitro and in vivo. T_H1 cells have been shown to be capable of IL-4 production without losing IFN-γ production, and some T cells co-produce IL-4 and IFN-γ[102,216]. All conventional memory T helper cell types can change their expression profile under specific microenvironmental conditions[102–105]. T_H1 and T_H2 cells have been both found to be co-expressing GATA3 and TBX21[217]. Also, the epigenetic modifications of T helper cells are not as straightforward as the cytokine expression profiles and do not follow the T helper cell type patterns[218]. So either the memory T helper cell types are not as well defined as previously thought, or the differences between those types are far more minor, or these phenotypes are much more overlapping.

Another way of interpreting the single-cell data in light of these contradictory findings has been gaining support in recent years. The so-called polarized continuum or heterogeneous continuum model does not separate memory T helper cells into distinct groups, rather these cells can occupy any possible state between the known extremes (conventional T helper types). In all of the published scRNA-seq datasets discussed in this work, we see that there are no clear and distinct clusters of T helper subsets. Even if some areas of these heterogeneous populations express higher levels of hallmark cytokines, the borders between supposed T helper subsets are by no means clear.

It is, of course, possible to run clustering algorithms on those continuous populations to try to find subsets. Such analysis is, in fact, commonplace, but the resulting clusters might not have anything to do with actual biology. This is how we end up with multiple T_H1 subtypes or “unknown” clusters of T helper cells that do not seem to fall into any conventional

category[112,113,115,195].

All of the above speaks in favor of the continuum model, but there could be a few technical considerations why we cannot distinguish between different types of T helper cells on a transcriptional level. By definition, T cell types are assigned based on their cytokine expression profiles[219], and in our datasets, we only saw different T helper cell types when we limited the cell type assignment to a handful of cytokines. It could be that these lineages specific cytokines are all the difference there is between the T helper cell types.

Another possibility is that there is more differential expression between T helper cell types but all of those transcripts are moderately or lowly expressed and not reliably picked up with current scRNA-seq methods. However, Heimberg et al. showed in their 2016 paper that gene expression data is inherently low dimensional. That means that genes tend not to vary in their expression individually, but instead, they are co-expressed together in modules, thus the lower dimensionality. The implication of this is that RNA-seq data should be surprisingly robust to technical noise; for example, as few as 100 transcripts per cell would suffice to distinguish between oligodendrocytes and neurons[117]. That is around 100 fold fewer features/genes than an average droplet-based sequencing platform manages or 1000 fold less than plate-based methods like SCRB-seq[151]. Unless differences between T helper cell types are exceedingly small, any current generation single-cell sequencing platform should, therefore be able to find it.

Even though the low sensitivity is unlikely to be an issue, considering our scRNA-seq capabilities, we still performed a low-input bulk RNA sequencing on T helper cells. The bulk nature of this experiment helped us validate our scRNA-seq findings and combine it with additional read-outs for the same samples (ELISA). We did not find definite evidence of T_H archetypes of in vitro differentiated $CD4^+$ T cells on population-level either. The increased sensitivity revealed what we already noticed in single-cell data: a continuum of samples (or cells in scRNA-seq) that differ only slightly from each other. Even when we could validate our results with complementary methods such as ELISA and assign a putative phenotype to our samples, we could not verify these findings in RNA-seq data when considering the entire transcriptome. Only when we limited the dataset to a few key cytokines were we finally able to tell our positive controls apart.

Taking all of that into account, it seems that distinct T helper types might be more accurately described as a continuum of cells with different active transcriptional programs

rather than a set cell type. Interestingly we see a similar heterogeneous continuum of cells in the T helper cell counterparts, the ILCs [88–90]. Both fields have struggled to define the cell types solely based on scRNA-seq data. Instead, there's a growing number of studies that try to model the response of individual cells as dictated by pathogens and environmental cues they come in contact with[91,196]. This would also make more biological sense as it would not be very efficient to have a specific cell type for each occasion (pathogen), as there is a potentially unlimited number of them. Instead, a "Jack-of-all-trades" could be a more capable system.

One final topic to consider is the difference between RNA and protein-level read-outs. We have almost exclusively looked at transcriptomics data in this work, but the conventional assays for T helper cell types have been performed on protein level. One issue that we currently cannot reconcile is the fact that under the flow cytometry conditions, we see quite clear differences between different skewing or KO conditions on a cellular level. There could be as of yet unknown post-translational modifications that influence these read-outs. While single-cell whole proteome studies are being developed, they are yet not available for the type of studies conducted here. However, partial single-cell proteomics studies using CyTOF also reveal much the same picture of continuous heterogeneous populations[195,220–222]. As a quick side note, all of the CyTOF data as well as scRNA-seq data reviewed here always shows a clear distinction between $CD4^+$ and $CD8^+$ T cells. Also naive T helper cells separate quite clearly in a non-overlapping manner from memory T cells. The reason why CyTOF data resembles scRNA-seq data might be because of the increased number of parameters measured in any given experiment. As long as the read-outs are limited to only a few parameters, it is possible to distinguish between T helper cell types. Yet, when multi-parametric read-outs are considered in their entirety, these differences vanish.

T cell research has constantly undergone changes and revisions. The T_H1-T_H2 dichotomy that lasted for almost 20 years was abandoned with the discovery of the third lineage, T_H17 cells. In light of the data presented here and recently published work, it might be time to let go of the T helper cell archetypes, especially when all the assays are increasingly more high-throughput and multi-parametric. It seems ill-advised to assign something so complex as a cell type solely by the expression of a few genes. Especially when considering that the said cell types have never shown clear distinction to begin with[99].

It is unlikely that this matter will be settled any time soon, but there seems to be more acceptance that T cell-extrinsic factors might have more to do with their function and properties than intrinsic factors and debatable subtypes[115,196]. This study has hopefully shed some light on this matter. Of course, it is necessary to consider this project's limitations, namely that we only worked with in vitro samples and differentiated the cells under non-polarizing conditions. Yet our results, as well as other published results[113], indicate that there would be few if any, significant differences between different T helper phenotypes whether they are brought up under polarizing or non-polarizing conditions.

6 Summary

RNA sequencing, along with single-cell RNA sequencing, has been gaining popularity in the last decade and a half. It has almost become a routine method, and numerous commercial and non-commercial methods to perform transcriptomics analysis exist. Despite it being so widespread, setting up an RNA sequencing platform, especially a single-cell RNA sequencing platform, is not without its own pitfalls. Although much more straightforward to implement, commercial solutions impose their own restrictions i.e., the choice of reagents, need of microfluidic chips, and constrictions to the general experimental setup. On the other hand, homemade protocols offer much more flexibility in regards to initial investments as well as experimental design and the possibility to combine it with other instruments for more in-depth analysis.

In this work, we showed the capabilities of our improved single-cell sequencing platform – SCRB-seq, in detecting transcriptional changes in BLaER1 macrophages after LPS stimulation. We could further show that our modifications to the original protocol substantially improved the sequencing results and that integrating different datasets (all produced with the same protocol) was feasible and increased the statistical power of RNA-seq analysis.

The properties of SCRB-seq also allow it to be used for low-input bulk RNA sequencing experiments. There are several benefits to such a setup. The main one is the reduced cost achieved by drastically reducing the hands-on time during the library preparation, multiplexing the samples (reducing reagent and material cost), and tag-based sequencing of the libraries. The resulting digital gene expression read-out is especially beneficial for experiments with a large sample sizes, such as screens.

We used this low-input setup to characterize the PRR stimulation in BLaER1 macrophages in a knockout-dependent manner. With three different stimulation conditions (LPS, dsDNA, and unstimulated) and 13 different genotypes, we successfully prepared and sequenced 150 samples. In an unbiased fashion, we could uncover the importance of the chosen factors involved in TLR4 and cGAS-STING signaling. Moreover, we could uncover the redundancy of certain kinases in the NF-κB pathway (*CHUCK* and *IKBKB*), and in the interferon signaling pathway (*IKBKE* and *TBK1*), respectively.

The main focus of this work was to sequence and analyze $CD4^+$ T cells' activation and differentiation and find time-dependent regulators of cell fate. To characterize these processes, we conducted two scRNA-seq experiments and one low-input bulk sequencing experiment. For both for single-cell and bulk sequencing, we found that restimulation is required to capture cytokine profiles characteristic for T helper cells. We were, however, not successful in detecting conventional T helper subtypes on a single-cell level, neither in resting nor in restimulated populations. Instead, our data revealed that the T helper cells, brought up under non-polarizing conditions, exhibit a continuous and heterogeneous population of cells, where the main drivers of variation are restimulation and the differentiation time. This was in clear contrast to the results obtained by flow cytometric analysis, in which distinct populations of T helper subtypes could be discerned.

In bulk sequencing data, we could confirm that the expression of hallmark cytokines (*IL-4, IL-13, IFN-γ*) was indeed dependent on T cell-fate master transcription factors *GATA3* and *TBX21*. As such, we could reliably distinguish between T helper types on population-level, but only when limiting our analysis to those cytokines. In fact, when the whole transcriptome was compared in an unbiased fashion, deficiency of these master regulators had little to no impact and a meaningful grouping into subtypes would not have been possible. In fact, there was more variation between wild-type samples than between the knockouts.

In all three datasets, we noticed a high expression of cytotoxic genes such as *GZMA*, *GZMB*, *GZMH*, and *GNLY*, as well as an almost ubiquitous expression of *CSF2*. The expression of these genes remained unaltered in different knockout conditions and was only dependent on differentiation time. This would be an interesting area for further investigation as these transcripts could possibly serve as additional markers for T helper differentiation.

In summary, this study has shed some light into difficulties identifying T helper types in complex samples using scRNA-seq. Two of the main problems: overlapping cytokine production, and the continuous population of T cells do not seem to be caused by shallow sequencing depth, but is instead appear to be an inherent property of T helper cells on a single-cell level. In the future, combining transcriptomics assays with additional readouts such as single-cell proteomics, epigenomics, and metabolomics could help to reconcile the differences we see on single cell vs. population level.

7 Bibliography

1. Pearson, H. What is a gene? *Nature* **441**, 398–401 (2006).
2. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* **10**, 155–159 (2009).
3. He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* **5**, 522–531 (2004).
4. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
5. Crick, F. The Biological Replication of Macromolecules. (1958).
6. Alberts, B. *et al. Molecular Biology of the Cell*. (Garland Science, 2014).
7. Coffin, J. M. & Fan, H. The Discovery of Reverse Transcriptase. *Annu. Rev. Virol.* **3**, 29–51 (2016).
8. Alwine, J. C., Kemp, D. J. & Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5350–5354 (1977).
9. Miller, B. R., Wei, T., Fields, C. J., Sheng, P. & Xie, M. Near-infrared fluorescent northern blot. *RNA* **24**, 1871–1877 (2018).
10. Mishima, E. *et al.* Immuno-Northern Blotting: Detection of RNA Modifications by Using Antibodies against Modified Nucleosides. *PLoS ONE* **10**, (2015).
11. Mullis, K. *et al.* Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 263–273 (1986).
12. Brock, T. D. & Freeze, H. Thermus aquaticus gen. n. and sp. n., a nonsporulating extreme thermophile. *J. Bacteriol.* **98**, 289–297 (1969).
13. Saiki, R. K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491 (1988).

14. Higuchi, R., Fockler, C., Dollinger, G. & Watson, R. Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions. *Bio/Technology* **11**, 1026–1030 (1993).

15. Vogel, F. A Preliminary Estimate of the Number of Human Genes. *Nature* **201**, 847–847 (1964).

16. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

17. Ensembl genome browser 103. https://www.ensembl.org/index.html.

18. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270**, 467–470 (1995).

19. Lashkari, D. A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 13057–13062 (1997).

20. Bumgarner, R. DNA microarrays: Types, Applications and their future. *Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al* **0 22**, Unit-22.1. (2013).

21. Etienne, W., Meyer, M. H., Peppers, J. & Meyer, R. A. Comparison of mRNA gene expression by RT-PCR and DNA microarray. *BioTechniques* **36**, 618–626 (2004).

22. DNA Sequencing Costs: Data. *Genome.gov* https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

23. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

24. Bainbridge, M. N. *et al.* Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**, 246 (2006).

25. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).

26. Angerer, P. *et al.* Single cells make big data: New challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* **4**, 85–91 (2017).

27. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).

28. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

29. Zhang, X. *et al.* Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. *bioRxiv* 313130 (2018) doi:10.1101/313130.

30. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. *Characterization of directed differentiation by high-throughput single-cell RNA-Seq*. http://biorxiv.org/lookup/doi/10.1101/003236 (2014).

31. Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343**, 776–779 (2014).

32. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).

33. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

34. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

35. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* 1–7 (2020) doi:10.1038/s41587-020-0497-0.

36. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).

37. Lahens, N. F. *et al.* IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol.* **15**, R86 (2014).

38. Durruthy-Durruthy, R. & Ray, M. Using Fluidigm C1 to Generate Single-Cell Full-Length cDNA Libraries for mRNA Sequencing. in *Disease Gene Identification: Methods and Protocols* (ed. DiStefano, J. K.) 199–221 (Springer, 2018). doi:10.1007/978-1-4939-7471-9_11.

39. Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. Direct Comparative Analysis of 10X Genomics Chromium and Smart-seq2. *bioRxiv* 615013 (2019) doi:10.1101/615013.

40. 10x Genomics. 10x Genomics Chromium. *10x Genomics* https://www.10xgenomics.com/ (2021).

41. Lebrigand, K., Magnone, V., Barbry, P. & Waldmann, R. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat. Commun.* **11**, 4025 (2020).

42. Zheng, Y.-F. *et al.* HIT-scISOseq: High-throughput and High-accuracy Single-cell Full-length Isoform Sequencing for Corneal Epithelium. *bioRxiv* 2020.07.27.222349 (2020) doi:10.1101/2020.07.27.222349.

43. Gupta, I. *et al.* Single-cell isoform RNA sequencing (ScISOr-Seq) across thousands of cells reveals isoforms of cerebellar cell types. *bioRxiv* 364950 (2018) doi:10.1101/364950.

44. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).

45. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133–138 (2009).

46. Mikheyev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* **14**, 1097–1102 (2014).

47. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).

48. Bagnoli, J. W. *et al.* Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 2937 (2018).

49. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).

50. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).

51. Kovaka, S., Fan, Y., Ni, B., Timp, W. & Schatz, M. C. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat. Biotechnol.* 1–11 (2020) doi:10.1038/s41587-020-0731-9.

52. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 90 (2018).

53. Briggs, R. & King, T. J. Transplantation of living nuclei from blastula cells into enucleated frogs' eggs. *Proc. Natl. Acad. Sci.* **38**, 455–463 (1952).

54. Waddington, C. H. & Kacser, H. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*. (Allen & Unwin, 1957).

55. Lorberbaum, D. S. & Barolo, S. Gene Regulation: When Analog Beats Digital. *Curr. Biol.* **23**, R1054–R1056 (2013).

56. Barolo, S. & Posakony, J. W. Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. *Genes Dev.* **16**, 1167–1181 (2002).

57. Tintori, S. C., Nishimura, E. O., Golden, P., Lieb, J. D. & Goldstein, B. A Transcriptional Lineage of the Early C. elegans Embryo. *Dev. Cell* **38**, 430–444 (2016).

58. Sun, Y. *et al.* Single-cell RNA profiling links ncRNAs to spatiotemporal gene expression during C. elegans embryogenesis. *Sci. Rep.* **10**, 18863 (2020).

59. Loeffler-Wirth, H., Binder, H., Willscher, E., Gerber, T. & Kunz, M. Pseudotime Dynamics in Melanoma Single-Cell Transcriptomes Reveals Different Mechanisms of Tumor Progression. *Biology* **7**, (2018).

60. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

61. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Comput. Biol.* **14**, e1006245 (2018).

62. Shah, D. T-cell development in thymus. (2020).

63. Spits, H. Development of αβ T cells in the human thymus. *Nat. Rev. Immunol.* **2**, 760–772 (2002).

64. Janeway, C. *Janeway's immunobiology*. (Garland Science, 2017).

65. Seder, R. A. & Ahmed, R. Similarities and differences in CD4 + and CD8 + effector and memory T cell generation. *Nat. Immunol.* **4**, 835–842 (2003).

66. Kumar, B. V., Connors, T. & Farber, D. L. Human T cell development, localization, and function throughout life. *Immunity* **48**, 202–213 (2018).

67. Hashimoto, K. *et al.* Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *Proc. Natl. Acad. Sci.* **116**, 24242–24251 (2019).

68. Corthay, A. A Three-cell Model for Activation of Naïve T Helper Cells. *Scand. J. Immunol.* **64**, 93–96 (2006).

69. Brunner, M. C. *et al.* CTLA-4-Mediated Inhibition of Early Events of T Cell Proliferation. *J. Immunol.* **162**, 5813–5820 (1999).

70. Jenkins, M. K., Taylor, P. S., Norton, S. D. & Urdahl, K. B. CD28 delivers a costimulatory signal involved in antigen-specific IL-2 production by human T cells. *J. Immunol. Baltim. Md 1950* **147**, 2461–2466 (1991).

71. Frauwirth, K. A. & Thompson, C. B. Activation and inhibition of lymphocytes by costimulation. *J. Clin. Invest.* **109**, 295–299 (2002).

72. Gutcher, I. & Becher, B. APC-derived cytokines and T cell polarization in autoimmune inflammation. *J. Clin. Invest.* **117**, 1119–1127 (2007).

73. Constant, S., Pfeiffer, C., Woodard, A., Pasqualini, T. & Bottomly, K. Extent of T cell receptor ligation can determine the functional differentiation of naive CD4+ T cells. *J. Exp. Med.* **182**, 1591–1596 (1995).

74. Kaplan, M. H., Sun, Y. L., Hoey, T. & Grusby, M. J. Impaired IL-12 responses and enhanced development of Th2 cells in Stat4-deficient mice. *Nature* **382**, 174–177 (1996).

75. Takeda, K. *et al.* Essential role of Stat6 in IL-4 signalling. *Nature* **380**, 627–630 (1996).

76. Afkarian, M. *et al.* T-bet is a STAT1-induced regulator of IL-12R expression in naïve CD4+ T cells. *Nat. Immunol.* **3**, 549–557 (2002).

77. Zhou, L. *et al.* IL-6 programs T(H)-17 cell differentiation by promoting sequential engagement of the IL-21 and IL-23 pathways. *Nat. Immunol.* **8**, 967–974 (2007).

78. Roberts, A. I. *et al.* The role of activation-induced cell death in the differentiation of T-helper-cell subsets. *Immunol. Res.* **28**, 285–293 (2003).

79. Pawelec, G., Sansom, D., Rehbein, A., Adibzadeh, M. & Beckman, I. Decreased proliferative capacity and increased susceptibility to activation-induced cell death in late-passage human CD4+ TCR2+ cultured T cell clones. *Exp. Gerontol.* **31**, 655–668 (1996).

80. Oberg, H. H., Lengl-Janssen, B., Kabelitz, D. & Janssen, O. Activation-induced T cell death: resistance or susceptibility correlate with cell surface fas ligand expression and T helper phenotype. *Cell. Immunol.* **181**, 93–100 (1997).

81. Garrod, K. R. *et al.* Dissecting T Cell Contraction In Vivo Using a Genetically Encoded Reporter of Apoptosis. *Cell Rep.* **2**, 1438–1447 (2012).

82. Romagnani, S. TH1 and TH2 in Human Diseases. *Clin. Immunol. Immunopathol.* **80**, 225–235 (1996).

83. Abbas, M., Moussa, M. & Akel, H. Type I Hypersensitivity Reaction. in *StatPearls* (StatPearls Publishing, 2021).

84. Harrington, L. E. *et al.* Interleukin 17–producing CD4 + effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. *Nat. Immunol.* **6**, 1123–1132 (2005).

85. Wu, X., Tian, J. & Wang, S. Insight Into Non-Pathogenic Th17 Cells in Autoimmune Diseases. *Front. Immunol.* **9**, (2018).

86. Vivier, E. *et al.* Innate Lymphoid Cells: 10 Years On. *Cell* **174**, 1054–1066 (2018).

87. von Burg, N., Turchinovich, G. & Finke, D. Maintenance of Immune Homeostasis through ILC/T Cell Interactions. *Front. Immunol.* **6**, (2015).

88. O'Sullivan, T. E. Dazed and Confused: NK Cells. *Front. Immunol.* **10**, (2019).

89. Suffiotti, M., Carmona, S. J., Jandus, C. & Gfeller, D. Identification of innate lymphoid cells in single-cell RNA-Seq data. *Immunogenetics* **69**, 439–450 (2017).

90. Bielecki, P. *et al.* Skin-resident innate lymphoid cells converge on a pathogenic effector state. *Nature* **592**, 128–132 (2021).

91. Mazzurana, L. *et al.* Tissue-specific transcriptional imprinting and heterogeneity in human innate lymphoid cells revealed by full-length single-cell RNA-sequencing. *Cell Res.* 1–15 (2021) doi:10.1038/s41422-020-00445-x.

92. Chtanova, T. *et al.* T follicular helper cells express a distinctive transcriptional profile, reflecting their role as non-Th1/Th2 effector cells that provide help for B cells. *J. Immunol. Baltim. Md 1950* **173**, 68–78 (2004).

93. Vinuesa, C. G., Tangye, S. G., Moser, B. & Mackay, C. R. Follicular B helper T cells in antibody responses and autoimmunity. *Nat. Rev. Immunol.* **5**, 853–865 (2005).

94. Dardalhon, V. *et al.* Interleukin 4 inhibits TGF-β-induced-Foxp3+T cells and generates, in combination with TGF-β, Foxp3− effector T cells that produce interleukins 9 and 10. *Nat. Immunol.* **9**, 1347–1355 (2008).

95. Kaplan, M. H. Th9 cells: differentiation and disease. *Immunol. Rev.* **252**, 104–115 (2013).

96. Skapenko, A. *et al.* GATA-3 in Human T Cell Helper Type 2 Development. *J. Exp. Med.* **199**, 423–428 (2004).

97. Kanhere, A. *et al.* T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nat. Commun.* **3**, 1268 (2012).

98. Unutmaz, D. RORC2: the master of human Th17 cell programming. *Eur. J. Immunol.* **39**, 1452–1455 (2009).

99. Kelso, A. Th1 and Th2 subsets: paradigms lost? *Immunol. Today* **16**, 374–379 (1995).

100. Hegazy, A. N. *et al.* Interferons direct Th2 cell reprogramming to generate a stable GATA-3(+)T-bet(+) cell subset with combined Th2 and Th1 cell functions. *Immunity* **32**, 116–128 (2010).

101. Lucey, D. R., Clerici, M. & Shearer, G. M. Type 1 and type 2 cytokine dysregulation in human infectious, neoplastic, and inflammatory diseases. *Clin. Microbiol. Rev.* **9**, 532–562 (1996).

102. Messi, M. *et al.* Memory and flexibility of cytokine gene expression as separable properties of human T H 1 and T H 2 lymphocytes. *Nat. Immunol.* **4**, 78–86 (2003).

103. Smits, H. H. *et al.* IL-12-induced reversal of human Th2 cells is accompanied by full restoration of IL-12 responsiveness and loss of GATA-3 expression. *Eur. J. Immunol.* **31**, 1055–1065 (2001).

104. Harbour, S. N., Maynard, C. L., Zindl, C. L., Schoeb, T. R. & Weaver, C. T. Th17 cells give rise to Th1 cells that are required for the pathogenesis of colitis. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7061–7066 (2015).

105. Lu, K. T. *et al.* Functional and epigenetic studies reveal multistep differentiation and plasticity of in vitro-generated and in vivo-derived follicular T helper cells. *Immunity* **35**, 622–632 (2011).

106. Das, A. *et al.* Effector/memory CD4 T cells making either Th1 or Th2 cytokines commonly co-express T-bet and GATA-3. *PloS One* **12**, e0185932 (2017).

107. Zhang, J. *et al.* A novel subset of helper T cells promotes immune responses by secreting GM-CSF. *Cell Death Differ.* **20**, 1731–1741 (2013).

108. Herndler-Brandstetter, D. & Flavell, R. A. Producing GM-CSF: a unique T helper subset? *Cell Res.* **24**, 1379–1380 (2014).

109. Duhen, T., Geiger, R., Jarrossay, D., Lanzavecchia, A. & Sallusto, F. Production of interleukin 22 but not interleukin 17 by a subset of human skin-homing memory T cells. *Nat. Immunol.* **10**, 857–863 (2009).

110. Fujio, K., Okamura, T. & Yamamoto, K. The Family of IL-10-secreting CD4+ T cells. *Adv. Immunol.* **105**, 99–130 (2010).

111. Kimmel, J. C. *et al.* Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Res.* **29**, 2088–2103 (2019).

112. Szabo, P. A. *et al.* Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat. Commun.* **10**, 4706 (2019).

113. Cano-Gamez, E. *et al.* Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4 + T cells to cytokines. *Nat. Commun.* **11**, 1801 (2020).

114. Zakharov, P. N., Hu, H., Wan, X. & Unanue, E. R. Single-cell RNA sequencing of murine islets shows high cellular complexity at all stages of autoimmune diabetes. *J. Exp. Med.* **217**, (2020).

115. Zemmour, D., Kiner, E. & Benoist, C. CD4+ teff cell heterogeneity: the perspective from single-cell transcriptomics. *Curr. Opin. Immunol.* **63**, 61–67 (2020).

116. Cooley, S. M., Hamilton, T., Deeds, E. J. & Ray, J. C. J. A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data. *bioRxiv* 689851 (2019) doi:10.1101/689851.

117. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low-dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* **2**, 239–250 (2016).

118. Kobak, D. & Linderman, G. C. UMAP does not preserve global structure any better than t-SNE when using the same initialization. *bioRxiv* 2019.12.19.877522 (2019) doi:10.1101/2019.12.19.877522.

119. Becht, E. *et al.* Evaluation of UMAP as an alternative to t-SNE for single-cell data. *bioRxiv* 298430 (2018) doi:10.1101/298430.

120. Rapino, F. *et al.* C/EBPα induces highly efficient macrophage transdifferentiation of B lymphoma and leukemia cell lines and impairs their tumorigenicity. *Cell Rep.* **3**, 1153–1163 (2013).

121. Gaidt, M. M. *et al.* Human Monocytes Engage an Alternative Inflammasome Pathway. *Immunity* **44**, 833–846 (2016).

122. Schmidt, T., Schmid-Burgk, J. L. & Hornung, V. Synthesis of an arrayed sgRNA library targeting the human genome. *Sci. Rep.* **5**, 14987 (2015).

123. Linder, A. *et al.* CARD8 inflammasome activation triggers pyroptosis in human T cells. *EMBO J.* **39**, (2020).

124. Schmid-Burgk, J. L. *et al.* OutKnocker: a web tool for rapid and simple genotyping of designer nuclease edited cell lines. *Genome Res.* **24**, 1719–1723 (2014).

125. DeAngelis, M. M., Wang, D. G. & Hawkins, T. L. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* **23**, 4742–4743 (1995).

126. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *GigaScience* **7**, (2018).

127. Girardot, C., Scholtalbers, J., Sauer, S., Su, S.-Y. & Furlong, E. E. M. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinformatics* **17**, 419 (2016).

128. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

129. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).

130. R core team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, 2020).

131. RStudio Team. *RStudio: Integrated Development for R.* (RStudio, 2020).

132. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

133. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

134. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer International Publishing, 2016). doi:10.1007/978-3-319-24277-4.

135. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**, 137–145 (2020).

136. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

137. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **65**, 631-643.e4 (2017).

138. Power Analysis of Single Cell RNA-Sequencing Experiments. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5376499/.

139. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).

140. Liu, T., Zhang, L., Joo, D. & Sun, S.-C. NF-κB signaling in inflammation. *Signal Transduct. Target. Ther.* **2**, 1–9 (2017).

141. Lu, Y.-C., Yeh, W.-C. & Ohashi, P. S. LPS/TLR4 signal transduction pathway. *Cytokine* **42**, 145–151 (2008).

142. Ivashkiv, L. B. & Donlin, L. T. Regulation of type I interferon responses. *Nat. Rev. Immunol.* **14**, 36–49 (2014).

143. Newton, K. & Dixit, V. M. Signaling in Innate Immunity and Inflammation. *Cold Spring Harb. Perspect. Biol.* **4**, (2012).

144. Chow, J., Franz, K. M. & Kagan, J. C. PRRs are watching you: Localization of innate sensing and signaling regulators. *Virology* **0**, 104–109 (2015).

145. Motwani, M., Pesiridis, S. & Fitzgerald, K. A. DNA sensing by the cGAS–STING pathway in health and disease. *Nat. Rev. Genet.* **20**, 657–674 (2019).

146. Oivanen, M., Kuusela, S. & Lönnberg, H. Kinetics and Mechanisms for the Cleavage and Isomerization of the Phosphodiester Bonds of RNA by Brønsted Acids and Bases. *Chem. Rev.* **98**, 961–990 (1998).

147. Dallas, A., Vlassov, A. V. & Kazakov, S. A. Principles of Nucleic Acid Cleavage by Metal Ions. in *Artificial Nucleases* (ed. Zenkova, M. A.) vol. 13 61–88 (Springer Berlin Heidelberg, 2004).

148. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).

149. *Data Analysis for Omic Sciences: Methods and Applications*. (Elsevier, 2018).

150. Thomas, P. D. *et al.* PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Res.* **13**, 2129–2141 (2003).

151. Svensson, V. *et al.* Power Analysis of Single Cell RNA-Sequencing Experiments. *Nat. Methods* **14**, 381–387 (2017).

152. Daniel, Z., Peter, A., Mikael, K. & Lukas, V. Performance comparison of reverse transcriptases for single-cell studies. *bioRxiv* 629097 (2019) doi:10.1101/629097.

153. Zimmerman, S. B. & Pheiffer, B. H. Macromolecular crowding allows blunt-end ligation by DNA ligases from rat liver or Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **80**, 5852–5856 (1983).

154. Marguerat, S. & Bähler, J. Coordinating genome expression with cell size. *Trends Genet.* **28**, 560–565 (2012).

155. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).

156. Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun.* **11**, 774 (2020).

157. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6**, 25533 (2016).

158. Tandonnet, S. & Torres, T. T. Traditional versus 3′ RNA-seq in a non-model species. *Genomics Data* **11**, 9–16 (2016).

159. Delhase, M., Hayakawa, M., Chen, Y. & Karin, M. Positive and Negative Regulation of IκB Kinase Activity Through IKKβ Subunit Phosphorylation. *Science* **284**, 309–313 (1999).

160. Israël, A. The IKK Complex, a Central Regulator of NF-κB Activation. *Cold Spring Harb. Perspect. Biol.* **2**, (2010).

161. Takeuchi, O., Hemmi, H. & Akira, S. Interferon response induced by Toll-like receptor signaling. *J. Endotoxin Res.* **10**, 252–256 (2004).

162. Kane, M. *et al.* Identification of Interferon-Stimulated Genes with Antiretroviral Activity. *Cell Host Microbe* **20**, 392–405 (2016).

163. Schoggins, J. W. & Rice, C. M. Interferon-stimulated genes and their antiviral effector functions. *Curr. Opin. Virol.* **1**, 519–525 (2011).

164. Svensson, V., Beltrame, E. da V. & Pachter, L. Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. *bioRxiv* 762773 (2019) doi:10.1101/762773.

165. Mosmann, T. R., Cherwinski, H., Bond, M. W., Giedlin, M. A. & Coffman, R. L. Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *J. Immunol. Baltim. Md 1950* **136**, 2348–2357 (1986).

166. Sekiya, T. & Yoshimura, A. In Vitro Th Differentiation Protocol. in *TGF-β Signaling: Methods and Protocols* (eds. Feng, X.-H., Xu, P. & Lin, X.) 183–191 (Springer, 2016). doi:10.1007/978-1-4939-2966-5_10.

167. Swain, S. L. T-Cell Subsets: Who does the polarizing? *Curr. Biol.* **5**, 849–851 (1995).

168. Basu, S., Campbell, H. M., Dittel, B. N. & Ray, A. Purification of Specific Cell Population by Fluorescence Activated Cell Sorting (FACS). *J. Vis. Exp. JoVE* (2010) doi:10.3791/1546.

169. Tung, P.-Y. *et al.* Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).

170. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, (2019).

171. Golstein, P. & Griffiths, G. M. An early history of T cell-mediated cytotoxicity. *Nat. Rev. Immunol.* **18**, 527–535 (2018).

172. Navarro, M. N. & Cantrell, D. A. Serine-threonine kinases in TCR signaling. *Nat. Immunol.* **15**, 808–814 (2014).

173. Isakov, N. & Altman, A. Protein Kinase Cθ in T Cell Activation. *Annu. Rev. Immunol.* **20**, 761–794 (2002).

174. Ai, W., Li, H., Song, N., Li, L. & Chen, H. Optimal Method to Stimulate Cytokine Production and Its Use in Immunotoxicity Assessment. *Int. J. Environ. Res. Public. Health* **10**, 3834–3842 (2013).

175. Green, D. R., Droin, N. & Pinkoski, M. Activation-induced cell death in T cells. *Immunol. Rev.* **193**, 70–81 (2003).

176. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

177. Somerville, T. D. D. *et al.* ZBED2 is an antagonist of interferon regulatory factor 1 and modifies cell identity in pancreatic cancer. *Proc. Natl. Acad. Sci.* **117**, 11471–11482 (2020).

178. Weitzman, J. B., Fiette, L., Matsuo, K. & Yaniv, M. JunD protects cells from p53-dependent senescence and apoptosis. *Mol. Cell* **6**, 1109–1119 (2000).

179. Wang, D. The Transcription Factor Runx3 Establishes Chromatin Accessibility of cis-Regulatory Landscapes that Drive Memory Cytotoxic T Lymphocyte Formation. 23.

180. Chen, L.-F. Tumor suppressor function of RUNX3 in breast cancer. *J. Cell. Biochem.* **113**, 1470–1477 (2012).

181. Lin, C.-C. *et al.* Bhlhe40 controls cytokine production by T cells and is essential for pathogenicity in autoimmune neuroinflammation. *Nat. Commun.* **5**, 3551 (2014).

182. Konopacki, C., Pritykin, Y., Rubtsov, Y., Leslie, C. S. & Rudensky, A. Y. Transcription factor Foxp1 regulates Foxp3 chromatin binding and coordinates regulatory T cell function. *Nat. Immunol.* **20**, 232–242 (2019).

183. Sedlmaier, A. *et al.* Overexpression of hepatoma-derived growth factor in melanocytes does not lead to oncogenic transformation. *BMC Cancer* **11**, 457 (2011).

184. Jennings, E. *et al. Differential Nr4a1 and Nr4a3 expression discriminates tonic from activated TCR signalling events in vivo*. http://biorxiv.org/lookup/doi/10.1101/767566 (2019) doi:10.1101/767566.

185. Odagiu, L. *et al.* Early programming of CD8+ T cell response by the orphan nuclear receptor NR4A3. *Proc. Natl. Acad. Sci.* **117**, 24392–24402 (2020).

186. O'Shea, J. J. & Paul, W. E. Regulation of T H I differentiation – controlling the controllers. *Nat. Immunol.* **3**, 506–508 (2002).

187. Jenner, R. G. *et al.* The transcription factors T-bet and GATA-3 control alternative pathways of T-cell differentiation through a shared set of target genes. *Proc. Natl. Acad. Sci.* **106**, 17876–17881 (2009).

188. Kaminuma, O. *et al.* GATA-3 suppresses IFN-gamma promoter activity independently of binding to cis-regulatory elements. *FEBS Lett.* **570**, 63–68 (2004).

189. Bhat, M. Y. *et al.* Comprehensive network map of interferon gamma signaling. *J. Cell Commun. Signal.* **12**, 745–751 (2018).

190. Schroder, K., Hertzog, P. J., Ravasi, T. & Hume, D. A. Interferon-γ: an overview of signals, mechanisms and functions. *J. Leukoc. Biol.* **75**, 163–189 (2004).

191. Kawka, E. *et al.* Regulation of chemokine CCL5 synthesis in human peritoneal fibroblasts: a key role of IFN-γ. *Mediators Inflamm.* **2014**, 590654 (2014).

192. Galli, E. *et al.* GM-CSF and CXCR4 define a T helper cell signature in multiple sclerosis. *Nat. Med.* **25**, 1290–1300 (2019).

193. Patil, V. S. *et al.* Precursors of human CD4+ cytotoxic T lymphocytes identified by single-cell transcriptome analysis. *Sci. Immunol.* **3**, (2018).

194. Tibbitt, C. A. *et al.* Single-Cell RNA Sequencing of the T Helper Cell Response to House Dust Mites Defines a Distinct Gene Expression Signature in Airway Th2 Cells. *Immunity* **51**, 169-184.e5 (2019).

195. Rasouli, J. *et al.* A distinct GM-CSF+ T helper cell subset requires T-bet to adopt a TH1 phenotype and promote neuroinflammation. *Sci. Immunol.* **5**, (2020).

196. Kiner, E. *et al.* Gut CD4 + T cell phenotypes are a continuum molded by microbes, not by T H archetypes. *Nat. Immunol.* **22**, 216–228 (2021).

197. Philips, R. M. & R. » How big is a human cell? http://book.bionumbers.org/how-big-is-a-human-cell/.

198. Hamilton, J. A. Colony-stimulating factors in inflammation and autoimmunity. *Nat. Rev. Immunol.* **8**, 533–544 (2008).

199. Sheng, W. *et al.* STAT5 programs a distinct subset of GM-CSF-producing T helper cells that is essential for autoimmune neuroinflammation. *Cell Res.* **24**, 1387–1402 (2014).

200. Bratke, K., Kuepper, M., Bade, B., Virchow, J. C. & Luttmann, W. Differential expression of human granzymes A, B, and K in natural killer cells and during CD8+ T cell differentiation in peripheral blood. *Eur. J. Immunol.* **35**, 2608–2616 (2005).

201. Suni, M. A. *et al.* CD4(+)CD8(dim) T lymphocytes exhibit enhanced cytokine expression, proliferation and cytotoxic activity in response to HCMV and HIV-1 antigens. *Eur. J. Immunol.* **31**, 2512–2520 (2001).

202. Takeuchi, A. & Saito, T. CD4 CTL, a Cytotoxic Subset of CD4+ T Cells, Their Differentiation and Function. *Front. Immunol.* **8**, (2017).

203. De Giovanni, M. *et al.* Spatiotemporal regulation of type I interferon expression determines the antiviral polarization of CD4 + T cells. *Nat. Immunol.* **21**, 321–330 (2020).

204. Pennock, N. D. *et al.* T cell responses: naïve to memory and everything in between. *Adv. Physiol. Educ.* **37**, 273–283 (2013).

205. Parmar, K., Blyuss, K. B., Kyrychko, Y. N. & Hogan, S. J. Time-Delayed Models of Gene Regulatory Networks. *Comput. Math. Methods Med.* **2015**, (2015).

206. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

207. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).

208. Geginat, J. *et al.* Plasticity of Human CD4 T Cell Subsets. *Front. Immunol.* **5**, (2014).

209. Zhu, J. *et al.* Conditional deletion of Gata3 shows its essential function in T(H)1-T(H)2 responses. *Nat. Immunol.* **5**, 1157–1165 (2004).

210. Zhuang, Y. *et al.* A continuous T-bet expression is required to silence the interleukin-4-producing potential in T helper type 1 cells. *Immunology* **128**, 34–42 (2009).

211. Hoefig, K. P. & Heissmeyer, V. Posttranscriptional regulation of T helper cell fate decisions. *J. Cell Biol.* **217**, 2615–2631 (2018).

212. Konno, S. *et al.* Interferon-gamma enhances rhinovirus-induced RANTES secretion by airway epithelial cells. *Am. J. Respir. Cell Mol. Biol.* **26**, 594–601 (2002).

213. Liu, J., Guan, X. & Ma, X. Interferon regulatory factor 1 is an essential and direct transcriptional activator for interferon {gamma}-induced RANTES/CCl5 expression in macrophages. *J. Biol. Chem.* **280**, 24347–24355 (2005).

214. Millward, J. M., Caruso, M., Campbell, I. L., Gauldie, J. & Owens, T. IFN-γ-Induced Chemokines Synergize with Pertussis Toxin to Promote T Cell Entry to the Central Nervous System. *J. Immunol.* **178**, 8175–8182 (2007).

215. Ferber, I. A. *et al.* GATA-3 significantly downregulates IFN-gamma production from developing Th1 cells in addition to inducing IL-4 and IL-5 levels. *Clin. Immunol. Orlando Fla* **91**, 134–144 (1999).

216. Maggi, E. *et al.* Reciprocal regulatory effects of IFN-gamma and IL-4 on the in vitro development of human Th1 and Th2 clones. *J. Immunol.* **148**, 2142–2147 (1992).

217. Peine, M. *et al.* Stable T-bet+GATA-3+ Th1/Th2 Hybrid Cells Arise In Vivo, Can Develop Directly from Naive Precursors, and Limit Immunopathologic Inflammation. *PLOS Biol.* **11**, e1001633 (2013).

218. Wei, G. *et al.* Global Mapping of H3K4me3 and H3K27me3 Reveals Specificity and Plasticity in Lineage Fate Determination of Differentiating CD4+ T Cells. *Immunity* **30**, 155–167 (2009).

219. Saravia, J., Chapman, N. M. & Chi, H. Helper T cell differentiation. *Cell. Mol. Immunol.* **16**, 634–643 (2019).

220. Kourelis, T. V. *et al.* Mass cytometry dissects T cell heterogeneity in the immune tumor microenvironment of common dysproteinemias at diagnosis and after first line therapies. *Blood Cancer J.* **9**, 1–13 (2019).

221. Barcenilla, H., Åkerman, L., Pihl, M., Ludvigsson, J. & Casas, R. Mass Cytometry Identifies Distinct Subsets of Regulatory T Cells and Natural Killer Cells Associated With High Risk for Type 1 Diabetes. *Front. Immunol.* **10**, (2019).

222. Hartmann, F. J. *et al.* Single-cell metabolic profiling of human cytotoxic T cells. *Nat. Biotechnol.* **39**, 186–197 (2021).

8 List of abbreviations

AB – Antibody
AICD – Activation induced cell death
APC – Antigen presenting cell
bp – base pair
CCL – C-C Motif Chemokine Ligand
CD – Cluster of differentiation
cDNA – Complementary DNA
cGAMP – Cyclic guanosine monophosphate–adenosine monophosphate
cGAS – Cyclic GMP-AMP synthase
CSF – Colony stimulating factor
CXCL – C-X-C motif chemokine ligand
DC – Dendritic cell
DDT – Dichlorodiphenyltrichloroethane
DE – Differential expression
DGE – Digital gene expression
DNA – Deoxyribonucleic acid
dNTPS – Deoxyribo nucleoside triphosphates
DP – Double positive
DR – Dimensionality reduction
dsDNA – double-stranded DNA
EDTA – Ethylenediaminetetraacetic acid
ELISA – Enzyme linked immunosorbent assay
EST – Expressed sequence tag
FACS – Fluorescence-activated cell sorting
FCS – Fetal calf serum
FDR – False discovery rate
GATA3 – GATA binding protein 3
gDNA – Genomic DNA
GFP – Green fluorescent protein
GNLY – Granylosin
GO – Gene ontology
GZM – Granzyme
HEPES – 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HVG – Highly variable gene
IFN – Interferon
CHUCK – (IKKA) Inhibitor of nuclear factor kappa B kinase complex
HT-DNA – Herring Testes DNA
IKBKB – Inhibitor of nuclear factor kappa B kinase complex subunit beta (*IKBKβ)*

IKBKE – Inhibitor of nuclear factor kappa B kinase complex subunit epsilon
IL – Interleukin
iPSCs – Induced pluripotent stem cell
IRF – Interferon regulatory factor
IVT – In vitro transcription
Kb – Kilo bases
KO – Knockout
LDR – Linear dimensionality reduction
lncRNA – long-noncoding RNA
LPS – Lipopolysaccharide
MACS – Magnetic assisted cell sorting
MAMP – Microbe associated molecular patterns
MARS-seq – Massively Parallel Single-Cell RNA-Seq
Mb – Mega bases
mcSCRB-seq – Molecular crowding SCRB-seq
MD2 – Myeloid Differentiation factor 2
MHC – Major histocompatibility complex
miRNA – micro RNA
mRNA – Messanger RNA
MYD88 – Myeloid Differentiation Primary Response gene 88
NDR – Nonlinear dimensionality reduction
NGS – Next generation sequencing
GO – Gene Ontology
OASL – Oligoadenylate Synthetase Like
ONT – Oxford nanopore technologies
ORF – Open reading frame
PBMCs – Peripheral blood mononuclear cell
PBS – Phosphate buffered saline
PCA – Principal component analysis
PCR – Polymerase chain reaction
PEG – Polyetylene glycol
PMA – phorbol-12-myristate-13-acetate
PRR – Pattern recognition receptor
QC – Quality control
qPCR – Quantitative PCR
RNA – Ribonucleic acid
RNA-seq – RNA sequencing
RORC – RAR-related orphan receptor C
RT – Reverse Transcription
RT-qPCR – Reverse Transcription quantitative PCR
SCRB-seq – Single-cell RNA barcoding and sequencing
SEM – Standard error of the mean

siRNA – small interfering RNA
SMART – Switching Mechanism at the 5′ end of RNA Template
SMRT – Single molecule real time
SNN – Sharing nearest neighbor
SNPs – Single nucleotide polymorphisms
SP – Single positive
SPRI – Solid Phase Reversible Immobilisation
STING – Stimulator of interferon genes
TAE – tris, acetic acid and EDTA
TBK1 – TANK-binding kinase 1
TBX21 – T-Box Transcription Factor 21
TCR – T cell receptor
TE – Tris EDTA
T_{FH} – T follicular helper cells
T_H1 – T helpers type 1
T_H17 – T helpers type 17
T_H2 – T helpers type 2
T_H9 – T helpers type 9
TI – Trajectory inference
TIR – Toll-interleukin-1 receptor
TLR – Toll like receptor
TNF – Tumor necrosis factor
TRAF – TNF Receptor Associated Factor
TREG – T regulatory cell
TICAM1 – (TRIF) TIR-domain-containing adapter-inducing interferon-β
tRNA – Transport RNA
tSNE – t-distributed stochastic neighbor embedding
TSO – Template switching oligo
UMAP – Uniform Manifold Approximation and Projection
UMI – Unique molecular identifier

9 Acknowledgments

First of all, I would like to thank Veit Hornung for finding me a place in the lab and having time to help me with my scientific endeavors. Without his guidance, help, and problem-solving skills, I would not have managed to get this far! Of course, it would be remiss of me not to mention all the social activities he encouraged us to undertake, ski holidays, hikes, and summer retreats.

I would also like to express my gratitude to Lucas Jae for agreeing to be the second reviewer on my defense committee.

Furthermore, I would like to thank all my scientific collaborators, especially Thomas and Andreas, for working with me over the years and providing me with lots of interesting samples to study. Also, Christoph Ziegenhain, Johannes Bagnoli, and the rest of AG Enard, whose help with troubleshooting was immeasurable.

I would also like to thank all my previous and current colleagues who made working in this lab awesome! I will never forget the fun times we had.

I especially want to thank:

Dennis for his help and cooperation, but especially for all the cakes and cookies he baked!

Fionan for being the best office mate and for occasionally taking me snowboarding!

Che for his help and scientific advice; I’ve never met anyone more enthusiastic about science.

I would also like to thank the Graduate school of Quantitative Biosciences Munich for funding my stipend for the first half of my Ph.D. and for organizing some of the best conferences and retreats ever!

I also want to thank my parents, who have always encouraged and supported me in every way possible!

Lastly, I want to thank Karl.

You helped me survive the rough times and made the good times great! I am lucky to have you in my life!