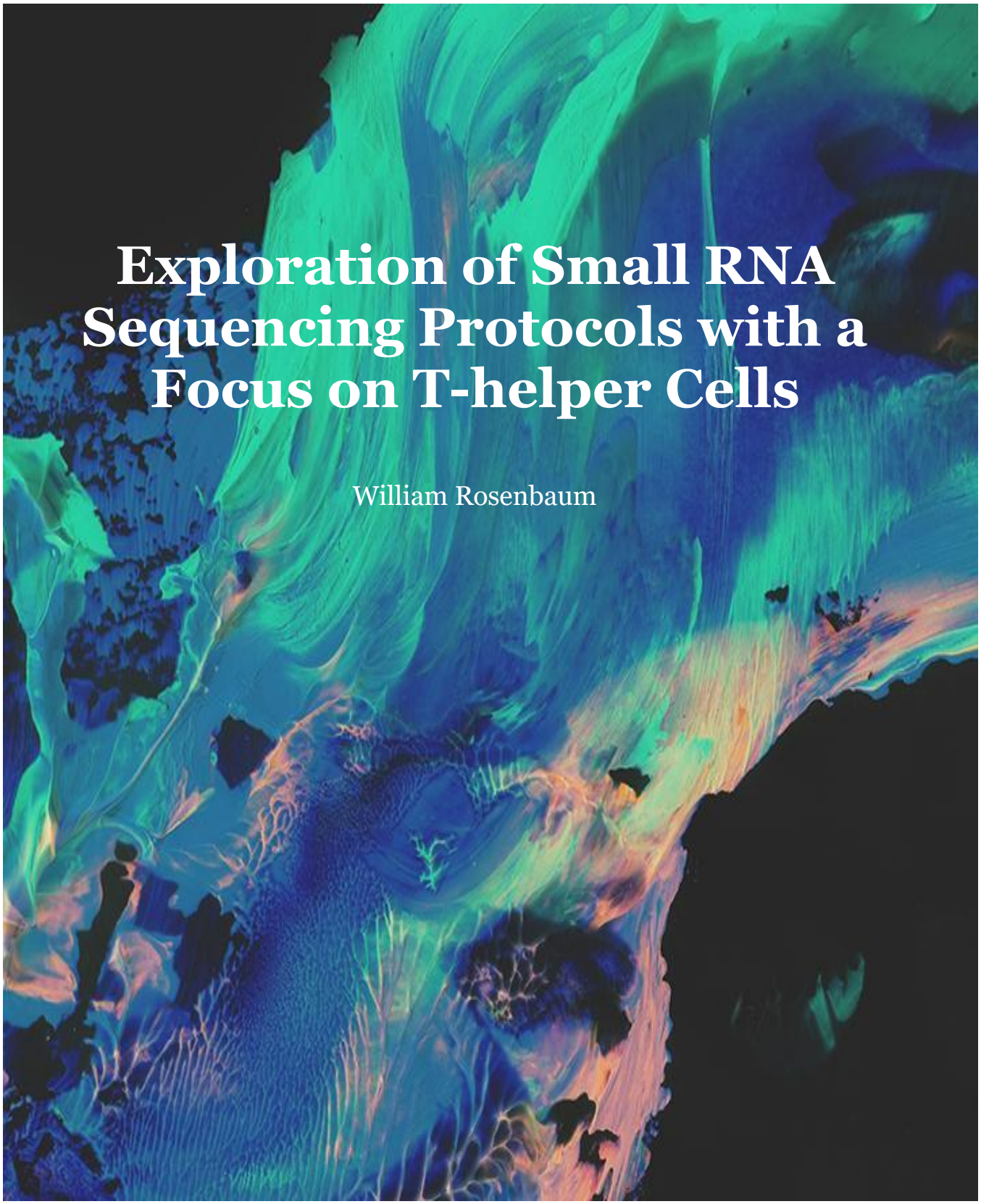
UMEÅ UNIVERSITY



Exploration of Small RNA Sequencing Protocols with a Focus on T-helper Cells

William Rosenbaum

Umeå University, 901 87 Umeå, Sweden
Master Degree Thesis Project in Molecular Biology, 30 ECTS
Date: 20220314
Performed at: MIMS
Supervisor: Johan Henriksson

Abstract

T helper cells contribute to the immune system by maintaining and sustaining antitumor potency. Utilizing the capacity of T helper cells in cancer immunotherapies would aid to enhance the efficiency of these treatments. Reliable methods which support the categorization and knowledge about T helper cell differentiation are therefore of importance. Evidence suggests that microRNAs can be used for T helper cell categorization. The most informative method for microRNA profiling is sequencing. However, microRNA sequencing is associated with innate problems, such as adapter dimer abundance and computational obstacles. Efforts to address both of these issues were made in this master thesis. To investigate the microRNA profile in various kinds of T helper cells by sequencing, naïve CD4+ cells were isolated from blood and differentiated into designated subtypes. Exploration of different microRNA sequencing protocols showed that methods using magnetic beads, compared to time-consuming gel electrophoresis, is preferable to discard unwanted adapter dimers. To meet the challenge with reproducibility related to small RNA analysis, the Python package gentools was developed. By default, gentools uses software and parameters optimized for small RNA analysis. Reanalyzing public available small RNA sequencing data with gentools resulted in different interpretations compared to analysis done with other pipelines. This emphasizes the necessity of a standardized pipeline for small RNA analysis. The results presented in this thesis could help progress the development of small RNA sequencing protocols, both experimentally and computationally. This can improve methodology regarding T helper cell categorization, which could aid improving immunotherapies and other cancer treatments.

Author Contributions

Johan Henriksson supervised and conceived the study.

Sebastian Mihai did the first CD4+ differentiation and activation according to a protocol optimized by himself. He also performed some of the Bioanalyzer inspections. He did the first extraction of PBMC cells from blood.

Martin Selinger helped and supervised the qPCR. He helped with cell culture and PBMC extraction.

Hayoung Lee read the thesis and suggested improvements.

Kristina Krank helped with the cell culture.

SciLifeLab performed the sequencing.

The author (William Rosenbaum) participated in all and performed almost all practical work and wrote the whole thesis. He also conceived the idea of gentools, implemented and wrote the code and tested the program.

Introduction

The use of T cells with chimeric antigen receptors (CARs) have in recent years expanded as one of the driving therapeutic treatments in cancer immunotherapies. To increase the efficacy of immunotherapies, studies have focused on improving the cytotoxic activity of CD8+ cells [1,2]. Although the cytotoxic capacity of CD8+ cells is fundamental for generating successful antitumor immunity, CD4+ T helper cells (Th) are essential to maintain, sustain and even contribute directly to the antitumor potency [1,2]. By releasing cytokines, Th cells contribute to the adaptive immune system via upregulation of CD8+ cytolytic activity and antibody production in B cells [1,3–6]. To become activated, a naïve Th cell must have its T cell receptor stimulated together with the impact of other activation signals. Co-stimulatory molecules and cytokines bind to the cell in question and influence the trajectory of differentiation the activated Th cell follows [2,5–7]. Depending on the context, naïve Th cells differentiate into various subtypes, *e.g.* T helper type 1 cells (Th1), Th2, Th17 and regulatory T cells (Treg), all with unique functions and attributes [3,4,7].

Historically, Th cells have been categorized through their cytokine and interleukin (IL) production. For example, Th1 can be recognized through high expression of the transcription factor (TF) *Tbx21*, *IFN-γ*, IL-2 and cytokines [3,5,7]. Th2, on the other hand, expresses *IL-4* and the TF *GATA3* [5,7]. Th17 expresses *IL-17* and the retinoic acid–related orphan receptor *RORγ*, but not *GATA3* [3,5,8]. Treg cells are mostly characterized by the expression of *FOXP3* [3,5]. However, the distinct division between various subtypes of Th cells is not as clear as previously thought. Due to substantial heterogeneity within every subtype, expression of cytokines may vary dramatically between similar groups and for different points in time [3]. Also, an increased number of studies suggest that previously polarized Th cells exhibit a higher capacity of plasticity than previously thought. Many studies show that already polarized Th cells can undergo reprogramming upon cytokine activation, resulting in a different kind of subtype and phenotype [3–5]. Large differences between activation of Th cells *in vitro* and *in vivo* have also been reported, which questions the stringency of categorization of well-defined subtypes. Is the division between subtypes real, or is it an artefact derived from *in vitro* studies [3]? The aforementioned statements help to explain the ambiguity regarding Th cell classification. In turn, this affects our knowledge of activation and differentiation pathways in T helper cells, which is crucial for a deeper understanding of the adaptive immune system [7]. Therefore, novel tools of Th cell classification need to be developed and classification of Th subtypes solely based on gene expression needs to be scrutinized further [3–5,7].

Besides protein-coding RNA, small RNAs such as microRNAs (miRNA), play a central role in Th cell activation and function [3,6,9–13]. The group of miRNAs are represented by 19-24 nucleotide long single stranded sequences, which act as translational repressors [6,10,11,14]. The miRNAs work by binding to the 3'-UTR end of encoding RNA. This interaction causes the receiving strand to degrade and become non-functional, resulting in a down regulation of protein expression [6,11,12,14,15]. In this way, miRNAs affect gene expression, and therefore function, of activated and differentiated CD4+ cells [6,10,11]. Therefore, more knowledge about the specific composition of the miRNA pool in different Th cells would allow for more accurate classification of various subtypes, as well as the possibility for manipulation of desired Th cell properties necessary for enhancing the effects of immunotherapies [6].

RNA-sequencing (RNA-seq) is one of the most common techniques for miRNA profiling [13,15]. RNA-seq quantifies and detects novel isoforms of small RNA species in an untargeted way, compared to microarrays and reverse transcription-quantitative PCR (RT-qPCR) [13,15,16]. Although many different protocols for miRNA sequencing exist, all of them are associated with different technical difficulties. One overarching problem in most protocols is the formation of large amounts of adapter dimers in relation to the

desired library [13,15–18]. Due to the small size difference between the wanted product and the adapter dimers, the two distributions can be hard to distinguish from each other [16,17]. The presence of adapter dimers dilutes the overall sequencing library concentration, and can lead to complications during sequencing, or in the downstream analysis [17]. To circumvent this problem, many protocols involve size selection using gel electrophoresis and recovery of the desired library from a band cut from the gel, also known as the crush and soak method [16,17,19]. However, the gel-based methods come with innate limitations. First, a large proportion of the fragments are lost during the recovery process. Second, the method is technically demanding and time consuming, resulting in a low level of reproducibility. These limitations hinder the up-scaling of the protocol for automated workflows [16,17].

Due to the short length of small RNA, difficulties also arise during the analysis stage of the experiment [20]. Differences in miRNA expression between groups, based on counting and differential expression analysis, rely on mapping of the raw reads to a reference genome [20,21]. The repetitive nature of genomes in addition to high levels of non-templated modifications of small RNAs, complicate the alignment of small RNA to the reference genome. Many different alignment software exists, and the choice of program and parameters affect the outcome and downstream analysis of the small RNA data. These differences can affect the reproducibility of the analysis, which makes the comparison of analyses between different studies difficult [20,21]. Still, no consensus on how to analyze miRNA sequencing data to increase the reproducibility exists, even though the aforementioned problems are widely known [21].

To utilize the miRNA expression profile to further refine categorization of Th subtypes, reliable tools for small RNA-seq and downstream analysis of the sequencing data are needed. Unveiling hidden information in the miRNA expression repertoire in activated CD4+ cells could potentially result in novel ways to manipulate Th cells to enhance immunotherapies. The aim of this master thesis was to analyze the small RNA expression profile in activated Th0, Th1, Th2, Th17 and Treg subtypes. For convenience, the small RNA library preparation method used in this master thesis was based on the Small-seq protocol [9,19]. To improve the Small-seq protocol, the removal of adapter dimers and other contaminating fragments were done, using methods other than the gel-based crush and soak procedure. Additionally, to address the lack of reproducibility in examination of miRNA sequencing experiments, and to make the data analysis available for users with limited bioinformatics experience, a command line program built in Python was developed (https://github.com/willros/gentools). This tool, gentools, uses software and algorithms recommended by previous studies, which base their results on known concentration and abundance of miRNA from dilution series of spike-in oligos [20,21]. To obtain differentiated Th cells, CD4+ cells were isolated from blood and activated using different cytokines and antibodies. Libraries from extracted miRNA were produced using a modified version of the Small-seq protocol, and sequencing was performed by SciLifeLab. However, due to technical difficulties, sequencing failed, and no data were obtained. To assess the performance of gentools, single-cell miRNA sequencing data published by Faridani et al. 2016 [9], were analyzed and examined. Because gentools relies on different thresholds and algorithms, the reanalysis differs vastly from the result presented in the original paper, emphasizing the necessity for a standard analysis pipeline.

Materials and Methods

Isolation of peripheral blood mononuclear cells

Blood samples were obtained from four healthy male donors. Peripheral blood mononuclear cells (PBMC) were isolated using Ficoll-Paque PLUS density gradient centrifugation, as described elsewhere [14]. Briefly, fresh blood was mixed with Ficoll-

Paque medium and centrifuged at 400 rcf for 40 minutes at 20°C. The mononuclear cell layer was washed with 3 x volume of phosphate buffered-saline (PBS) and centrifuged at 400 rcf for 20 minutes at 20°C. The washing was repeated two times. Finally, the isolated PBMC were cryopreserved in 90% fetal bovine serum and 10% dimethyl sulfoxide (DMSO) medium, and stored at -150°C.

Activation and differentiation of T cells

Isolated and cryopreserved PBMC cells from the four donors were thawed and resuspended in PBS. The cells were counted and centrifuged at 450 g for 5 minutes at 20°C. DMSO was washed out from the pellet by carefully resuspending the pellet into 5 mL of PBS. The cells were centrifuged again at 450 rcf for 5 minutes at 20°C, and the pellet was resuspended in 2 mL of MACS® BSA Stock Solution (Miltenyi Biotec) to a concentration of 50 x 10^6 cells/mL. The cells from each donor were pooled together for the next steps.

CD4+ T cells were isolated from the PBMC using the EasySep™ Human Naïve CD4+ T Cell Isolation Kit (STEMCELL) following the manufacturer's instructions. The EasySep™ Human Naïve CD4+ T Cell Isolation Kit relies on negative selection, which prevents pre-activation of the CD4+ cells. In brief, the cells were mixed with biotinylated antibody beads and were isolated using a magnet. The isolated CD4+ T cells were centrifuged and resuspended in ImmunoCult™-XF T Cell Expansion Medium mixed with different cytokines and antibodies (view Table S1 for a full list of molecules and concentrations used). Th0 cells were generated in the absence of cytokines. In addition, cells left in only medium were also cultivated, as described earlier [4]. A cell culture plate was coated with 100 µL of Anti-CD3 antibodies (Biolegend, 317326) (100 mg/µL) and incubated at 37°C for 2 hours prior to seeding 200 µL of respective cell solutions to the wells. The plate was incubated at 37°C for five days before the activated Th cells were harvested.

RNA extraction and quality control

After five days of stimulation, polarized Th0, Th1, Th2, Th17 and Treg cells were harvested and counted. Large RNA and miRNA from about 5 x 10^6 cells with a mean viability of 77% were extracted, using the E.Z.N.A.® Micro RNA Kit, following the manufacturer's instructions. Briefly, the RNA was extracted into different fractions using columns and several centrifugation steps. Concentration of the RNA was measured with the NanoDrop™ 2000 spectrophotometer. The sizes of the fragments were analyzed using an Agilent 2100 Bioanalyzer RNA Pico Kit.

Small RNA sequencing library preparation with Small-seq protocol

The protocol used for preparing the small RNA sequencing library was described by Hagemann-Jensen et al.[19]. However, some modifications were made to the protocol. First and foremost, the miRNA sequencing was performed on bulk RNA input, and not single cell as described in the protocol. Second, the small RNA extraction was performed with E.Z.N.A columns.
An RNA library preparation was made for five different biological samples: Th0, Th1, Th2, Th17 and Treg and technical duplicates were made for Th0 and Treg. Water was used as a negative no-cell control. Each reaction was performed in a 0.2 mL PCR tube. Oligonucleotides and primers that were used can be seen in Table 1.

The protocol followed is described in detail in Hagemann-Jensen et al., 2018 and Faridani et al., 2016 [9,19]. First, 2 µL of the 3' adapter ligation mix (2 µM RA3, 8% PEG8000, 0.8x T4 RNA ligase reaction buffer, 10 units/ µL T4 RNA ligase 2, truncated KQ and 0.8 units/µL of Recombinant RNase inhibitor) were added to each sample. The reactions were vortexed and incubated at 30°C for 6 hours and 4°C for 10 hours. To digest potential free 3' adapters, 3 µL of 3' adapter digestion mix (5 µM reverse

transcription primer (RTP), 0.3 units/µL Lambda exonuclease, and 3.12 units/µL of 5′ deadenylase) were added to each PCR tube. The tubes were vortexed and incubated at 30°C for 15 minutes and 37°C for an additional 15 minutes. After this, the 5' adapters were ligated by adding 2 µL of the 5' adapter ligation mix (1 µM RA5, 0.7 mM ATP, Tris-buffered, 0.25x T4 RNA ligase reaction buffer, and 0.45 units/µL of T4 RNA ligase 1), and the samples were incubated at 30°C for 1 hour. Next, reverse transcription (RT) was accomplished through addition of 5 µL RT reaction mix (1.3x Taq DNA polymerase buffer, 8 mM dithiothreitol (DTT), 0.5 mM of each dNTP, 0.27 units/µL Recombinant RNase inhibitor, and 6.67 units/µL SuperScript II Reverse Transcriptase), followed by incubation at 42°C for 1 hour and 70°C for 15 minutes. After RT, the first PCR amplification was performed. A 10 µL volume of 1x Phusion HF buffer, 0.04 units/µL Phusion Hot Start II, 0.15 mM of each dNTP and 1 µM RP1 was added. The tubes were placed in a thermocycler and the following program was completed: 98°C for 30 seconds, 13 cycles of 98°C for 10 seconds, 60°C for 30 seconds and 72°C for 30 seconds and finally 72°C for 5 minutes. Lastly, 1 µL of the first PCR product was mixed with 23 µL of 1x Phusion HF buffer, 0.02 units/µL Phusion Hot Start II, 0.2 mM of each dNTP and 0.8 µM of RP1. To barcode each sample, 1 µL of either SR0001-SR0008 (Table 1) was added to each reaction before the following PCR program was performed: 30 seconds incubation at 98°C, 13 cycles of 98°C for 10 seconds, 67°C for 30 seconds and 72°C for 30 second and finally 72°C for 5 minutes [9 19]. The final PCR products were then purified using Zymo Clean & Concentrator Kit-5. After cleaning, the libraries were analyzed using a Bioanalyzer and Qubit, and then pooled together in equimolar concentrations. The pooled library was again analyzed on the Bioanalyzer and the libraries was diluted to 20 nM according to the following formula: Molarity (nM) = (concentration (ng/µL x 10^6) / (Average molecular length [bp] x 660 [g/mol]). Ultimately, the library was sent to SciLifeLab, for sequencing on an Illumina MiSeq 2000 (150 SE, v4 chemistry). For a more detailed depiction of the method, see https://github.com/willros/master_thesis/blob/main/smallseq_method.html.

Table 1. *Primers and adapters used in the library preparation*. The sequence highlighted in bold represents different barcodes. All adapters and oligos were designed and first used by Hagemann-Jensen et al [19]. The adapters and oligos were ordered from IDTDNA.

Name	**Sequence (5' to 3')**
5' adapter	NH2-rGrUrUrCrArGrArGrUrUrCrUrArCrArGrUrCrCrGrArCrGrArUrCrHrHrHrHrHrHrHrCrA
3' adapter	rAppTGGAATTCTCGGGTGCCAAGG-ddC
5.8s rRNA mask	TCGGCAAGCGACGCTCAGACAGGCGTAGCCCCGGGAGGAACCCGGGGCCGCAAGTGCGTTCGAAGTGTCGATGAT-biotin
RT primer/Reverse primer	biotin-CCTTGGCACCCGAGAATTCCrA
Index primer (8 different)	CAA GCA GAA GAC GGC ATA CGA GAT TT**NNNNNT**G TGA CTG GAG TTC CTT GGC ACC CGA GAA TTC CA
Forward primer	AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA

Small RNA library preparation with QsRNA-seq protocol

The experimental design was the same as described in the generation of the Small-seq library. The QsRNA-seq protocol, described in detail by Fishman and Lamm (2019), [22], was followed with some modifications. The same oligos and adapters as in the Small-seq protocol were used (Table 1). Since the length of the adapters used in the experiment differs from those originally used in the QsRNA-protocol, PEG and isopropanol concentration in the SPRI-based separation step was modified according to the recommendations in the protocol [22]. For the reverse transcription and PCR, the procedure described by Hagemann-Jensen et al., 2018 [19] was followed. For the final cleaning of the PCR product, the different libraries were pooled in equimolar concentrations and a double-sided size-selection using SPRI-beads was performed, as described in the QsRNA-protocol. The quality and quantity of the pooled library was inspected using Qubit and Bioanalyzer.

Microscopy images

Images of differentiating cells were taken five days after activation using a ZOE Fluorescent Cell Imager (Biorad).

Gels and electrophoresis

For the crush and soak method, a 10% tris borate ethylenediaminetetraacetic acid polyacrylamide (TBE-PAGE) gel was prepared, and electrophoresis was performed at 100 V for 60 minutes in a 1% TBE buffer. The gel was put on a UV-table to extract the desired fraction (155-250 bp).

To visualize the library composition, a 3% TBE-agarose gel was prepared, and electrophoresis was performed at 80 V for 90 minutes.

qRT-PCR

cDNA was prepared from 500 ng of total RNA from Th0, Th1, Th2, Th17 and Treg cells, using the SuperScript II RT kit, according to the manufacturer's instructions. Briefly, each sample was mixed with 1 µL of oligo (dT) primer (50 µM) and 1 µL of dNTP (10 mM) and was topped up to 12 µL with ddH2O. The reactions were heated at 65°C for 5 minutes and then put on ice. Four µL of 5X First-Strand Buffer and 2 µL of 0.1 M DTT was added and the reactions were heated for 2 minutes at 42°C. One microliter of SuperScript II RT (200 units) was added. The reactions were incubated at 42°C for 50 minutes and then 70°C for 15 minutes. qPCR was performed using Q5® High-Fidelity 2X Master Mix, according to the manufacturer's instructions. In short, the cDNA was diluted 1:4 with ddH2O and 2 µL of the diluted cDNA was mixed with 0.75 µL of 10 µM Forward and Reverse primer and 1.5 µL of Invitrogen SYBR I green (100x). A 7.5 µL volume of Q5® High-Fidelity 2X Master Mix was added and ddH2O was added to a total volume of 15 µL. PCR was carried out by 40 cycles of 98°C for 30 seconds, 60°C for 10 seconds and 72°C for 10 seconds.

Table 2. *Table of primers used in RT-qPCR*. All primers were ordered from OriGene.

Gene	Forward (5' to 3')	Reverse (5' to 3')
TBX21	ATTGCCGTGACTGCCTACCAGA	GGAATTGACAGTTGGGTCCAGG
FOXP3	GGCACAATGTCTCCTCCAGAGA	CAGATGAAGCCTTGGTCAGTGC
RORγ	GAGGAAGTGACTGGCTACCAGA	GCACAATCTGGTCATTCTGGCAG
GATA3	ACCACAACCACACTCTGGAGGA	TCGGTTTCTGGTCTGGATGCCT
IL4	CCGTAACAGACATCTTTGCTGCC	GAGTGTCCTTCTCATGGTGGCT
IL17a	CGGACTGTGATGGTCAACCTGA	GCACTTTGCCTCCCAGATCACA

The difference in gene expression for every gene was calculated through the ΔΔCt method [23], with *GAPDH* used as a housekeeping gene. To calculate the log2 fold difference, all subtypes were compared to the Tho subtype. See https://github.com/willros/master_thesis/blob/main/qPCR.ipynb for the complete code used.

Data accession

Raw sequencing reads from the Small-seq publication [9] were downloaded from the NCBI Sequence Read Archive with the accession number SRP074776.

Data analysis

The raw sequencing reads from the Small-seq publication were analyzed using gentools (https://github.com/willros/gentools). A description and documentation of the analysis is stated below.

To pre-process the raw sequencing reads downloaded from the NCBI Sequence Read Archive , the raw reads were trimmed from the unique molecular identifier (UMI) sequence using UMI-tools version 1.1.1 (https://umi-tools.readthedocs.io/en/latest/QUICK_START.html) [24], with the following parameters: umi_tools_extract: - input: raw, - mode: extract, - extract-method: regex, - bc-pattern: (?P<discard_1>.*)(?P<umi_1>[ACT]{8}CA).

Since the junction between the 5' adapter and the sequence of interest is demarcated by a CA nucleotide pair, and the UMI is defined by a stretch of eight A,C or T, the above setting ensures that only reads with a 5' adapter ligation are passed on as input to the downstream analysis.

The surviving reads were trimmed from adapter sequences using cutadapt version 3.5 (https://github.com/marcelm/cutadapt) [25], with the following parameters: cutadapt: - input: umi_tools_extract, - adapter: TGGAATTCTCGGGTGCCAAGG, - minimum-length: 18, - maximum-length: 41, - error-rate: 0.1, - overlap: 1, - trimmed_only?: Y

Mapping of reads shorter than 18 nt cannot be done with a high level of confidence, which is why the minimum length was set to 18 [26]. Moreover, the parameters filter only reads confirmed adapter sequences, ensuring that the reads actually come from a potential small RNA.

Next, the reads were aligned to the human reference genome with bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml) [27]. The index was downloaded from https://genome-idx.s3.amazonaws.com/bt/GRCh38_noalt_as.zip. As recommended by Ziemann et al., 2016[20], the following parameters were used: bowtie2: - k: 100, - local: very-sensitive-local, -x : GRCh38_noalt_as. The aligned reads were sorted and indexed using samtools version 1.14 and deduplicated using UMI-tools with the parameters: umi_tools_dedup: - mode: dedup, - method: unique. The deduplicated reads were counted using featureCounts from the subread package (version 2.0.1) [28] (http://subread.sourceforge.net/), using the gene transfer format file https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_39/gencode.v39.annotation.gtf.gz as follows: featureCounts: - M:, - O:, - a: gencode.v39.annotation.gtf.gz.

Normalization and differential expression analysis were performed using DESeq2 version 3.14 [29] from the R bioconductor package (https://bioconductor.org/packages/release/bioc/html/DESeq2.html), which utilizes the negative binomial distribution for modeling of the data [21]. The count matrix output from featureCounts was used as input.

Graphs, Transcripts per million and Clustering

The outputs from gentools were analyzed by the interactive web hosted analysis application https://share.streamlit.io/willros/gentools_streamlit/main/app.py. Transcripts per million (TPM) normalization was performed using the Python package bioinfokit version 2.0.8 and scikit-learn version 1.0.2 was used to perform PCA. As described by Wu et al. [30], genes with a transcripts per million (TPM) value > 0 were labeled as detected.

Data availability

Gentools can be downloaded from https://github.com/willros/gentools. The interactive web application can be found at https://share.streamlit.io/willros/gentools_streamlit/main/app.py
Other code and files used in this master thesis can be found at https://github.com/willros/master_thesis.

Results

Differentiation of Activated CD4+ subtypes

To determine the miRNA expression profile in various subtypes of CD4+ cells, peripheral blood mononuclear cells (PBMC) were extracted from the blood of four healthy adult male donors. CD4+ cells were isolated and activated *in vitro* using a cocktail of different antibodies and cytokines (Figure 1). The whole experimental workflow is depicted in Figure 1.

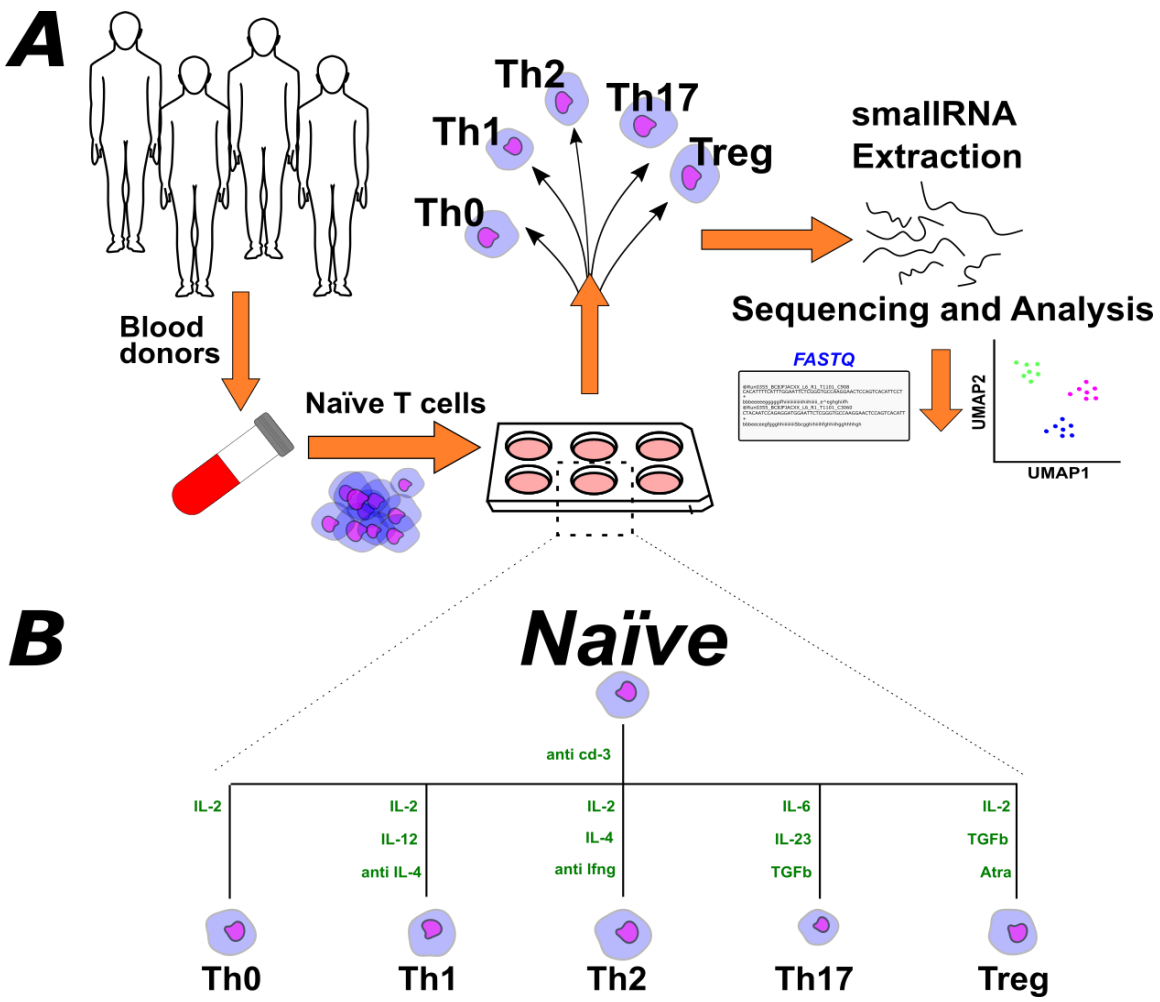


Figure 1. *Schematic representation of the overall workflow and CD4+ cell activation*. **(A)** Naïve CD4+ cells were extracted from blood donated from four healthy individuals. The cells were activated into different subtypes using cytokines and antibodies. After five days, miRNA was extracted from each subtype and sequencing libraries were prepared, sequenced, and analyzed. **(B)** Cytokines and antibodies used in the activation of the Th subtypes.

To maintain a pool of non-differentiated T cells *in vivo,* naïve cells are prevented from becoming activated by stimulation of different ILs and non-activated cells are kept from proliferation [31]. This phenomenon was exploited to confirm whether the activation by cytokines and antibodies had the desired effect on the naïve CD4+ cells. Naïve CD4+ cells were treated with antibodies and cytokines (Figure 1B) or left entirely untreated. After five days the putative subtypes were counted and assessed by microscope (Figure 2).

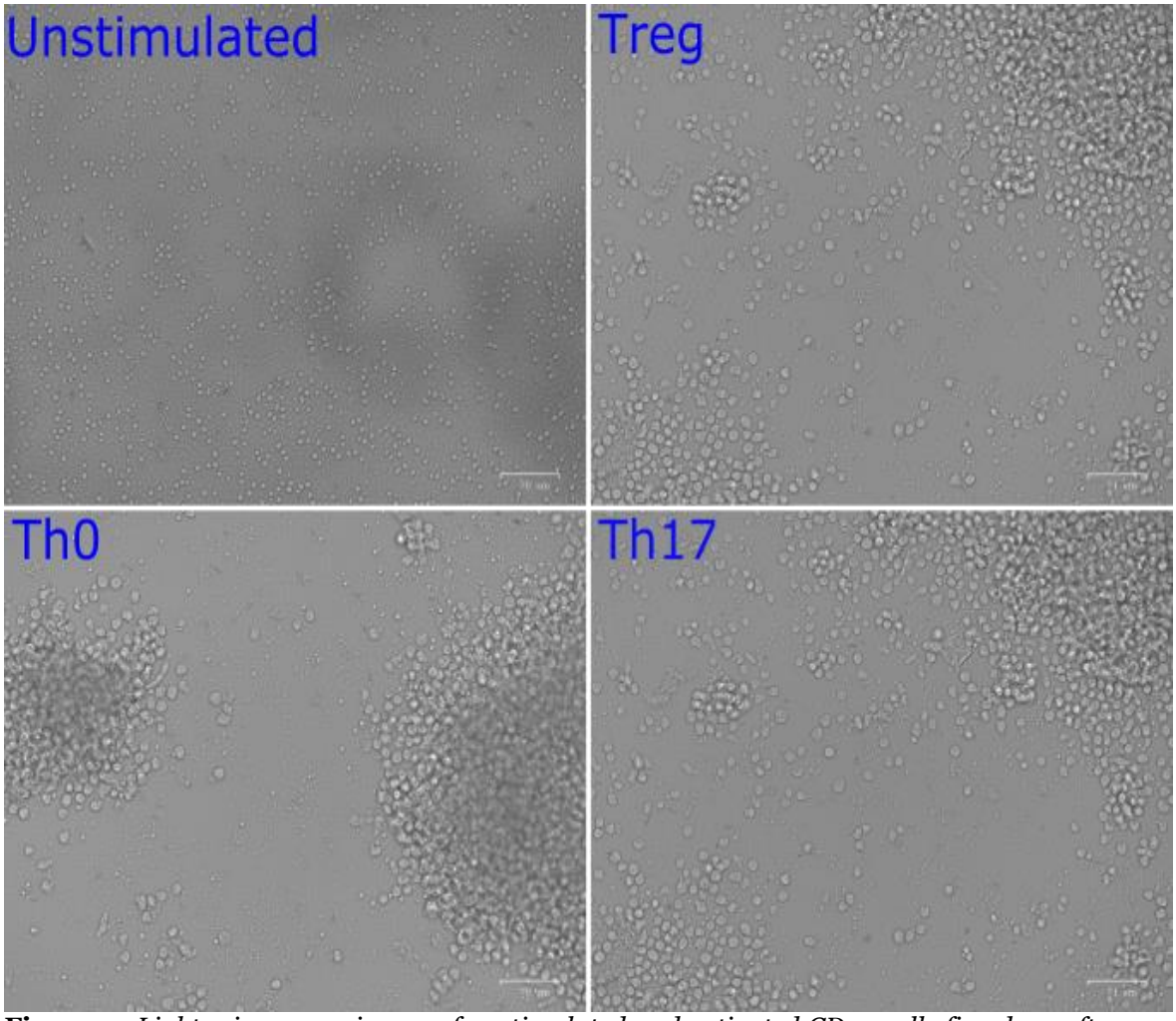


Figure 2. *Light microscopy image of unstimulated and activated CD4+ cells five days after seeding***.** The cells were left untreated (Unstimulated) or treated with chemicals according to Figure 1B.

Activated Th cells were more numerous compared to the unstimulated cells, when counted with a Biorad TC20 Automated Cell Counter. Additionally, the activated cells differed in shape and clustered together more, compared to the unstimulated cells (Figure 2). The proliferation seen in the cells treated with ILs and antibodies suggest that the activation was successful and had the desired effect.

To further examine the effect of activation, qRT-PCR was performed. RNA from Th0, Th1, Th2, Th17 and Treg was extracted and the genes *TBX21, RORγ, IL-17, IL-4, GATA3* and *FOXP3* were used to categorize the different subtypes (Figure 3). Although this method of categorization of Th cell is not as reliable as previously suggested [3–5], the qRT-PCR result still provides reasonably useful information about the states of the cells. The difference in expression between the various subtypes was calculated through the ΔΔCt method [23], with *GAPDH* used as a housekeeping gene. The levels of gene expression in the different subtypes were compared to the levels of gene expression in the Th0 subtype.

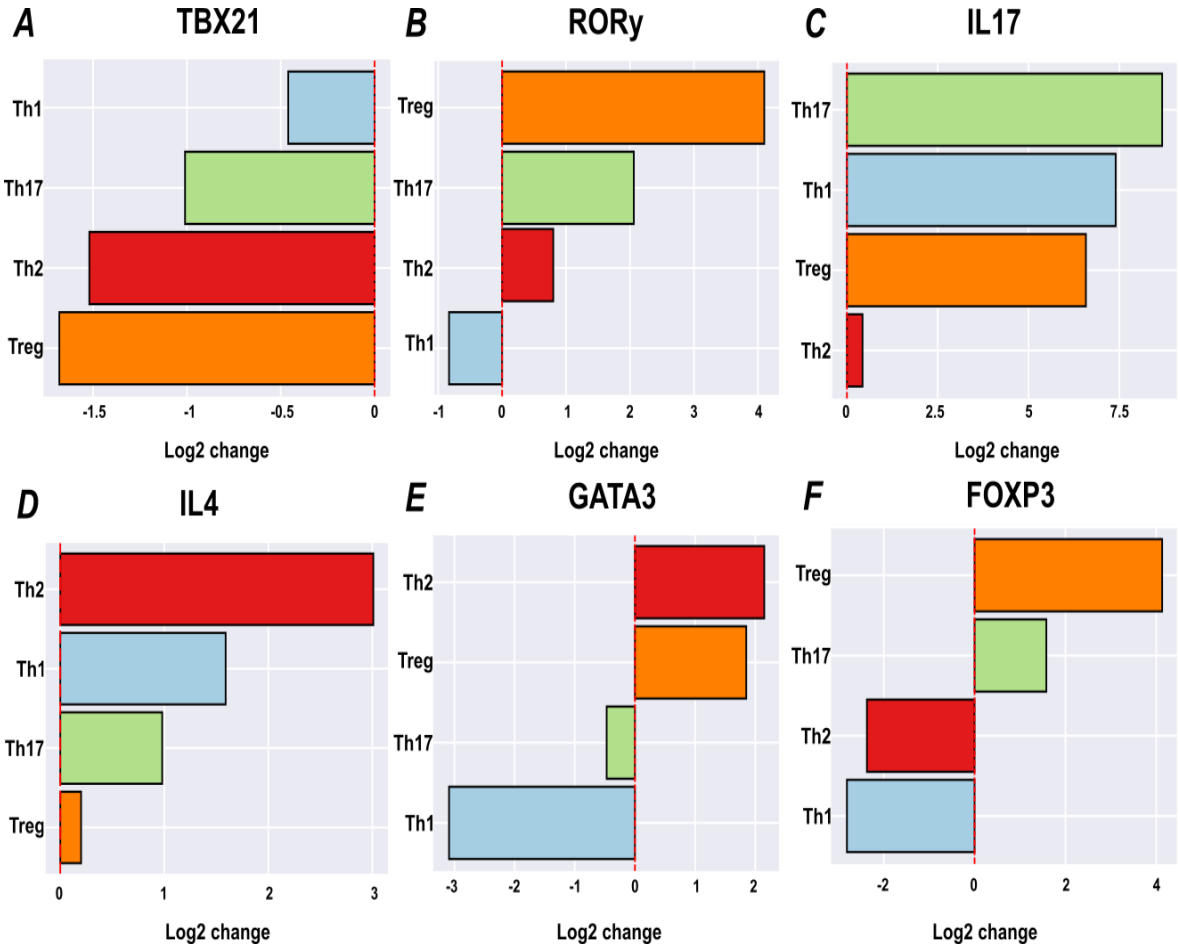


Figure 3. *Results from the qRT-PCR to detect the genes TBX21, RORγ, IL-17, IL-4, GATA3 and FOXP3 for categorizing the subtypes Th0, Th1, Th2, Th17 and Treg of T helper cells. Each Th cell subtype was compared against Th0 with the ΔΔCt method* [23]. *GAPDH was used as the housekeeping gene to generate the log2 fold change of expression.* **(A)** *TBX21* was expressed at a lower level in all subtypes compared to its expression in the Th0 cells. Of Th1, Th17, Th2 and Treg, Th1 cells had the highest level of TBX*21* expression. **(B)** *RORγ* expression was highest in the Treg cells. **(C)** *IL17* expression was highest in the Th17 cells. **(D)** *IL-4* expression was higher in all subtypes, compared to its expression in Th0 cells. Th2 cells had the highest expression of *IL-4*. **(E)** *GATA3* was expressed at a higher level in the Th2 and Treg cells and at a lower level in the Th17 and Th1 cells compared to its expression in the Th0 cells. **(F)** *FOXP3* expression was higher in the Treg and Th17 cells and lower in the Th2 and Th1 cells, compared to its expression in the Th0 cells.

Expression of the marker genes in each subtype corresponds well with their expression profile proposed in the literature (Figure 3). Th2 cells have a high expression of *IL-4* and *GATA3* [5,7] and Th17 cells express *IL-17* and *RORγ*, but not *GATA3* [3,5,8]. Treg cells had the highest expression of *FOXP3,* which corresponds well with findings in previous studies [3,5]. According to the literature, Th1 cells are expected to have a high expression of *Tbx21* [3,5,7]. However, this was not seen in this experiment. Taken together, the microscopy images, counting of cells and the qRT-PCR results indicate that the cellular differentiation was successful.

Refinement of Library Generating Protocol

Before extracting small RNA from the T helper cells as input for the sequencing libraries, the viability of the activated Th cells was measured. The mean viability was approximately 77% (Table S2), as estimated by the Biorad TC20 Automated Cell Counter. The cells were lysed, and small RNAs were isolated from all differentiated subtypes. To determine the fragment sizes of the extracted RNA, the samples were run

on a Bioanalyzer RNA 6000 pico assay, which confirmed that the miRNA was in the desired range of about 50-200 base pairs (bp) (Figure 4A).

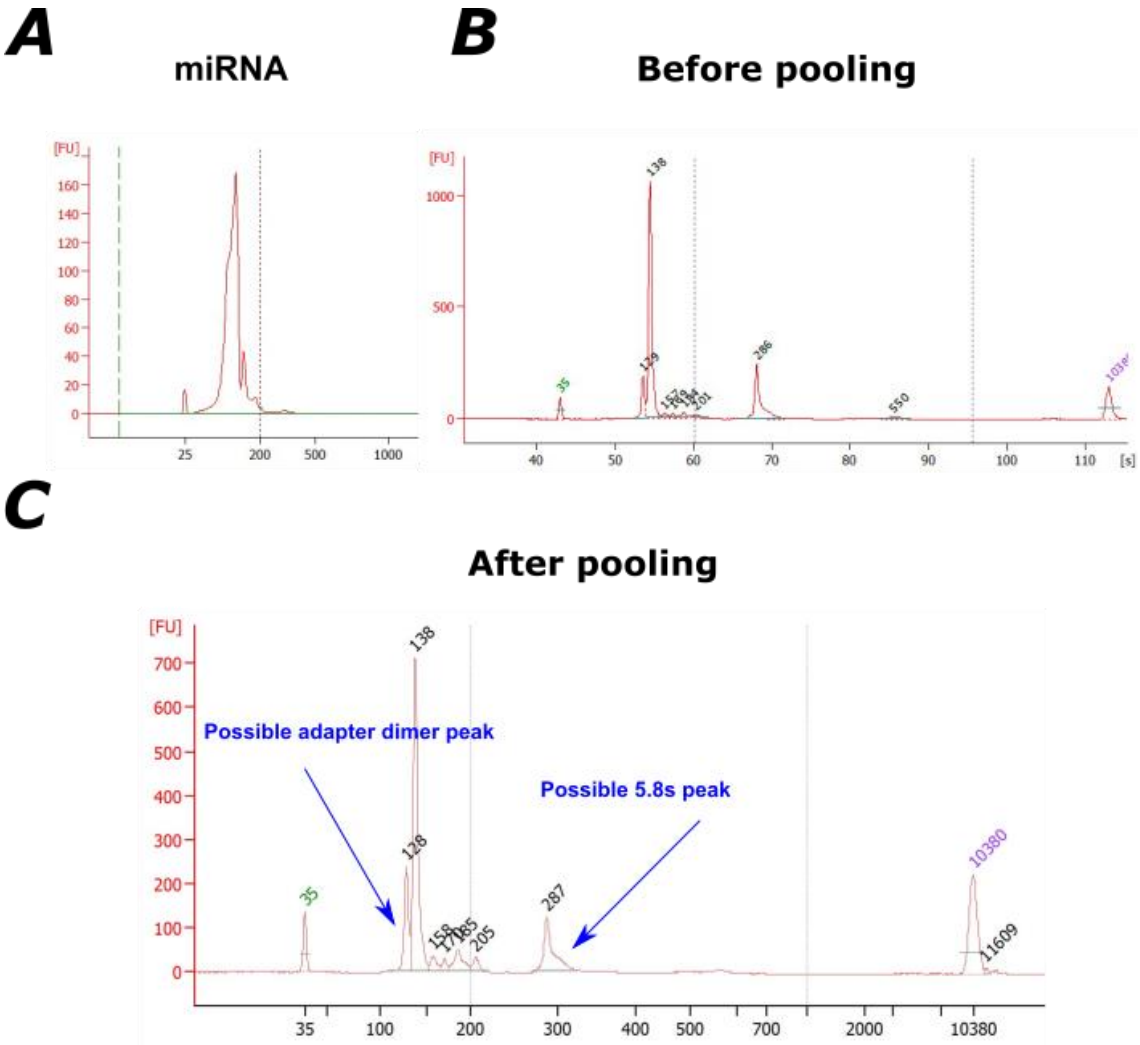


Figure 4. *The miRNA fragment distribution and Bioanalyzer results from libraries prepared with the Small-seq protocol. The Y axis represents fluorescence units (FU) and the X axis represents the number of bp.* **(A)** Representative profile of miRNA fragment distribution, using E.Z.N.A. Micro RNA Kit for extraction. **(B)** Representative profile of the prepared sequencing library before pooling. **C)** Libraries from all activated Th cells pooled in equimolar concentrations. Blue arrows indicate plausible explanations for some of the peaks.

After miRNA isolation, libraries for sequencing were prepared using the Small-seq protocol [9,19]. Since the Small-seq protocol was originally developed as a single cell protocol, adjustments in input miRNA concentration were made. The Small-seq protocol is based on the TruSeq Small RNA Library Prep Kit from Illumina, and the included reference guide recommends a minimum of 10–50 ng miRNA. Despite being optimized for single cell use, Zheleznyakova et al. (2021) successfully utilized Small-seq on bulk miRNA from blood plasma and cerebrospinal fluids [12]. To evaluate the best input concentration, a dilution series ranging from 1.5 ng to 50 ng of total miRNA was made. Interestingly, the final concentrations of the different libraries were not that variable. The library prepared from 1.5 ng miRNA had a concentration of 2.9 ng/µL, whereas the library prepared from 50 ng had a concentration of 3.6 ng/µL, as assessed by Qubit.

Prior to size selection with the crush and soak method, the DNA libraries prepared from each subtype were analyzed using a Bioanalyzer High Sensitivity DNA assay (Figure 4B). Libraries from the various subtypes of Th cells had a sharp peak around 138 bp and a

peak around 280 bp (Table S3). After pooling the libraries in equimolar concentrations, the profile of the fragment distribution was similar to before pooling (Figure 4C). This result corresponds well with the Bioanalyzer profile shown in Hagemann-Jensen et al., 2018 [19]. Table 3 displays all the potential products that theoretically could be generated from the Small-seq protocol. Amplification products of adapter dimers are in theory 130 bp long and it is likely that this pool constitutes most of the peak around 130 bp, highlighted by a blue arrow in Figure 4C. The other major peak at 280 bp is potentially composed of 5.8S rRNA, which is also seen in the libraries generated by Hagemann-Jensen et al., 2018 [19].

Table 3. *Expected products utilising oligos and primers from the Small-seq protocol.* A similar table can be seen in the article from Fishman and T Lamm 2019 [22], but with different oligos, and therefore different products and adapters, compared to the table below.

Pre ligation	**Size (nt)**
Small RNA	19-27
3' adapter	21
5' adapter	36
Intermediate products	
RNA + 3' adapter	40-48
5' adapter + RNA + 3' adapter	76-84
5' adapter + 3' adapter (adapter dimer)	57
Amplification products	
Desired library	149-157
Adapter dimer	130
5.8S rRNA product	282

The results presented above show that omitting the size selection electrophoresis step leads to libraries with high amounts of contaminating fragments, especially adapter dimers (Figure 4). The desired library, around 149-157 bp (Table 3 and Figure 4), are present only in low amounts, indicating that size selection needs to be done.

To remove the unwanted adapter dimer and 5.8S rRNA product, the crush and soak method was performed. The Small-seq library was loaded onto a 10% TBE-PAGE gel and the presumed library RNA molecules were separated (Figure 5A). However, the desired library RNA molecules were too scarce and no material could be cut out from the gel piece.

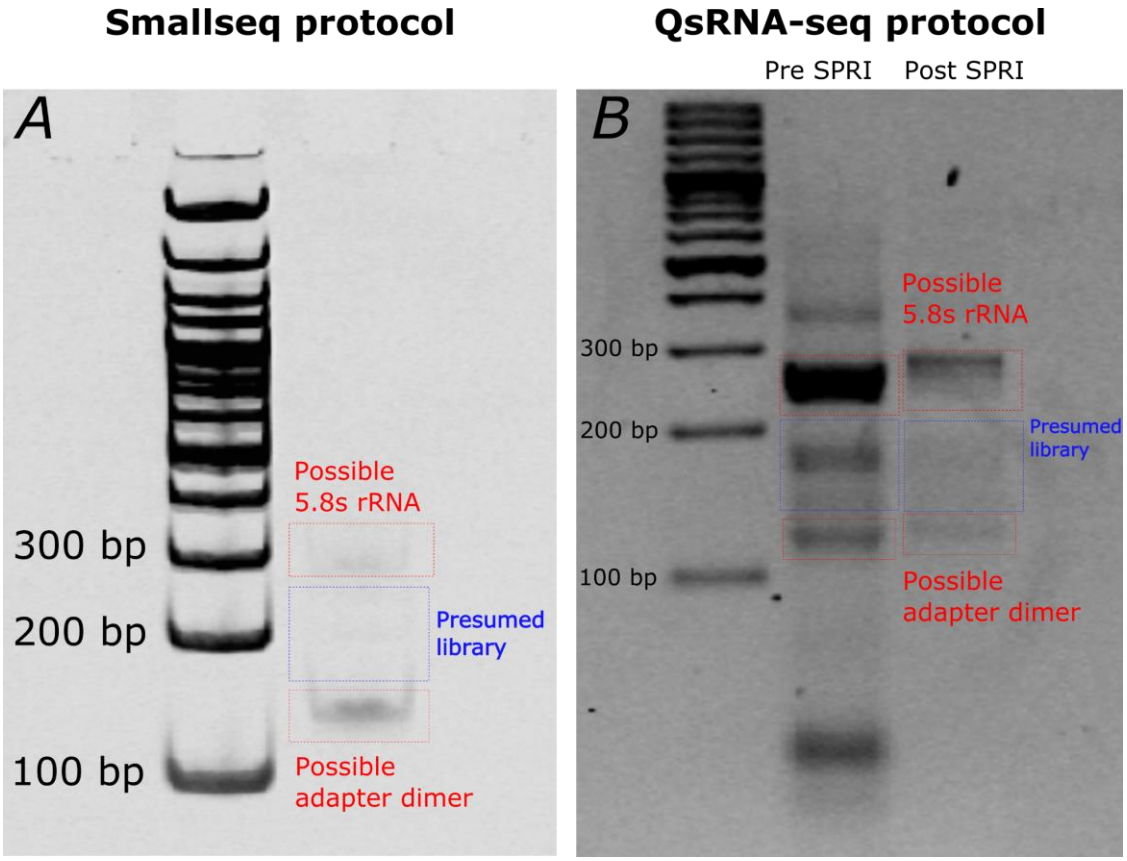


Figure 5. *Fragment sizes of sequencing libraries prepared with Small-seq and QsRNA-seq protocol.* **(A)** 10% TBE-PAGE gel of the same Small-seq library in Figure 4C. Two bright bands (highlighted with dashed red rectangles) correspond with the presumed adapter dimer and 5.8S rRNA peak seen in Figure 4C. The dashed blue rectangle shows the theoretical location of the library of interest (around 145-255 bp) and where the gel should be cut. The library was prepared with 50 ng miRNA as input and the gel was run at 100 V for 60 minutes. **(B)** Pre SPRI: 3% TBE-agarose gel on library prepared with the QsRNA-seq protocol before size selection with SPRI beads. A clear band is visible around 280 bp (highlighted in the red rectangle), which possibly could be the 5.8S rRNA. The presumed library (around 145-255 bp) is highlighted in the blue rectangle. Post SPRI: 3% TBE-agarose gel of the same sample to the left after a two-sided SPRI size selection. The putative 5.8S rRNA (red rectangle) is less visible compared to the corresponding band before size selection, making the concentration of the presumed library relatively higher.

To improve the composition and output of the sequencing libraries, and to eliminate the need of gel electrophoresis for size selection, modifications were made to the Small-seq protocol. Still using adapters and oligos developed for the Small-seq (Table 1), steps to remove free adapters and adapter dimers with SPRI beads and isopropanol crowding, as described in the QsRNA-seq protocol [16], were applied. Libraries were prepared using the same experimental design as described above, with 50 ng of miRNA as input from the different subtypes of cells. By using SPRI-beads and isopropanol, the QsRNA-seq protocol is able to separate DNA fragments with a resolution of 20 nt [16,22]. To remove fragments over 200 nt and under 100 nt in length, caused by amplification of unwanted products in the PCR step, a final cleaning consisting of a two-sided SPRI size selection was performed. Figure 5B depicts the pooled library pre and post the final size selection. Compared to the library produced from the Small-seq, the QsRNA-seq protocol resulted in less adapter dimers, but with more putative 5.8S rRNA products (Figure 5). After size selection, the concentration of the 5.8S rRNA and adapter dimer products were reduced, seen by less visible bands around 300 and 130 bp, respectively (Figure 5B). The concentration of the library decreased from 34 ng/µL pre size selection to 24 ng/µL post size selection, as measured by Qubit. This indicates that little material was lost.

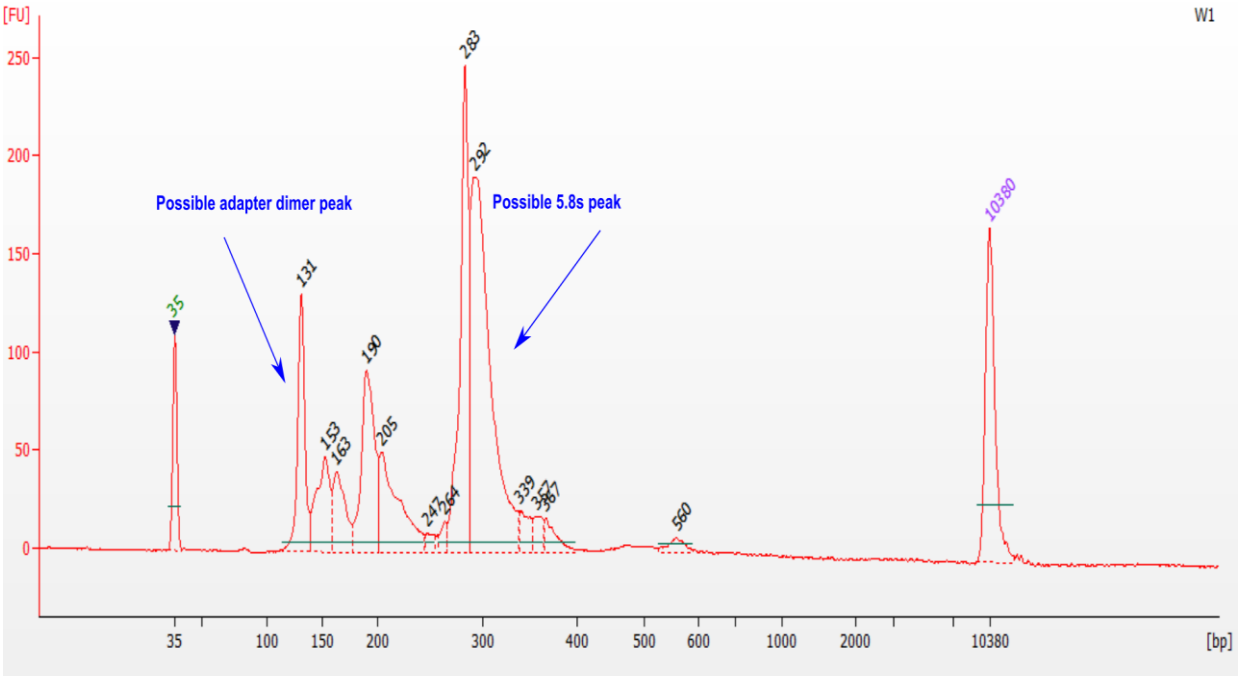


Figure 6. *Bioanalyzer results from the library prepared with the QsRNA-seq protocol after SPRI size selection. The Y axis represents fluorescence units (FU) and the X axis represents the number of bp*. The possible adapter dimer peak around 130 bp was still present, but at a lower concentration compared to the Small-seq protocol. The putative 5.8S rRNA peak was still present at a relatively high concentration. The peaks correspond well with the bands visible in Figure 5B.

The QsRNA-seq library was further analyzed on the Bioanalyzer to inspect the library at a higher resolution (Figure 6). The corresponding profile had a much higher ratio of desired library molecules to contaminating fragments in comparison to the Small-seq library (Figure 5 and Table S4). Despite being more flexible, user friendly and reproducible, the QsRNA-seq protocol resulted in a cleaner library with less contaminants in comparison to the Small-seq protocol.

Due to time constraints, only the Small-seq generated library was subjected to sequencing. Unfortunately, because of the large fraction of adapter dimers, the flow cell of the Illumina machine was over-clustered. Therefore, no miRNA sequencing data were obtained from the different CD4+ subtypes.

Development of Gentools to Analyze Data From small RNA Sequencing Experiments

Analysis of sequencing data is complex and often depends on multiple command line tools (CLI) and software [21,32]. For that reason, scripting in languages such as bash or Python is often implemented to facilitate the automatization of running multiple CLI tools sequentially. To increase the reproducibility between studies, the software and parameters used must be well documented, easily accessible and comprehensible. However, scripts can be difficult to maintain, read and customize, and are therefore prone to errors and fall short regarding reproducibility [32].

Attempts were made to reproduce the results presented in Hagemann-Jensen et al., 2018 [19], using the same data and script-based pipeline suggested by the authors. However, the efforts were unsuccessful, and the same results could not be reproduced. To create a more reliable data analysis pipeline optimized for miRNA sequencing, gentools was developed, which can be installed through Python's package management system pip (https://github.com/willros/gentools) and used in the command line. By default, gentools uses bowtie2 [27] with parameters optimized for miRNA alignment, recommended by Ziemann et al., 2016 [20]. Other tools, such as cutadapt [25] and UMI-

tools [24], inspired from the pipeline described by Faridani et al., 2016 [9,19] (https://github.com/eyay/smallseq), were also used. The default configuration of gentools was customized for the library preparation protocols used in this thesis. Differential expression analysis is performed by DESeq2 [29]. Inspired by a bioRxiv preprint [33], gentools utilizes a customizable configuration file that determines which software and parameters should be run in the pipeline. This makes the analysis reproducible and easy to use by users with limited experience. In addition, the configuration file also serves as a log file for the specific run, allowing for effortless documentation [33]. The output files generated by gentools can be used in an interactive web application (https://share.streamlit.io/willros/gentools_streamlit/main/app.py), where further analysis of the experiment can be made.

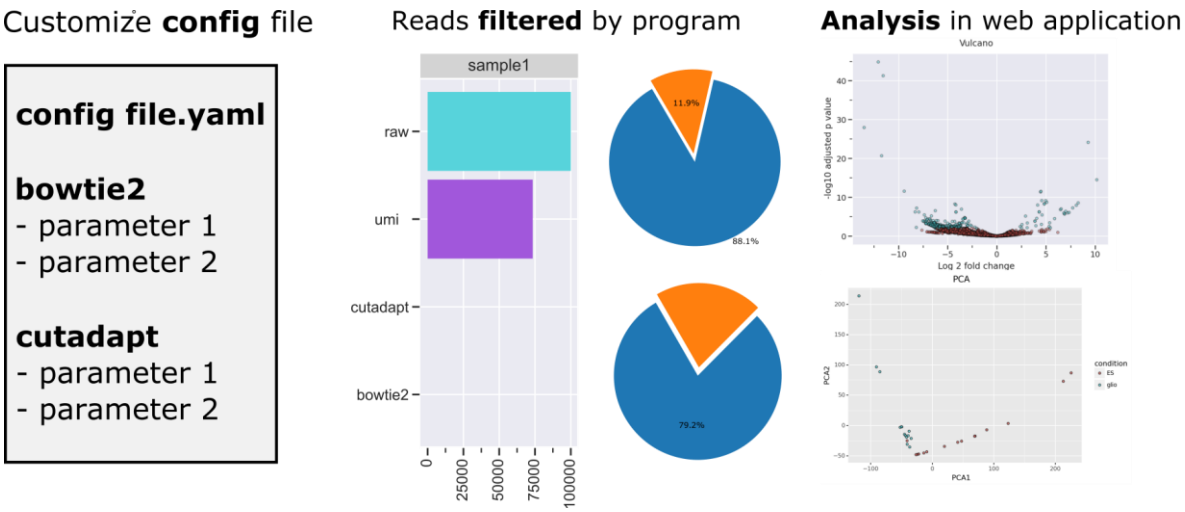


Figure 7. *Schematic figure of workflow and output generated by gentools*. A customizable configuration file determines which software and parameters to use. The pipeline is run and information about the pre-processing step is generated. Gentools outputs csv files, which can be passed to the interactive web application for further analyses.

The interactive web application of gentools takes input files generated by gentools and outputs analysis in the form of plots of principal component analysis (PCA) and differentially expressed genes, together with files containing information about transcript per million (TPM) (Figure 7 & 8). The data analyzed in Figure 8 are 15 fastq files from embryonic stem cells (SRR3495737-SRR3495751) and 15 fastq files from glioblastoma cells (SRR3495787-SRR3495801), deployed by Faridani et al., 2016 [9]. Fractions of gene types based on the TPM values are visualized.

Interestingly, the result generated by gentools and the interactive associated web application differ remarkably to the results generated from the pipeline published in the Small-seq publication, which are presented in the supplementary materials to the article [9]. Across all samples, the Small-seq publication authors report only a small part of the total reads as derived from rRNA, whereas in my analysis, the glioblastoma cells seem to contain a large fraction of reads stemming from rRNA (Figure 8). The fact that I was not able to reproduce the analysis presented by Faridani et al., 2016 [9], despite following their well-documented pipeline, underlines the difficulties of reproducibility in small RNA sequencing analysis. The development of gentools is an attempt to fill a gap that currently makes it impossible to compare results between different small RNA sequencing studies.

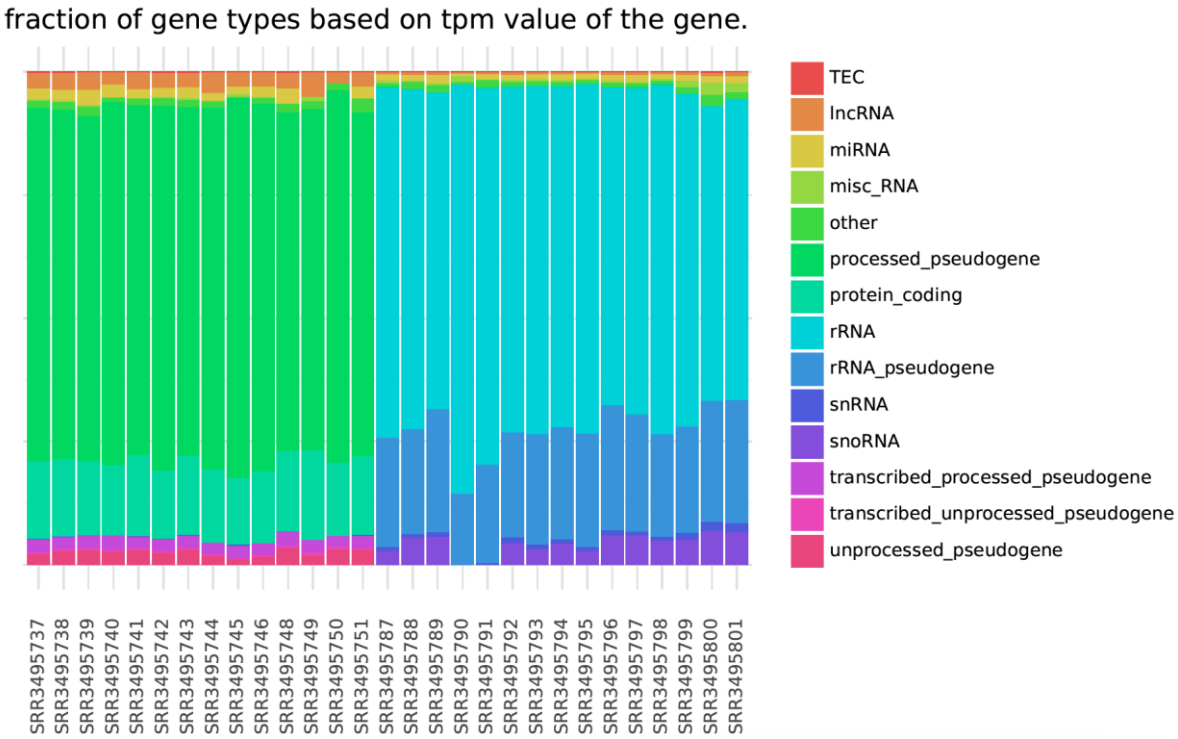


Figure 8. *Plot of fractions of gene types based on TPM value generated by the gentools interactive web application.* Number of gene categories can be specified by the user. Redundant gene types are lumped together in the category "other". Genes with a TPM value above 0.1 were categories as detected, as described by Wu et al., 2018 [30]. The figure was generated using gentools interactive application.

Discussion

To achieve deeper knowledge about Th cell activation and differentiation, novel and more sensitive methods for categorization of Th cell subtypes are needed. Finding ways to categorize and to regulate the behavior of CD4+ cells via miRNA, could aid improving immunotherapies and personal medicine [6,34]. In this master thesis, I have explored the possibilities for improving both small RNA sequencing protocol as well as bioinformatic pipelines, to gain more knowledge about small RNA in CD4+ cells. Although a lot of studies already have profiled the expression of small RNA in Th cells, most of them rely on RT-qPCR or microarray data [6,35], which neglect the possibility of finding novel or low expressed variants [36]. Other studies use mice as a model organism [11], which due to the large differences between Th cells in mice and human cannot provide reliable or useful extrapolations toward deepened knowledge about the human immune system [37].

Starting with high quality RNA as input material is important when generating libraries for RNA-seq. Since degraded products cannot be excluded from the library synthesis itself, and since these products can falsely be assigned as novel miRNA, the input RNA must be of high quality [16]. In this thesis, input miRNA used for creating sequencing libraries came from bulk extraction of small RNA, using a kit based on columns without isopropanol usage. Some studies propose the use of total RNA as input instead of isolated fractions of small RNA, due to the presumed loss of small RNA species associated with length separation [15]. A previous study claims that RNA extraction kits, which do not rely on isopropanol for precipitation, results in a non-biased extraction of small RNA [22]. This finding was confirmed in the present study by inspection with the Bioanalyzer (Figure 4A). Questions about using frozen rather than freshly isolated CD4+ cells were addressed by Satpathy et al, 2019 [38] where they concluded that no difference in the RNA isolated could be observed when comparing frozen and alive cells, at least regarding ATAC peaks. Precautions considering the above mentioned problems

were taken into account during the experimental work of this thesis, to ensure that the input RNA were of high quality.

To improve the accessibility and reproducibility of miRNA sequencing data analysis, gentools was developed. The default configurations of gentools are optimized for small RNA analysis and for the adapters and oligos implemented by the Small-seq protocol (Table 2) [9,19]. The parameters for UMI-tools used by gentools can be put in contrast to the more lenient Small-seq data analysis protocol, where the UMI are counted as the first eight nucleotides from the 5' end of each read, regardless of evidence indicating otherwise. This unprejudiced assumption runs the risk of producing false positives, counting reads which stem from other sources than small RNA, *e.g.,* artefacts from library preparation and sequencing errors. Gentools accounts for this by utilizing the CA nucleotide linker which separates the UMI from the putative miRNA. Through a regular expression search, everything to the left of the match is discarded whereas the right part is counted as true miRNA. This ensures that each read is a product from the library preparation and not just an arbitrary output from the sequencing platform. The default settings for cutadapt address the same concern and use the ligated 3' adapters as a verification that the read stems from miRNA and not from another source, by eliminating reads without signs of 3' adapters.

Although unaccompanied mapping to known miRNA databases, for example miRBase [39,40], while omitting mapping to the entire genome can be fast, it ignores the discovery of novel small RNA species and neglects the identification of transcripts. This makes alignment to the entire genome a preferred choice [20]. Since the origin of true miRNA can be ambiguous [39] and the transcription start sites of miRNAs is difficult to map compared to regular RNA transcription [41], allowance for multiple hits per read during alignment is important for the discovery of novel miRNAs. In an article from 2016, Ziemann et al. [20] tested the accuracy of many popular aligners with different parameters on simulated data. The 3'-end non-template extensions are common variations in small RNAs [20] and these extensions must be considered when mapping reads to the reference genome. Using bowtie2 with --local --very-sensitive-local parameter enables local alignment and allows mismatches at the 3' end of the read. Omitting this could lead to overlooking miRNA containing non templated extensions [20]. This, in addition to how well bowtie2 performed in the tests [20], makes bowtie2 a suitable choice for aligning small RNA reads.

Reanalysis of sampled data from the Small-seq article [9] resulted in a low frequency of mapped reads. Similar results can be seen in an article from 2021, where Hücker et al. [18] compared multiple miRNA sequencing methods. Across different protocols, they report that less than 10% of the total reads mapped to the human genome and merely 2% mapped to annotated miRNA [18], which underlines the problems associated with small RNA mapping [9,18,20,21]. Like many other miRNA sequencing articles, Hücker et al. also reported that several of the tested methods exhibit a high ratio of adapter dimers [18], like the results presented in this thesis (Figure 4-6). In a study from 2013, 't Hoen et al. [26] examined the reproducibility of miRNA protocol between different laboratories, and reported that a high level of variance can be observed between them. 't Hoen et al. also reported a high grade of variance between different samples regarding reads mapping to miRNA regions with 19% of mapped reads being the median [26]. Like the reanalyzed samples in Figure 8, where a lot of the reads seem to originate from rRNA, 't Hoen et al. also reported samples with a high level of rRNA content. These differences can potentially be explained by differences stemming from RNA extraction, which again underlines the importance of qualitative input RNA [26].

A recurring and major problem encountered in this master thesis was the presence of adapter dimers and rRNA fragments (Figure 4-6 and Table S3-S4). Therefore, as previously mentioned, the focus of future research should lie on resolving these issues [18]. Many interesting and novel technologies have been proposed as possible candidates to overcome these obstacles. The CRISPR/Cas9 system has successfully been used to eliminate occurrences of unwanted fragments from small RNA sequencing libraries [42].

Here, a targeting sgRNA, together with the Cas9 protein, cleaves unwanted products, *e.g.* adapter dimers, which leads to a higher concentration of the desired library and excludes the need for downstream size selection [42]. Other protocols make use of biotinylated beads, which hybridize to the sequences to be depleted [43]. This method is effective and accurate and allows flexibility regarding which sequence to target [43]. Ligation strategies based on circularization have also been shown to increase the quality and specificity of the library [44]. By utilizing intermolecular ligation of the small RNA target and a ligated 3' adapter, the need for a separate 5' adapter is eliminated, which circumvents the formation of adapter dimers [44]. In summary, all the above-mentioned techniques are putative solutions to problems regarding high abundance of unwanted sequences, either naturally occurring or formed as by-products from the protocols.

By attempts of refining the Small-seq [9,19] and the QsRNA-seq protocol [16,22], and the development of a computational pipeline built on establishment by previous research [20,21,45], this master thesis addresses some of the major problems innate to small RNA sequencing methods. The results may be a step in the right direction towards an improved methodology regarding Th cell categorization, which further could help enhance arising cancer treatments such as immunotherapies involving CAR T cells. Moreover, despite the CD4+ cell focus, the presented tools are not limited to this domain, but can be extended toward other systems and biological questions.

Societal Impact Statement

Cancer causes a lot of suffering for the individual and costs a lot of money for society. In recent years, novel cancer therapeutics have been developed that enhance the innate capacity of the patient's own immune cells, called T cells. These enhanced T cells are reintroduced into the bloodstream of the patient where they patrol the body in the search of cancer cells, to destroy them with their newly acquired abilities. Certain types of T cells have been used to improve cancer immunotherapy methods, due to their capacity to kill cancer cells. While these T cells are crucial for the riddance of cancer, T helper cells are also important for the immune response. The T helper cells support the killer cells during their battle by offering assistance. To understand how the helper cells contribute to the war against cancer, we must gain more knowledge about their complex life cycle and their states of differentiation.

Evidence suggests that a certain group of small molecules, called micro-RNA, are important for T helper cell development and for their role in the immune system. microRNAs can also provide us with information about the current state of the cell and can therefore be used as a tool for categorization.

To gain information about the current microRNA profile of each cell, a method called RNA sequencing is often used. For sequencing to work, the micro–RNA is extracted from the cells and from that a sequencing library is prepared. This workflow is unfortunately associated with some common problems. The first problem regards the purity of the library. The library often contains a large part of an unwanted fraction of molecules which stems from the chemicals used in the protocol. This fraction can complicate the analysis and the sequencing itself, making the results unusable. The second problem concerns the computational analysis of the data. The small size of the micro-RNA makes the analysis hard to interpret and reproduce across different times and laboratories. This master thesis addressed both problems and aimed to refine the method for future use. To achieve this, various protocols were tested and tweaked to optimize further their use. In addition, a computational tool was developed. This tool, called gentools, is easy to use and aims to streamline and standardize the microRNA analysis, to create reproducible results which can be compared across time and studies.

The findings made in this master thesis will provide future researchers with better tools to study the microRNA profile of T helper cells, which in turn could help to improve and develop immunotherapies.

Bioethical Statement

The acquisition of patient samples and blood was in accordance with institutional and national bioethical guidelines and regulation, approved by DNR 2016/53-31: In vitro-studier av humana B-celler I blod, tonsil och adenoid.

Acknowledgements

Thank you, Sebastian, for showing me around the lab and for showing me how to perform the protocols. Thank you, Martin, for answering everyday questions about the wet lab! Thanks, Hayoung, for reading and commenting on the thesis. Thanks, Kristina, for helping with the cell culture. Thanks, Debo, for your nice company and teaching me about AI and UMAPS! And thank you Johan Henriksson for supervising this thesis and taking me in as a master student and for letting me try different things and helping me develop my coding abilities. I had a great time!

References

1. Agarwal, S. *et al.* In Vivo Generation of CAR T Cells Selectively in Human CD4+ Lymphocytes. *Mol. Ther.* **28**, 1783–1794 (2020).
2. Tay, R. E., Richardson, E. K. & Toh, H. C. Revisiting the role of CD4+ T cells in cancer immunotherapy-new insights into old paradigms. *Cancer Gene Ther.* **28**, 5–17 (2021).
3. Zhu, J., Yamane, H. & Paul, W. E. Differentiation of effector CD4 T cell populations (*). *Annu. Rev. Immunol.* **28**, 445–489 (2010).
4. 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).
5. Martinez-Sanchez, M. E., Huerta, L., Alvarez-Buylla, E. R. & Villarreal Luján, C. Role of Cytokine Combinations on CD4+ T Cell Differentiation, Partial Polarization, and Plasticity: Continuous Network Modeling Approach. *Front. Physiol.* **9**, 877 (2018).
6. Diener, C. *et al.* Quantitative and time-resolved miRNA pattern of early human T cell activation. *Nucleic Acids Res.* **48**, 10164–10183 (2020).
7. Henriksson, J. *et al.* Genome-wide CRISPR Screens in T Helper Cells Reveal Pervasive Crosstalk between Activation and Differentiation. *Cell* **176**, 882–896.e18 (2019).
8. Liu, R. *et al.* MicroRNA-15b Suppresses Th17 Differentiation and Is Associated with Pathogenesis of Multiple Sclerosis by Targeting O-GlcNAc Transferase. *J. Immunol.* **198**, 2626–2639 (2017).
9. Faridani, O. R. *et al.* Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.* **34**, 1264–1266 (2016).
10. Rodríguez-Galán, A., Fernández-Messina, L. & Sánchez-Madrid, F. Control of Immunoregulatory Molecules by miRNAs in T Cell Activation. *Front. Immunol.* **9**, 2148 (2018).
11. Kirigin, F. F. *et al.* Dynamic microRNA gene transcription and processing during

T cell development. *J. Immunol.* **188**, 3257–3267 (2012).

12. Zheleznyakova, G. Y. *et al.* Small noncoding RNA profiling across cellular and biofluid compartments and their implications for multiple sclerosis immunopathology. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
13. Dard-Dascot, C. *et al.* Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genomics* **19**, 118 (2018).
14. Juzenas, S. *et al.* A comprehensive, cell specific microRNA catalogue of human peripheral blood. *Nucleic Acids Res.* **45**, 9290–9301 (2017).
15. Benesova, S., Kubista, M. & Valihrach, L. Small RNA-Sequencing: Approaches and Considerations for miRNA Analysis. *Diagnostics (Basel)* **11**, (2021).
16. Fishman, A., Light, D. & Lamm, A. T. QsRNA-seq: a method for high-throughput profiling and quantifying small RNAs. *Genome Biol.* **19**, 113 (2018).
17. Shore, S. *et al.* Small RNA Library Preparation Method for Next-Generation Sequencing Using Chemical Modifications to Prevent Adapter Dimer Formation. *PLoS One* **11**, e0167009 (2016).
18. Hücker, S. M. *et al.* Single-cell microRNA sequencing method comparison and application to cell lines and circulating lung tumor cells. *Nat. Commun.* **12**, 4316 (2021).
19. Hagemann-Jensen, M., Abdullayev, I., Sandberg, R. & Faridani, O. R. Small-seq for single-cell small-RNA sequencing. *Nat. Protoc.* **13**, 2407–2424 (2018).
20. Ziemann, M., Kaspi, A. & El-Osta, A. Evaluation of microRNA alignment techniques. *RNA* **22**, 1120–1138 (2016).
21. Tam, S., Tsao, M.-S. & McPherson, J. D. Optimization of miRNA-seq data preprocessing. *Brief. Bioinform.* **16**, 950–963 (2015).
22. Fishman, A. & T Lamm, A. QsRNA-seq: A protocol for generating libraries for high-throughput sequencing of small RNAs. *Bio Protoc* **9**, e3179 (2019).
23. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402–408 (2001).
24. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in

Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).

25. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
26. 't Hoen, P. A. C. *et al.* Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31**, 1015–1022 (2013).
27. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
28. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
29. 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).
30. Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M. & Wilke, C. O. Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics* **19**, 510 (2018).
31. Tsunetsugu-Yokota, Y. *et al.* Homeostatically Maintained Resting Naive CD4+ T Cells Resist Latent HIV Reactivation. *Front. Microbiol.* **7**, 1944 (2016).
32. Singh, U., Li, J., Seetharam, A. & Wurtele, E. S. pyrpipe: a Python package for RNA-Seq workflows. *NAR Genom Bioinform* **3**, lqab049 (2021).
33. Farrell, D. smallrnaseq: short non coding RNA-seq analysis with Python. doi:10.1101/110585.
34. Raabe, C. A., Tang, T.-H., Brosius, J. & Rozhdestvensky, T. S. Biases in small RNA deep sequencing data. *Nucleic Acids Res.* **42**, 1414–1426 (2014).
35. Fayyad-Kazan, H. *et al.* MicroRNA profile of circulating CD4-positive regulatory T cells in human adults and impact of differentially expressed microRNAs on expression of two genes essential to their function. *J. Biol. Chem.* **287**, 9910–9922 (2012).
36. Heinicke, F. *et al.* Systematic assessment of commercially available low-input miRNA library preparation kits. *RNA Biol.* **17**, 75–86 (2020).

37. Allen, T. M. *et al.* Humanized immune system mouse models: progress, challenges and opportunities. *Nat. Immunol.* **20**, 770–774 (2019).
38. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
39. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).
40. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–73 (2014).
41. Cha, M. *et al.* A two-stream convolutional neural network for microRNA transcription start site feature integration and identification. *Sci. Rep.* **11**, 5625 (2021).
42. Hardigan, A. A. *et al.* CRISPR/Cas9-targeted removal of unwanted sequences from small-RNA sequencing libraries. *Nucleic Acids Res.* **47**, e84 (2019).
43. Kraus, A. J., Brink, B. G. & Siegel, T. N. Efficient and specific oligo-based depletion of rRNA. *Sci. Rep.* **9**, 12281 (2019).
44. Barberán-Soler, S. *et al.* Decreasing miRNA sequencing bias using a single adapter and circularization approach. *Genome Biol.* **19**, 105 (2018).
45. Potla, P., Ali, S. A. & Kapoor, M. A bioinformatics approach to microRNA-sequencing analysis. *Osteoarthritis and Cartilage Open* **3**, 100131 (2021).

Appendix

Table S1. *Antibodies and chemicals used in the T helper cell differentiation and activation.* The table displays the volume (microliter) of each substance that was used in the different Th cell activation. For example, the medium prepared for activating Th0 cells consisted of 35 uL *IL-2*, 15 uL anti-cd28 and 4950 uL of immunocult medium.

	Th0	**Th1**	**Th2**	**Th17**	**Treg**
total media	5000.0	5000.0	5000.0	5000.0	5000.0
IL-2	35.00	50.00	50.00		50.00
IL-6				5.0	
IL-4			5.0		
IL-12		5.0			
TGFb				50.0	50.0
IL23				25.0	
Atra					50.0
anti-IL4		50.0			
anti-Ifng			50.0		
anti-cd28	15.0	15.0	15.0	15.0	15.0
immunocult	4950.0	4895.0	4895.0	4920.0	4850.0

Table S2. *Viability of cells used to extract the miRNA*. The cell count and viability of the activade Th cell used to extract the miRNA for the downstream experiments. The viability was estimated through an automated cell counter.

Cell type	**Cell count (cell/mL)**	**Viability**
T0	5.3 million cells/mL	90%
T1	4.9 million cells/mL	66%
T2	6.3 million cells/mL	80%
T17	4.6 million cells/mL	70%
Treg	4.9 million cells/mL	80%

Table S3. *Distribution and concentration of one representative library before pooling. Created with the original Small-seq protocol.* The majority of the library consists of products around 138 bp in length, which likely corresponds to adapter dimers.

Size (bp)	**Concentration (pg/µL)**
129	229
138	1260
157	45
169	26
184	49
201	39
286	328
550	20

Table S4. *Distribution and concentration of a representative library created with the modified QsRNA-seq protocol, before cleaning with SPRI-beads.* The majority of the library consists of sequences of around 290 in length, which likely corresponds to 5.8S rRNA.

Size (bp)	**Concentration (pg/µL)**
130	378
187	357
280	853
289	1 362