Additional file: 1
A time-resolved meta-analysis of consensus gene expression profiles during human T-cell activation

List of Figures

List of Tables

1   Data sources

1.1   Transcriptome datasets used for the development of the consensus gene signature

Table S1 and Figure S1 provide an overview of the datasets used for the discovery and verification of temporal consensus gene expression signatures of T cells.

1.2   *Pan T-cell Verification Set*

1.2.1   Preparation of peripheral mononuclear blood cells (PBMCs) and pan T-cell isolation

Buffy coats were obtained from the Springe Blutspendedienst and were kept at 4 °C until PBMC isolation. PBMCs were isolated by density centrifugation. The cell count and viability was determined with a hemocytometer and trypan blue staining. CD14+ monocytes, CD19+ B-cells and CD56+ natural killer cells were depleted using CD14, CD19 and CD56 magnetic cell sorting (MACS) beads (Miltenyi) following the manufactures instruction. The flow through was kept for CD2+ T-cell isolation. Therefore $1x10^7$ cells were resuspended and CD2 microbeads were added for 15 min at 4 °C After centrifugation (10 min, 300 xg, 4 °C) and resuspension of the cells in MACS-buffer. LS columns (Miltenyi) were placed into a magnet and cells were loaded onto the column through a pre-seperation filter (Miltenyi). The cell count was determined by trypan blue staining using a hemocytometer. Quality of the CD2+ isolation was confirmed by fluorescence-activated cell sorting (FACS).Isolated pan T-cells (0.5-1×10^6) were incubated with 2.5 µL human TruStain FcXTM-block for 10 min at 4 °C. Viable cells were stained with Zombie NIRTM for 15 min at RT. Afterwards, cells were stained with CD3-FITC, CD45-AF700, CD-BV510, CD4-PacificBlue, Cd45RO-PE/Cy7, CD45RA-APC, CD14PerpCy/Cy5 (BioLegend, San Diego, USA) or respective isotype controls for 30 min on ice. Zombie NIRTM was used as a viability marker due to manufactures instructions. Samples were washed with 2 mL FACS buffer (PBS supplemented with 0.25% bovine serum albumin and 2 mM ethylenediaminetetraaceticacid (EDTA)). After centrifugation at 350 xg for 10 min, cells were resuspended. Viable leukocytes were discriminated from cell debris by forward (FSC) and sideward scatter (SSC) properties as well as CD45+, Zombie NIR+ cells Cells were separated by size in FS (Forward Scatter) and by granularity in SS (Sidewards Scatter). Subsequently, duplicates of singular cells were eliminated via the disproportion of height (PEAK) and width (TOF). The singular leukocytes negative for the dead dye ZombieNIR were further differentiated into monocytes (CD14) and CD3 T-cells. NK cells and B cells were identified on CD3- cells using CD56 and CD19. CD3+ cells were further differentiated into T helper CD4+ and cytotoxic CD8+ cells. Naive T-cells were discriminated as CD45RA+ cells and memory T-cells as CD45RO+ cells.

For quality control, 200.000 cells were analyzed using a Beckman Coulter Navios 3 L 10 C. Flow cytometry was analyzed using Kaluza 2.1. software. Analysis were done using seven human blood donors. Pan T-cell isolation strategy resulted in a CD3+ T-cell purity of 97,7 ± 1,8%. Fractions of the T-cell populations are shown in Figure 5B (main part) Stimulation of the T-cells with either CD3/CD28 beads to activate T-cells via CD3 and CD28 receptors, Phorbol Myristate Acetate (PMA) to activate protein kinase C together with ionomycin, a calcium ionophore or Phytohämagglutinin (PHA) a mitogen resulted in proliferation and cytokine (IL-2) secretion of the cells. Viability was proven in ATP assay. Controls stimulation assays were performed with LPS and medium (see Figure S2).

1.2.2   RNA-seq sample preparation and sequencing

For RNA sequencing, up to $0.5x10^6$ treated or untreated T-cells per sample were collected. The cell pellets were re-suspended in Qiazol and stored at –80°C. RNA was extracted according to the miRNeasy mini protocol (Qiagen). After two steps of DNase-digestion (TURBO DNA free

Tab. S1: Overview of datasets included for discovery and verification. Green colored time points were used for the meta-analysis.

Discovery set

Study	Organisation	PMID	GEO	Layout	Reads (bp)	Platform	Strandness	Extraction	Sample source	All|Bio|Tech Repl.	0h naive	0.5h	1h	2h	4h	6h	8h	12h	24h	2d	3d	4d	5d	6d	induced to following phenotye:	
Tuomela et al.	Aalto University	26967054	GSE52260	SE	1x50	HiSeq 2000	N	total RNA	Cord blood	51|3|1	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y				Th0	Th17
Ullah et al.	Aalto University	29466736	GSE90569	PE	2x101	HiSeq 2000	N	polyA RNA	Cord blood	51|3|1	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y				Th0	iTreg
Schmidt-ind	Aalto University	29730990	GSE96538	PE	2x101	HiSeq 2500	Y	polyA RNA	PBMC	51|3|2*	Y	Y	Y	Y		Y		Y	Y	Y	Y	Y	Y		Th0	iTreg
Schmidt et al.	Karolinska Institut	29730990	GSE94396	PE	2x75	HiSeq 2500	N	total RNA	PBMC	27|3|1	Y		Y	Y		Y		Y	Y	Y	Y			Y	Th0	iTreg
Elo et al.	University of Turku	20620947	GSE17974			Affymetrix HG-U133 Plus 2.0			Cord blood	49|3|1	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y				Th0	Th2
Äijö et al.	University of Turku	23110343	GSE32959			Affymetrix HG-U133 Plus 2.0			Cord blood	27|3**|1	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y					Th1

SE = single-end; PE = paired-end; Th0 = activated CD4+ T-cells without differentiation cytokines; Bio Repl. = Biological replicates; Tech Repl. = Technical replicates; * For 2, 6 and 12h only one techn. replicate is available; ** For 72h only two biological replicate

GSE52260: Only human data were analyzed. Five samples were sequenced with 2x75 bp paired-end reads. As in Tuomela et al. those reads were truncated to 50 bp and only one of the paired-end reads was used.

GSE94396 : G01 (unstimulated), G02 (Th0) and G04 (iTreg) datasets were used. To be consistent with the treatment protocols for iTreg differentiation across datasets, we used only samples induced with TGF-β + ATRA reagents.

GSE17974: unstimulated and induces Th2 samples were analyzed

GSE32959: unstimulated and induces Th1 samples were analyzed. Note: The same pooled cells as for the Th2 (IL-4 stimulated) culture conditions by Elo et al. (GSE17974) were used.

Verification set

Study	Organisation	PMID	GEO	Layout	Reads (bp)	Platform	Strandness	Extraction	Sample source	All|Bio|Tech Repl.	0h	0.5h	1h	2h	4h	6h	8h	12h	24h	2d	3d	4d	5d	6d	induced to following phenotye:
Rade et al.	Fraunhofer ITEM/IZI		GSE197067	PE	2x101	HiSeq 2000	Y	total RNA	PBMC	44|4|1	Y					Y		Y	Y	Y	Y				Th0 & negative controls
Gutierrez-Arcelus et al.	Brigham and Women's Hospital, Harvard Medical School	32066938	GSE140244	PE	2x101	HiSeq 2000	Y	polyA RNA	PBMC	184|24|1	Y			Y	Y		Y	Y	Y	Y	Y				Th0

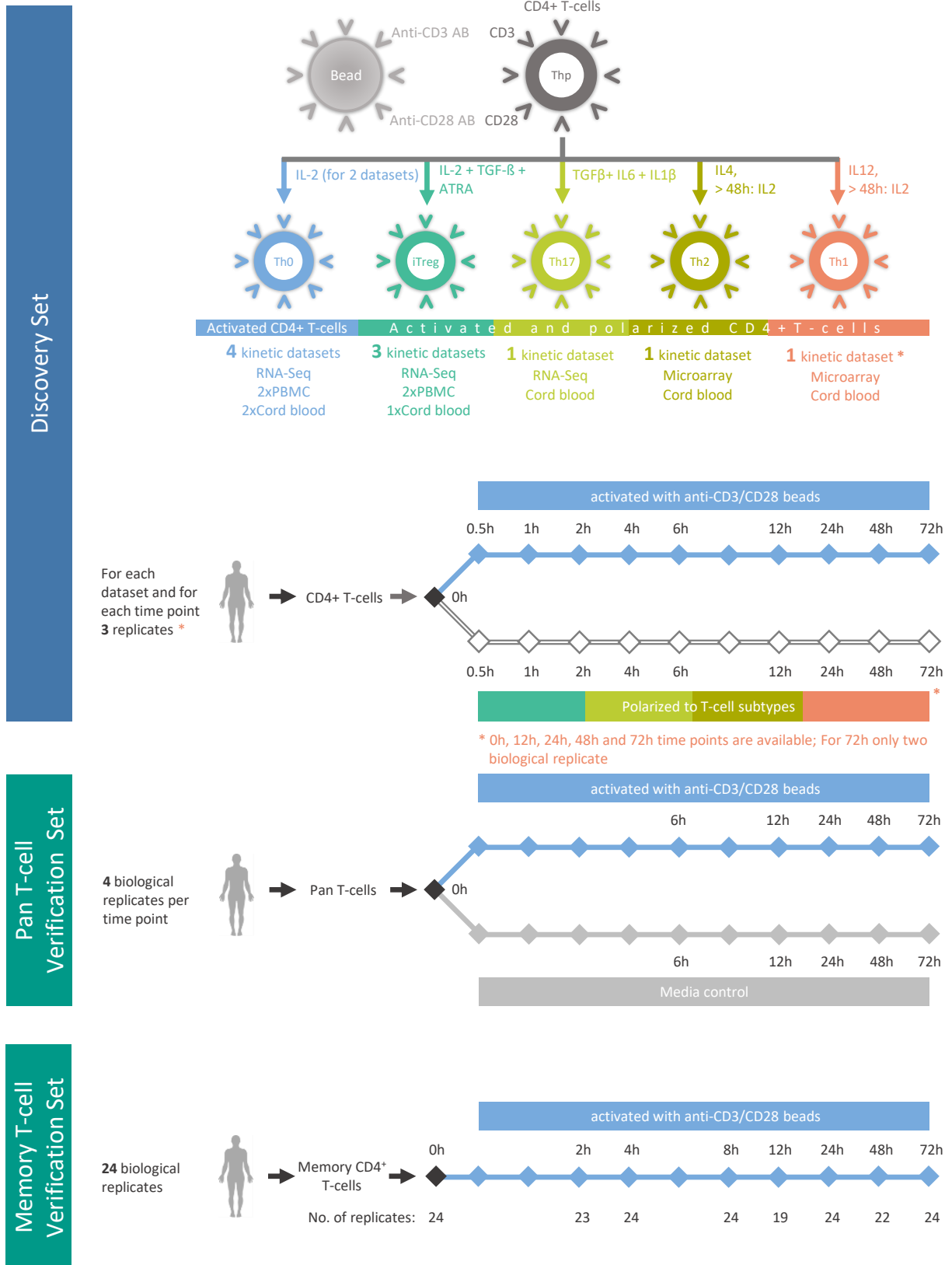GSE140244: Only replicate A was used (2 biological replicates have technical replicates)

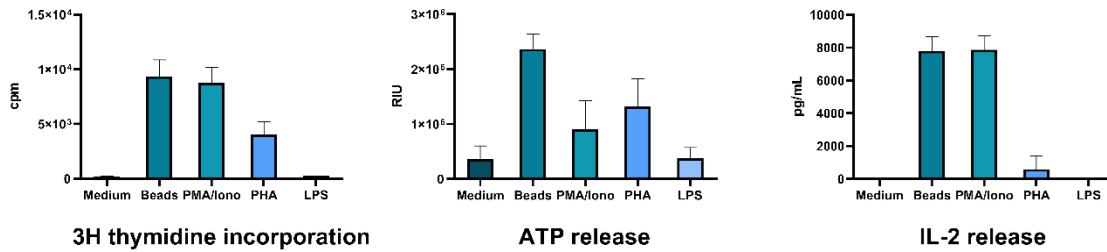Fig. S1: Overview of the datasets included for discovery and verification.

Fig. S2: *Pan T-cell Verification Set*: **Biological activity of isolated T-cells.** Proliferation capacity, viability and cytokine secretion was analyzed after three days stimulation with either CD3/28 receptor activating beads, PMA/Ionomycin or PHA using ATP assay, 3H thymidine proliferation assay and IL-2 ELISA. Controls stimulation assays were performed with LPS and medium. Assays were done using T cell isolations from three human blood donors in technical triplicates. Analysis were performed using GraphPad Prism 9.0.

Kit, Ambion), extracted RNA was quantified using a Qubit RNA-Kit and the DeNovix instrument (Biozym). RNA quality was analyzed on a Bioanalyzer 2100 instrument (Agilent Technologies). For subsequent RNA-sequencing analyses, 150 ng of total RNA per sample were used. Library preparation was conducted using Truseq-Stranded total RNA Sample Prep kit (Illumina, Inc, San Diego, CA) according to the manufacturers' protocol. Molarity of each library was calculated and equal amounts were pooled and used for sequencing (12 pM). Sequencing was performed with $2 \times$ 126-bp paired-end reads using HiSeq SBS Kit v4 chemistry on a HiSeq 2500 instrument (Illumina). Three flowcell à 8 lanes were sequenced with all 44-pooled libraries each.

2   Pre-processing

2.1   Microarray datasets

Each T-cell population dataset was processed separately as follows: normalization and probeset summarization was performed using the robust multi-array average algorithm (RMA) implemented in the `rma()` function of the `affy` v1.68.0 R package [87]. An updated probeset annotation chip definition file (CDF) based on GENCODE v29 and provided by BrainArray [88] (version 23.0.0 for Affymetrix HGU133Plus2 arrays) was used (File available at `http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/23.0.0/gencodeg.asp`). Each sample was evaluated by visual inspection of the array pseudo-images, density plots of probe intensities, Normalized Unscaled Standard Error (NUSE) and Relative Log Expression (RLE) implemented in the `affyPLM` v1.66.0 R package [89] (data not shown). An unspecific gene filtering step was incorporated into the analysis. The normalized intensities for at least 3 samples, which represent the smallest experimental group, must be larger than the 25th percentile across all gene intensities. Furthermore, low-intensity probes were filtered out based on interquartile range (IQR) as described in Chockalingam et al. [90]. The principal component analysis did not reveal any critical outlier (see Figure S5D-E).

2.2   RNA-Seq datasets

Demultiplexing of Illumina raw files from the *Pan T-cell Verification Set* was performed with the Illumina `bcl2fastq` software v2.19 [91]. The following steps were performed for all RNA-Seq datasets from the *Discovery Set* and *Pan T-cell Verification Set*: The single-end or paired-end FASTQ reads were trimmed and filtered using `AdaptorRemoval` v2.3.1 [92] with additional parameters to trim ambiguous bases (N) at 5'/3' termini (–trimns), remove low-quality bases (–trimqualities, –minquality 20) and keep reads with a minimum read length of 30bp (–minlength 30). Trimmed reads were mapped to the human reference genome version GRCh38/hg38 by `HISAT2` v2.1.0 [93]. Gene level quantification for the human reference gene annotation GENCODE (release 29 GRCh38.p12)

was obtained by using `HTSeq` v0.11.2 [94]. Sample QC was reported using `FastQC` v0.11.8 [95] to assess base call accuracy, `Preseq` v2.0.3 [96] to evaluate the library complexity. Duplication metrics were collected using `Picard tools` v2.18.29 (http://broadinstitute.github.io/picard/) function MarkDuplicates using BAM files generated by `HISAT2`. Picard's `CollectRnaSeqMetrics` was used to collect mapping percentages on intergenic, intronic, coding and UTR regions as well as gene body coverage. `RSeQC` v3.0.0 [97], was used to determine, read GC content, junction saturation, read pair inner distance, and strandness of reads. `FastQScreen` v0.14.0 [98] in conjunction with `bowtie2` [99] was conducted to assess RNA library composition. Aggregated data visualization for the secondary analysis and quality control were generated using the `MultiQC` [100] framework. A quality control summary report for the *Pan T-cell Verification Set* is shown in Figure S3. Two samples mapped with at least 25% against the human rRNA reference genome (Figure S3C). However, we did not observe a noteworthy difference of the number of uniquely mapping reads against the human genome reads or the number of reads assign to genes (Figure S3D-F) compared to the other samples. Principal component analysis of all samples did not reveal any critical outliers (Figure S3F).

Gene counts generated by `HTSeq` from different studies and based on the same CD4$^+$ T-cell population were concatenated. For each population gene counts were converted to count-per-million (CPM). Low expressed genes were filtered out from downstream analysis using the `filterByExpr()` function from the `edgeR` v3.32.1 `R/Bioconductor` package [101] with default settings. Genes having a CPM above c in m samples were kept, where m the size of the smallest experimental group and c is determined by 10 counts divided by the median library sizes times one million. Normalization factors to scale the raw library size were calculated using the function calcNormFactors() from `edgeR` with the Trimmed mean of M-Values (TMM) method. [102]. For T-cell populations consisting of multiple studies (Th0 and iTreg), log-transformed CPM expression data was adjusted for the factor *study* using the `removeBatchEffect()` function in the `R/Bioconductor limma` v3.46.0 package [54, 55]. After batch correction no significant correlation was observed between the first 50 principal components and potential confounding factors such as *study* or *sample source* (PBMCs or neonatal cord blood) (Figure S4). For principal component analysis of the T-cell populations see Figure S5. The batch corrected CPM expression data was used as input for non-negative matrix factorization.
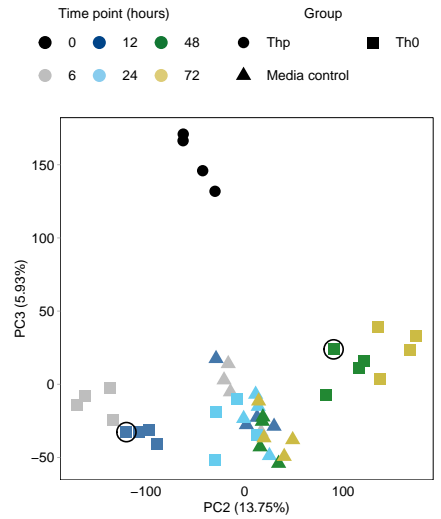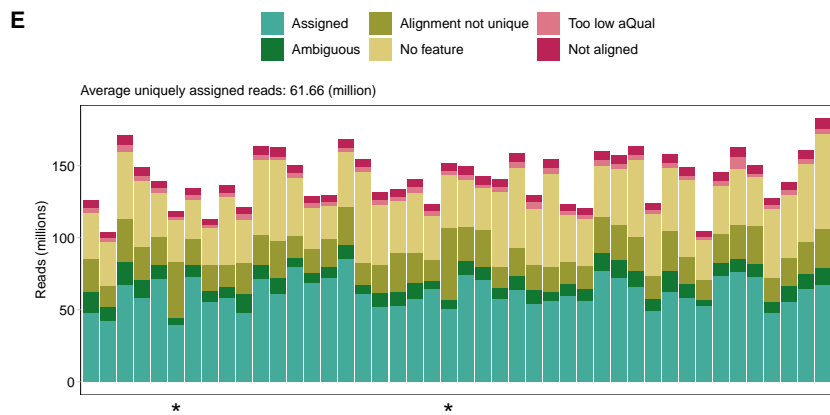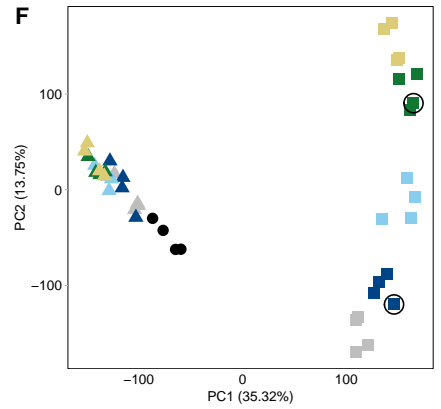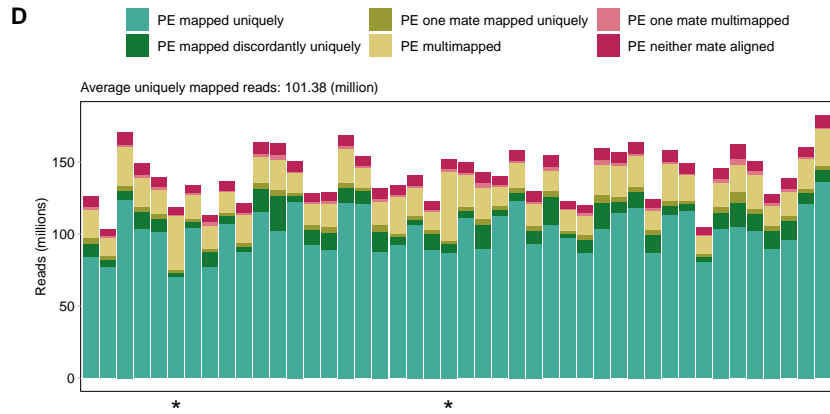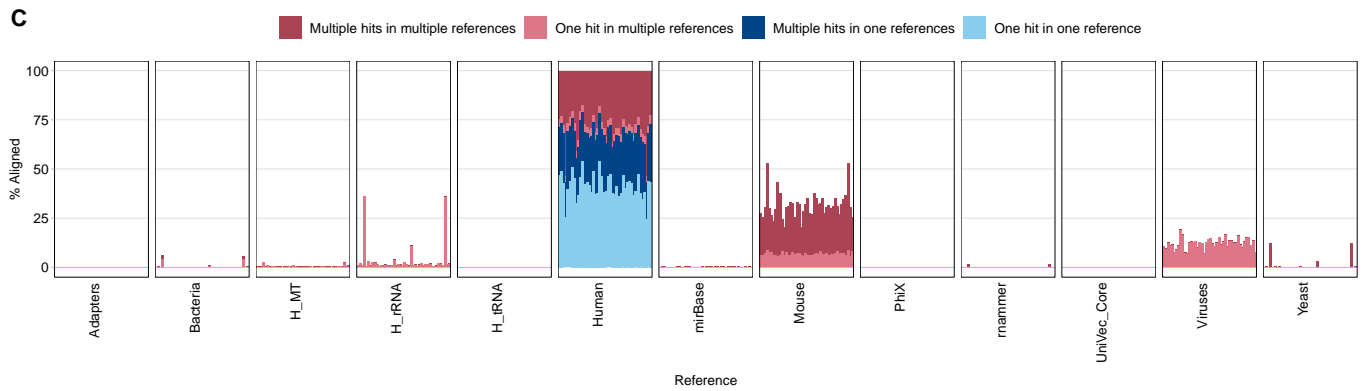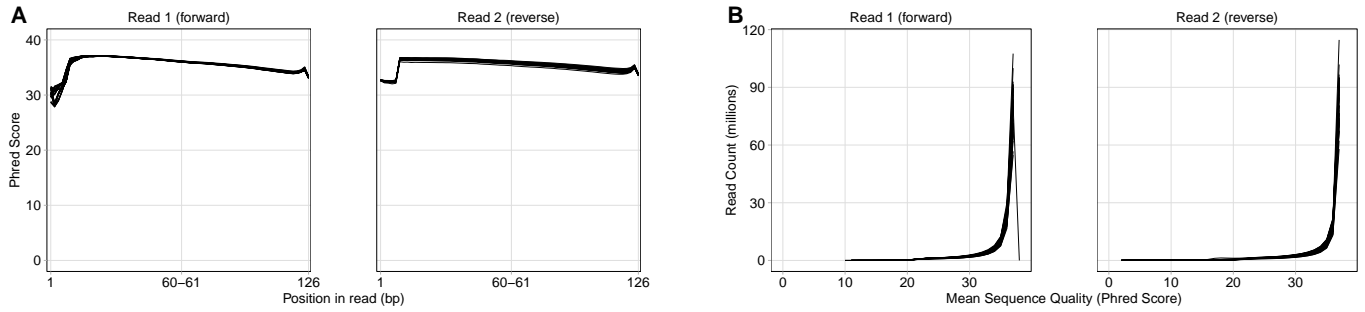
Fig. S3: *Pan T-cell Verification Set*: **Quality Control of samples with respect to sequencing library composition, read alignment and sample variance**. Following adapter trimming, each FASTQ file were assessed for average per base **(A)** and per sequence **(B)** quality as measured by Phred score. **(C)** To assess the sequencing library composition, each sample was subsampled to randomly 1 million trimmed paired-end reads. `FastQ Screen` in conjunction with `bowtie2` was conducted to detect possible contamination like for example bacteria and overrepresented fractions of RNA species like human rRNA. The y-axis depicts the percentage of first reads for each sample that aligned against the references from Table S2. Reads are classified into four distinct types indicating reads uniquely mapping in one sequence reference (one hit in one reference), reads with multiple mappings in one sequence reference (multiple hits in one reference), reads uniquely mapping in distinct sequence references (one hit in multiple references) and reads with multiple mappings in distinct sequence databases (multiple hits in multiple references). **(D)** Quality assessment of read alignment results for each sample. Mapped uniquely: read pairs aligned concordant 1 time. Multi-mapped: read pairs aligned concordant >1 times. Mapped discordantly uniquely: read pairs aligned discordantly 1 time. One mate mapped uniquely: one read of the pair maps to the genome 1 time. One mate multi-mapped: one read of the pair maps to the genome >1 times. **(E)** Results of gene quantification for each sample. Quantification was performed with `HTSeq`. For **(D)** and **(E)**: Samples mapped with at least 25% against the human rRNA reference depicted with asterisks. **(F)** Principal component analysis of variance-stabilized counts based on the 5000 most variable genes. The upper plot depicts first and second principal components, the bottom plot the second and third principal component. Samples are shaped as asterisks, if the subsampled reads mapped with at least 25% against the human rRNA reference.

Tab. S2: References used for FastQ Screen

| Reference	Source
Adapter sequences	https://github.com/csf-ngs/fastqc/blob/master/Contaminants/contaminant_list.txt
Bacteria	ftp://ftp.ncbi.nih.gov/genomes/refseq/bacteria/, Oct 2014
H_MT	human mitochondrial reference sequence from GRCh37/hg38
H_rRNA (human ribosomal RNA sequences)	NR_003286.1 (18S), NT_003287.1 (28S), NR_003285.2 (5.8S), V00589.1 (5S), NC_012920.1: gi|251831106:1671-3229 (MT 16S) and NC_012920.1: gi|251831106:648-1601 (MT 12S)
H_tRNA (human transfer RNA sequences)	http://gtrnadb.ucsc.edu/genomes/eukaryota/Hsapi38/hg38-tRNAs.fa
Human genome	GRCh38/hg38, reference chromosomes only
mirBase	miRNA sequences from mirBase v21
Mouse genome	UCSC/mm10
PhiX	gi|9626q372|ref|NC_001422.1| Enterobacteria phage phiX174 sensu lato, complete genome
RNAmmer (predicted rRNA sequences)	http://www.cbs.dtu.dk/services/RNAmmer/, v1.2
UniVec Core	ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/UniVec_Core build 8.0, May 2015
Viruses	ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral/, March 2014
Yeast	Genome assembly SacCer3
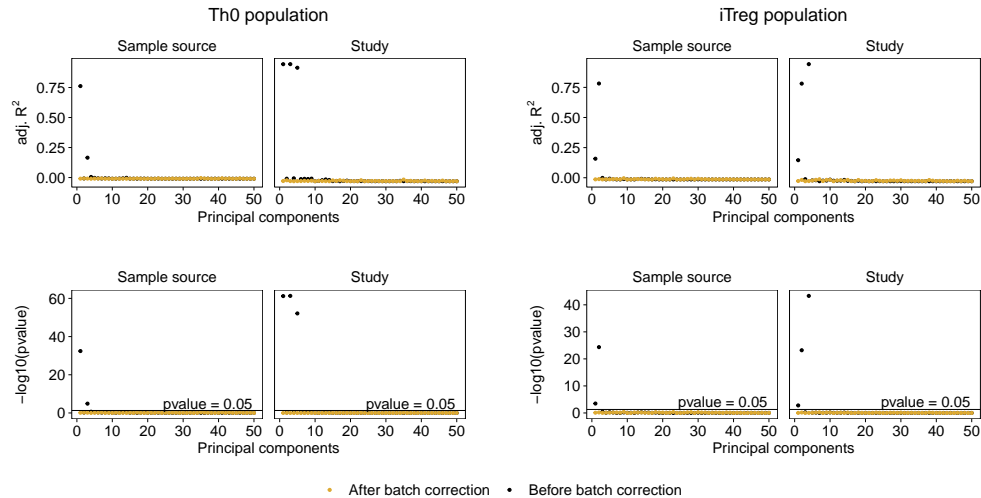
Fig. S4: *Discovery Set*: **Correlation analysis of potential confounding factors.** Analyses was performed for T-cell populations (Th0 and iTreg) consisting of multiple studies. Correlation and F-statistic based on linear regression with the factors *sample source* (PBMCs or cord blood) or *study* as covariate for the first 50 principal components. The 5000 most variable genes were used as input to PCA. Linear regression was performed before and after batch correction by adjusting for the factor *study* using the `removeBatchEffect()` function from the `limma` package. The top panels depicted the adjusted coefficient of determination and the bottom panels the p-values from the F-stastic calculated by the lm() function in `R`. After batch correction no significantly (p-value < 0.05) correlation could be observed between the factors *sample source* or *study* and the first 50 PCs.
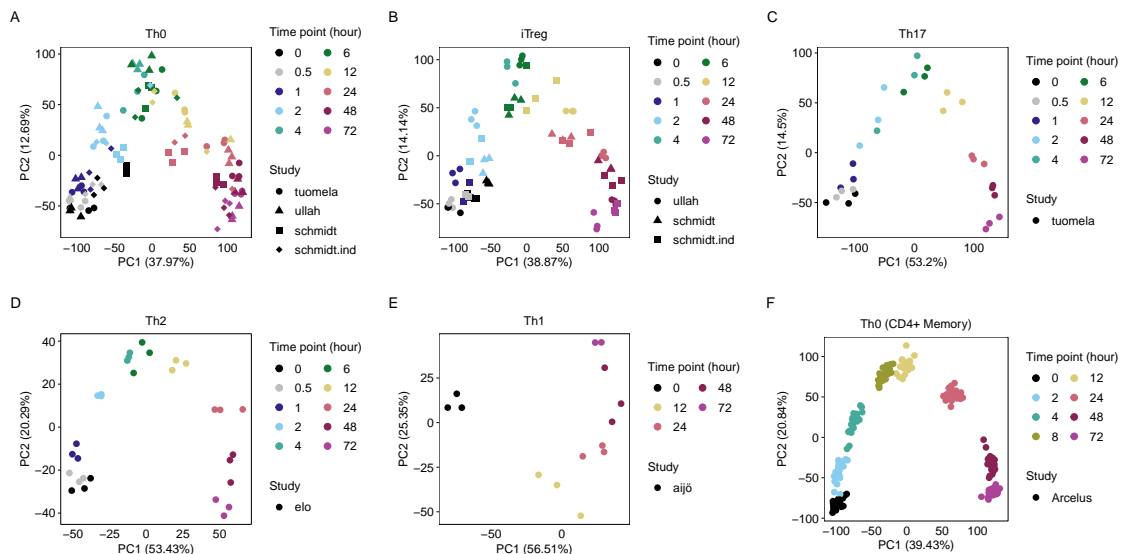
Fig. S5: *Discovery and Verification Sets*: **Principal component analysis (PCA).** PCA of time-course RNA-Seq and microarray datasets was performed separately for each T-cell population from the *Discovery Set* and *Verification Sets*. The 5000 most variable genes were used for PCA. For RNA-Seq data, we used log-transformed counts per million (CPM) values as input to PCA. In case of microarrays, we used the log-transformed normalized intensities using the multi-array average (RMA) normalization method. For T-cell populations (Th0 and iTreg) consisting of multiple studies, potential batch effects across the studies were removed using the `removeBatchEffect()` function from the `limma` package. The plot depicts the first and second principal components with samples colored according time point of activation and shaped by study. Plot **(A-E)** depict samples from the *Discovery Set* and plot **(F)** from the *Memory T-cell Verification Set*. A PCA plot from the *Pan T-cell Verification Set* is shown in Figure S3F. PCA of all samples did not reveal any critical outliers.

3   Differential gene expression analysis

As input to `limma` we used for RNA-Seq data the filtered gene counts (see subsection 2.2) and the normalized factors from the TMM method. In case of Microarrays we used the filtered normalized intensities from the RMA method (see subsection 2.1). For the T-cell populations Th1, Th2, Th17 from the *Discovery Set* and Th0 from both *Verifications Sets* a paired design of the form $\sim 0 + block + contrast$ was used, where *block* encodes the replicates and *contrast* encodes the pre- and post-treatment information for each time point. For the T-cell populations Th0 and iTreg from the *Discovery Set*, *block* encodes the different studies due to unbalanced paired design. In case of RNA-Seq data, the `voom` [103] function of `limma` was used to accommodated the mean-variance relationship using precision weights. This function allows to fit gene expression derived from RNA-Seq by linear regression models. By following this workflow, we ensure that for both types of datasets (microarray and RNA-Seq) identical workflows are applied for (i) fitting the gene-wise statistical models and for (ii) testing the null hypothesis that coefficients do not diverge from null. Linear modelling and empirical Bayes moderation to assess differential expression was performed using the `lmFit()` and `contrasts.fit()` function from `limma`. In addition, the `limma_confects()` function of the `Topconfects` v1.6.0 `R/Bioconductor` package [57] was used. This function calls the function `eBayes()` from `limma` itself and builds on the `TREAT` method [104]. `Topconfects` calculates for each gene a confident effect size or "confect", a confident inner bound of the calculated effect size by `limma` while maintaining a given false discovery rate (FDR). "Confects" in this case can be considered as moderated log2 fold changes and are used to rank significant differentially expressed (DE) genes which is a more conservative method than ranking by raw log2 fold changes. A gene was considered as DE if the FDR-adjusted p-value was <0.05. Volcano plot representation of differential expression analysis is shown in Figure S6 An overview of DE genes from the *Discovery Set* genes is shown in Figure S7.

We observed a greater difference in the number of DE genes between Th1/Th2 compared to the other T-cell populations, especially after 2 hours of activation (see blue vertical bars in Figure S7). Th1 and Th2 are based on gene expression arrays, while Th0, Th17 and iTreg are RNA-Seq data. For gene expression arrays 21302 Ensembl gene ID from GENCODE could be mapped to microarray propes and evaluated, whereas for RNA-Seq 58721 Ensembl gene ID were evaluated. This could explain the difference in the number of DE genes between the populations.

3.1 Volcano plots
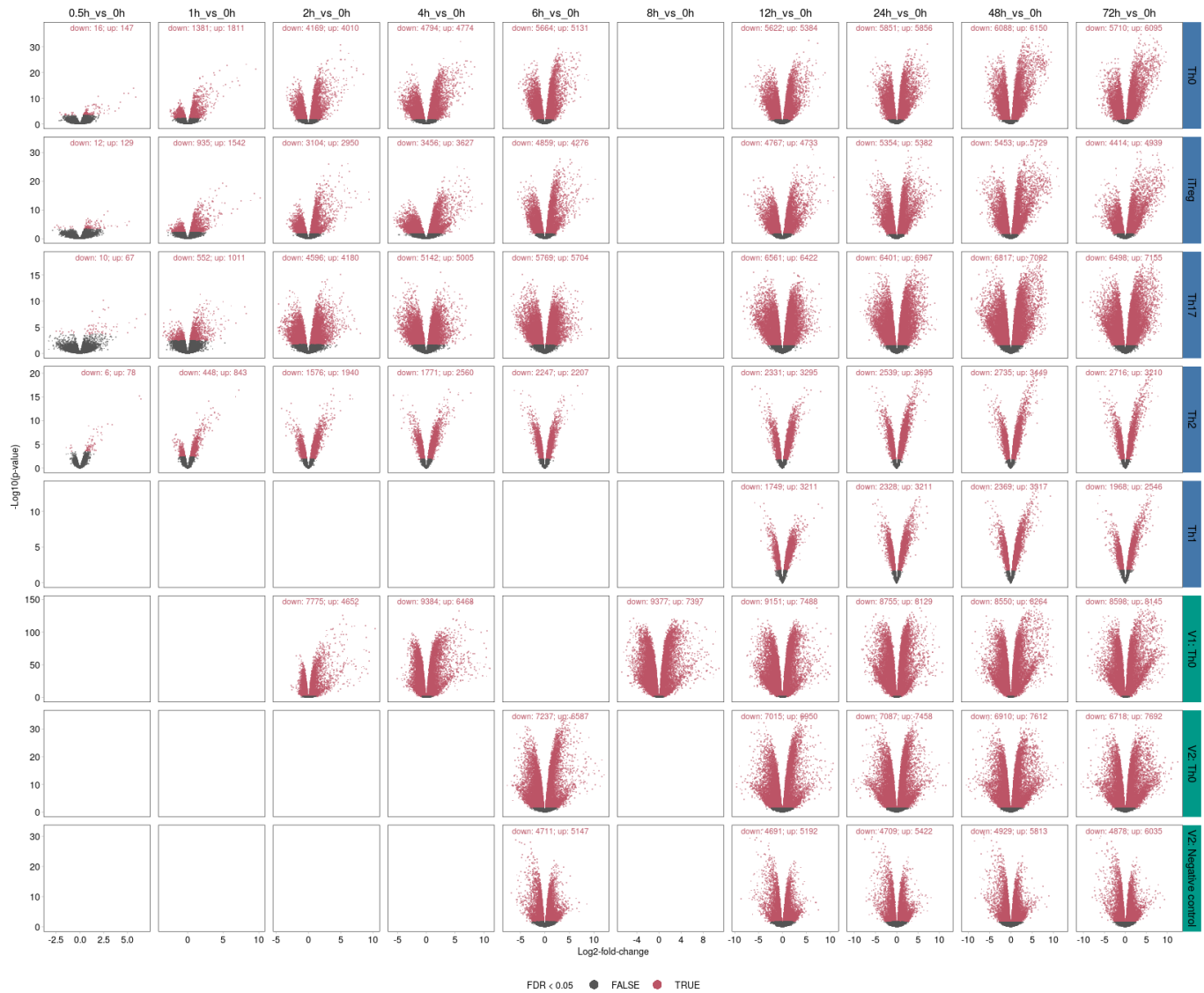
Fig. S6: *Discovery and Verification Sets*: **Volcano plots of DE genes for each contrast and T-cell population.** Red dots represent DE genes (FDR <0.05). The number of down- and up-regulated DE genes are shown at the top of each volcano plot. The y-axis denotes -log10 p-values while the x-axis represents log2 fold change values. Labels with blue color indicate the *Discovery Set*. Labels with green color indicate the *Verification Sets* (V1 = *Memory T-cell Verification Set*, V2 = *Pan T-cell Verification Set*.

3.2 *Discovery Set*: UpSet plots of DE genes
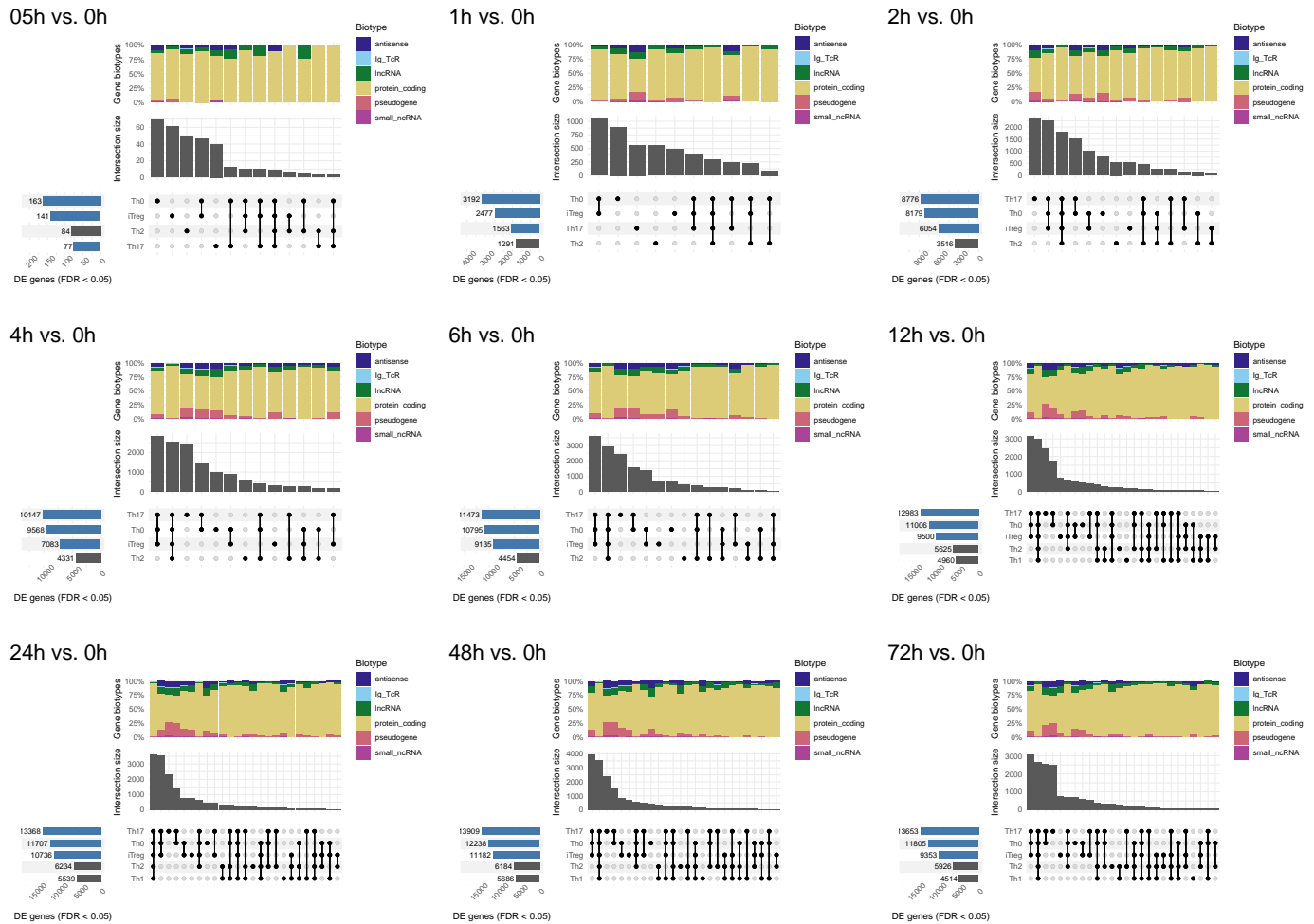
Fig. S7: *Discovery Set*: **Overview of DE genes for all CD4$^+$ T-cell populations** UpSet plots summarizes all DE genes (FDR <0.05) of activated CD4$^+$ T-cell populations compared to unactivated (0h) CD4$^+$ T-cell populations for each analyzed time point. In each panel, the bottom left horizontal bar graph indicates the total DE genes for each T-cell population. The blue horizontal bar graph indicate datasets from RNA-seq. The circles in each matrix represent unique and overlapping DE genes for the T-cell populations. Connected circles indicate a certain intersection of DE genes between the populations. The bar graph above the matrix in each panel summarizes the number of DE genes for each unique or overlapping combination. The top stacked barplot depicting the fraction of gene types (biotypes) for each unique or overlapping combination. With exception to the contrast "05h vs. 0h" the minimal number of observations in an intersection is set to 50
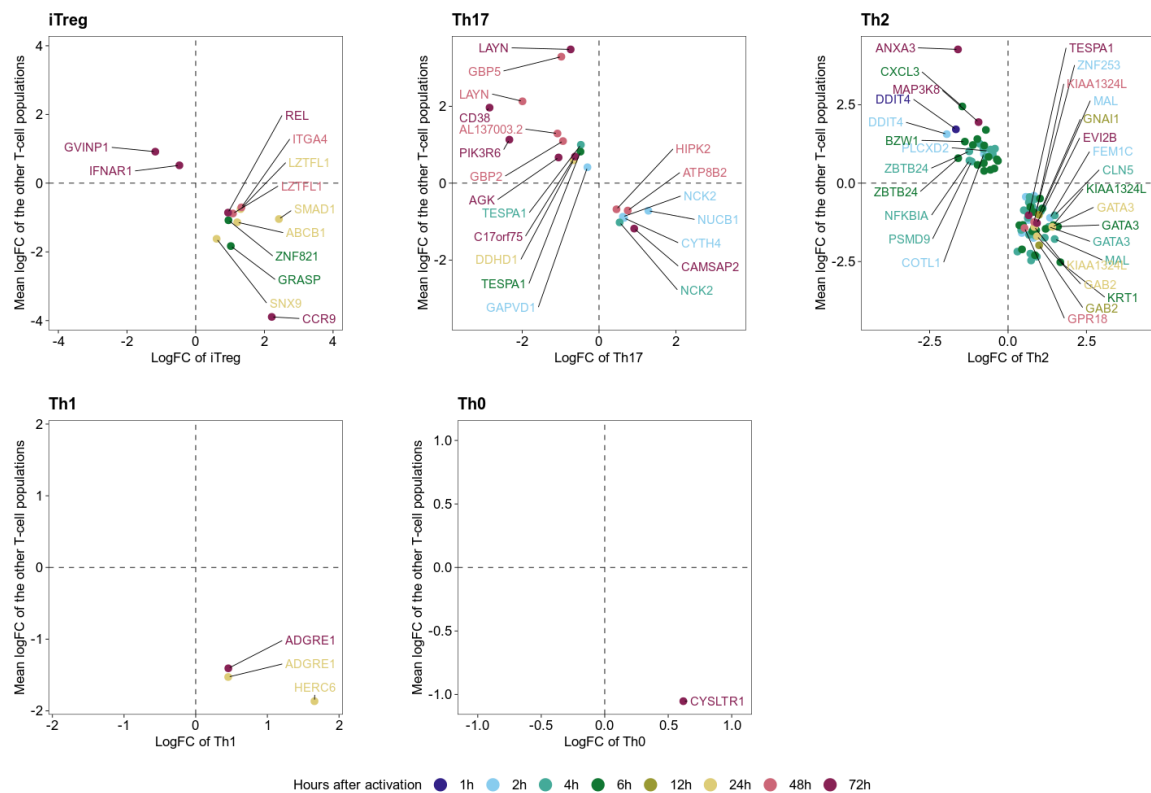
Fig. S8: *Discovery Set*: **Scatter plot of DE genes for each T-cell population showing opposite logFC compared to all other populations** For each T-cell population in the *Discovery Set* and at each time point of activation, we assessed differential gene expression between polarized or activated T-cells without differentiating cytokines and CD4+ T-cells before activation. We declared a gene as DE if the FDR-adjusted p-value was <0.05. Only genes that are significant in all populations at a given activation time point and have the same logFC direction in all but one population are shown. For Th2, due to a high number of DE genes, we annotated only the top three ranked genes (by logFC) for each time point (if available).

4   Meta-analysis

The calculation of the Hedges' g values [58, 59, 60] involves the calculation of the Cohen's d value:

$$\text{Cohen's } d_{ij} = \frac{\overline{x}_{ijPost} - \overline{x}_{ijPre}}{s_{ij}} \tag{1}$$

where x_{ijPost} and x_{ijPre} is the mean expression for each DE gene i of the post- and pre-treatment group in population j. Here we used the log2 fold changes as estimated in `limma`. For the pooled standard deviation s_{ij} we adopted the definition of the pooled variance as in the `limma` procedure, by borrowing information of variances from all tested genes, more specific, the square root of the posterior estimate of the variance as calculated by the empirical Bayes method from the eBayes() function. Since the Cohen's d value overestimates the effect size for studies with small sample sizes following correction factor was incorporated:

$$\text{Correction factor } J_{ij} = 1 - \frac{3}{4(n_{jPost} + n_{jPre}) - 9} \tag{2}$$

where n_{jPost} and n_{jPre} is the number of samples for the post- and pre-treatment group in the T-cell population j. The Hedges' g value for each DE gene can be computed as:

$$\text{Hedges' } g_{ij} = d_{ij} \times J_{ij} \tag{3}$$

The variance of the Hedges' g value was calculated by:

$$\text{V}g_{ij} = J_{ij}^2 \times \left(\frac{n_{jPost} + n_{jPre}}{n_{jPost} n_{jPre}} + \frac{d_{ij}}{2(n_{jPost} + n_{jPre})} \right) \tag{4}$$

Using the estimated Hedges'g values and the corresponding variances of the Hedges'g values, a meta-analysis with a random effects model was then performed. We analyzed each DE gene in at least 2 T-cell populations and time points of the *Discovery Set*. To assess the amount of heterogeneity post-hoc, the I^2 statistic was calculated. We observed increased heterogeneity ($I^2 > 75\%$) for 1548 genes with a significant combined effect size (FDR <0.05) across all 5 T-cell populations (Figure S9B). Considering the estimated standard deviation of log2 fold changes from the differential gene expression analysis, we observed that the estimated standard deviations are lower for the microarrays compared to the RNA-Seq datasets (Figure S10A). This results in higher Hedges'g values (equation 1 and 3) and consequently to higher estimated variances (equation 4 and Figure S10B). One explanation could be a wider dynamic range of the gene expression measurements between platforms. Gene expression intensities for microarrays are limited by background at the low end and signal saturation at the high end. On the other hand, RNA-Seq platforms provide discrete read counts and can quantify gene expression across a wider dynamic range, as it is determined only by the amount of sequencing obtained [105]. This larger dynamic range can lead to a higher variance of the read counts between replicates.
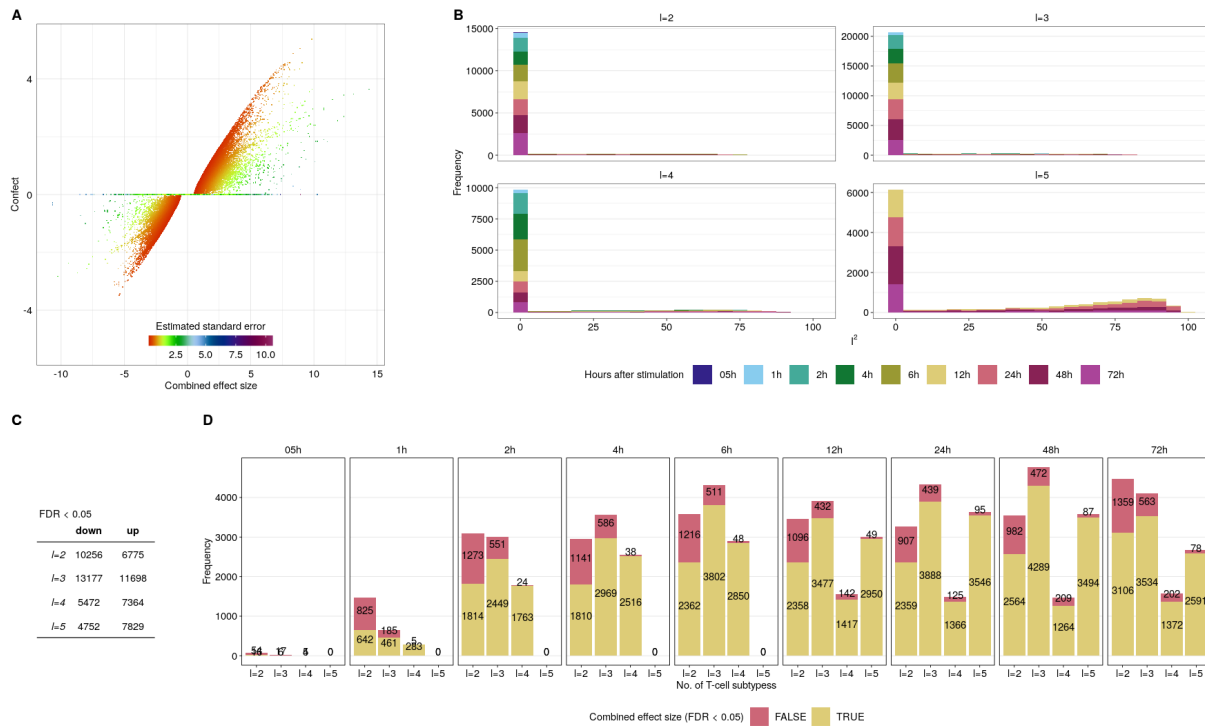
Fig. S9: *Discovery Set*: **Differentially expressed genes with a significant combined effect size across the CD4$^+$ T-cell populations identified by meta-analysis. (A)** Shown are DE genes with a combined effect size identified in the meta-analysis in at least 2 CD4$^+$ T-cell populations. The x-axis represents the combined effect size, the y-axis the "confect" value. Genes that do not show a significant combined effect size (FDR >0.05) have a "confect" value of 0. Genes are color-coded according to the estimated standard error calculated using the R package metafor. **(B)** To assess the amount of heterogeneity post-hoc, the I^2 statistic was calculated. The different plots indicate DE genes with a significant combined effect size across l T-Cell populations. **(C-D)** Number of DE genes with a significant combined effect size (FDR <0.05) across l populations are illustrated by yellow bar graphs
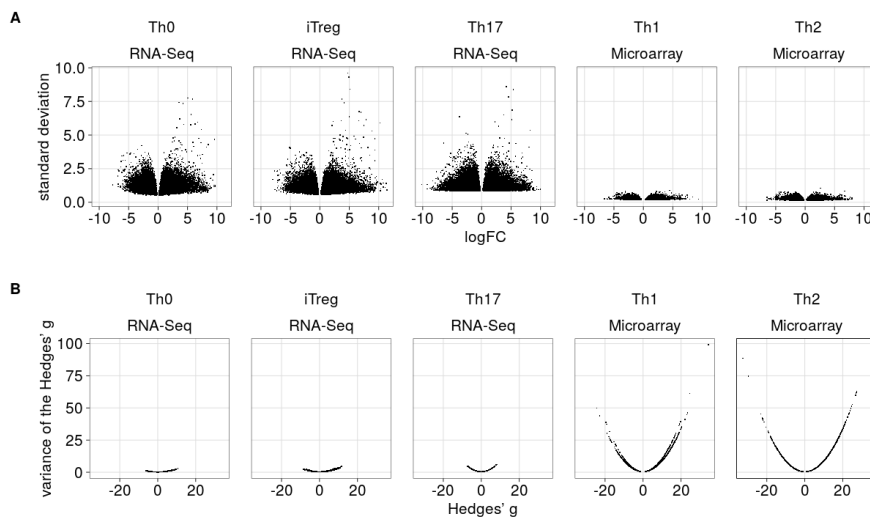

Fig. S10: *Discovery Set*: **Volcano plots of log2 fold changes, Hedges' g values and corresponding estimated variances**. **(A)** Log2 fold changes and estimated standard deviations of all DE genes from the differential gene expression analysis by limma for each T-cell population. **(B)** Hedges' g values (equation 3) and estimated variances of the Hedges' g values (equation 4) used as input for meta-analysis.

4.1 Highest ranked genes from the meta-analysis

1h

2h
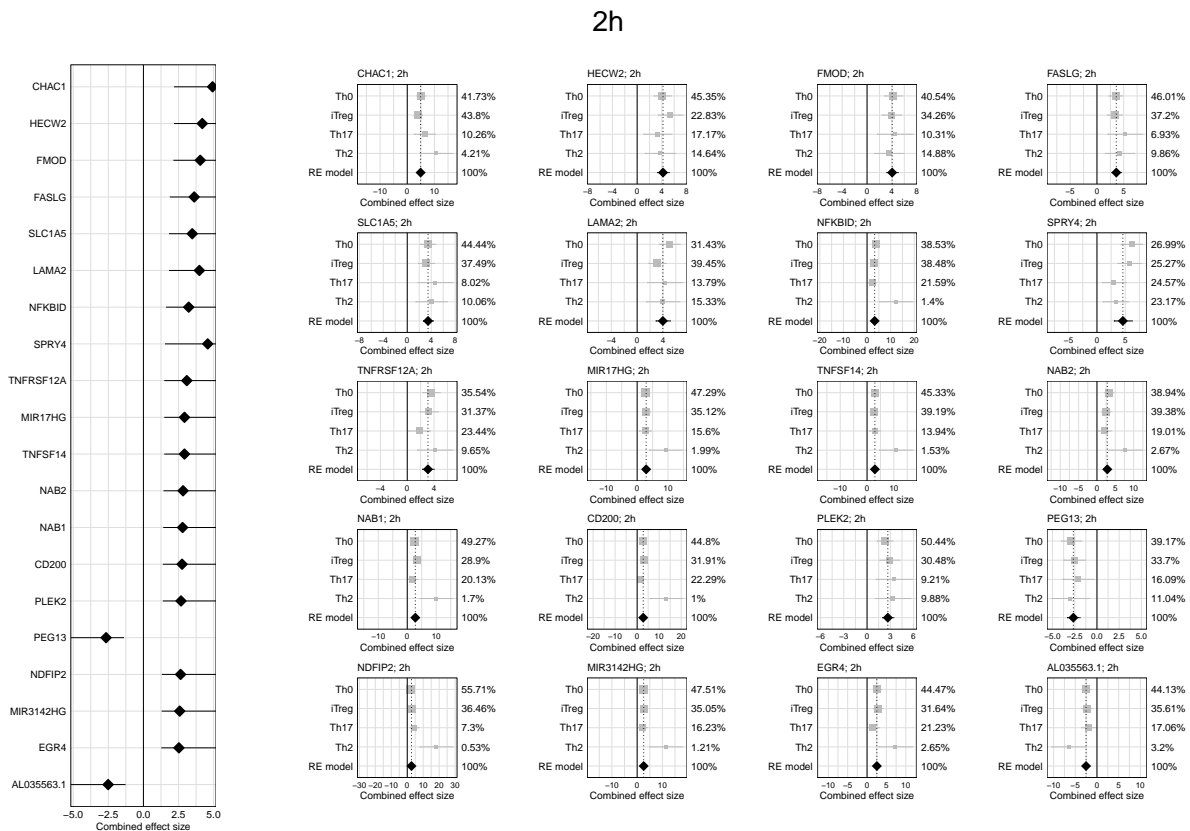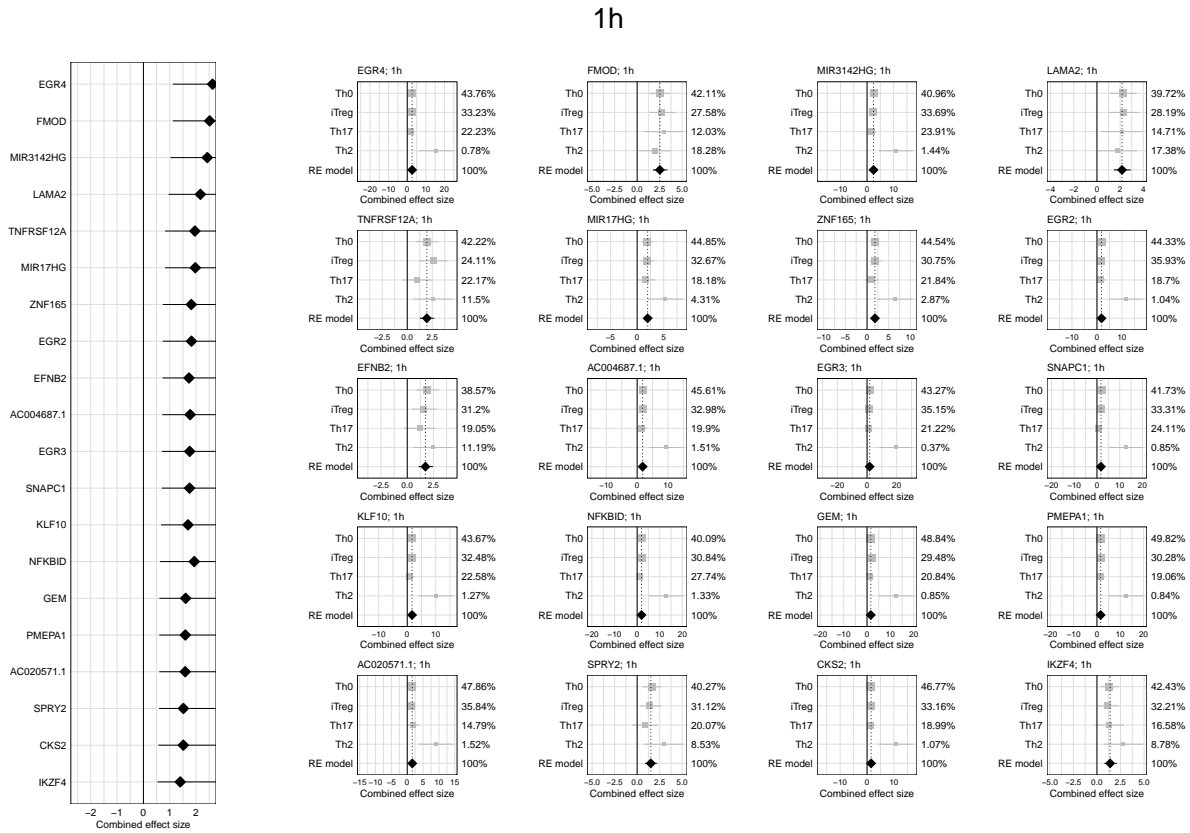
4h
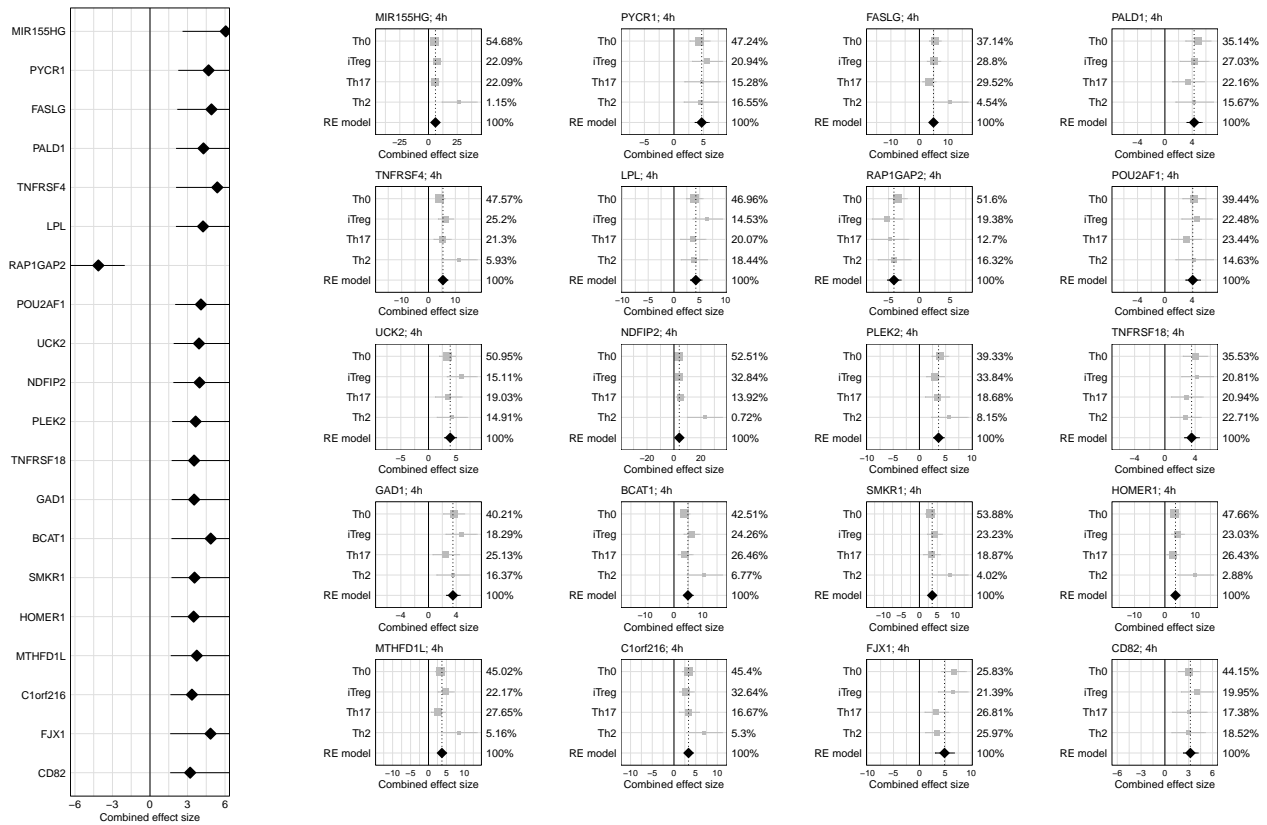
6h

12h

24h
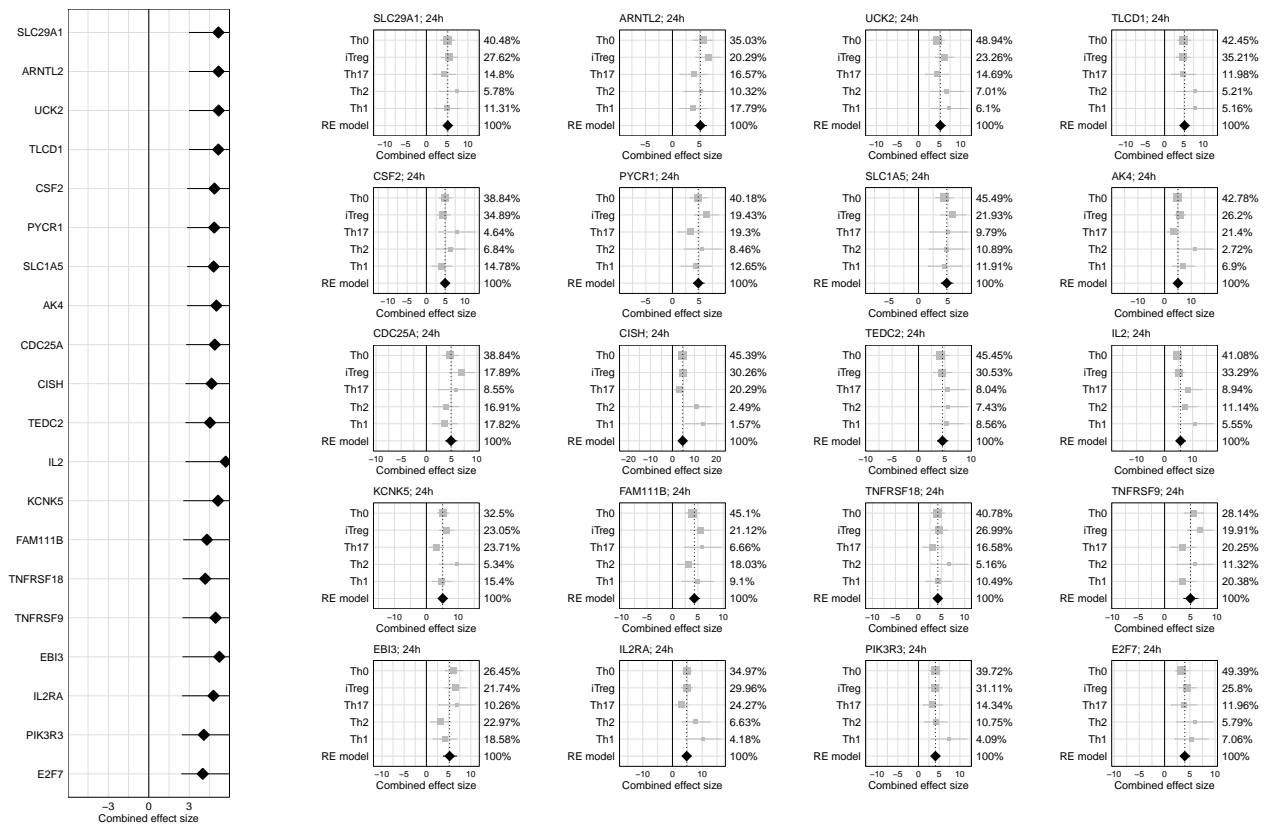
48h
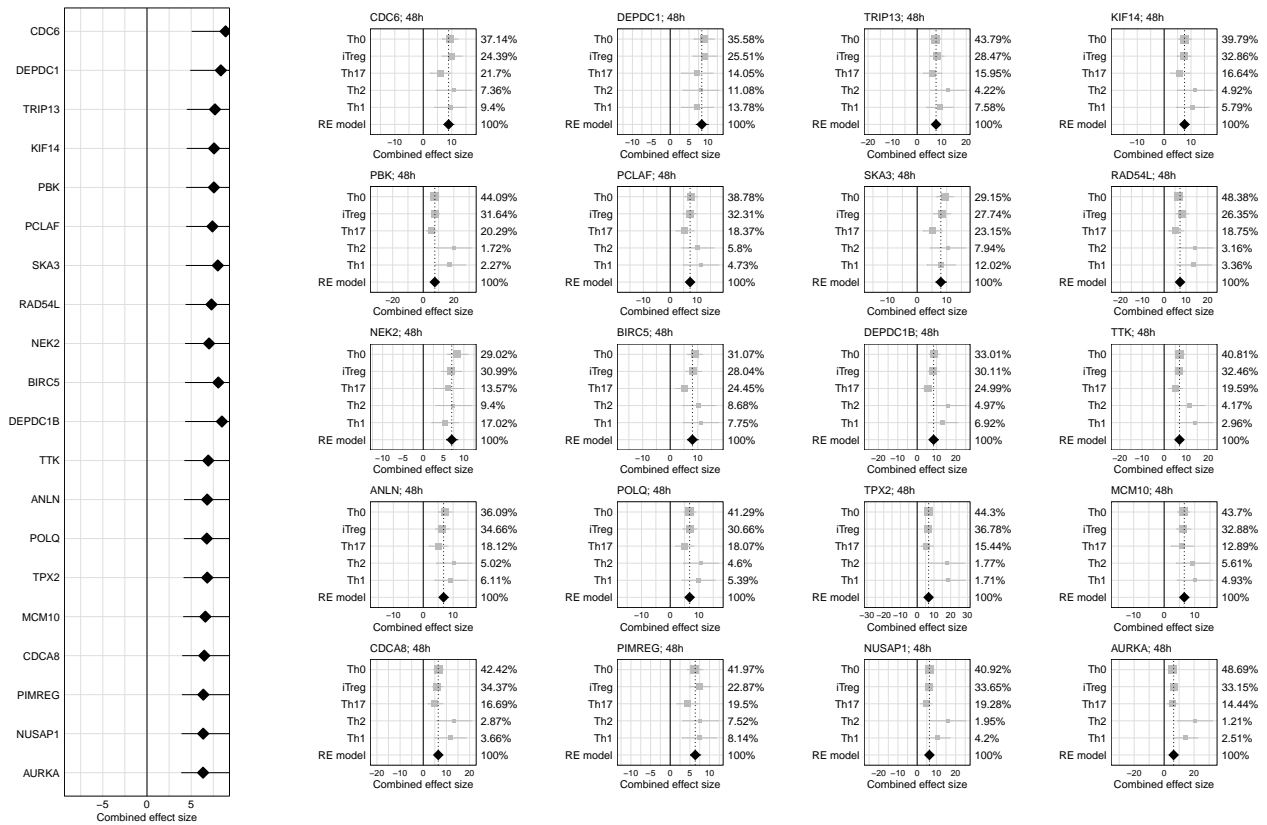
72h
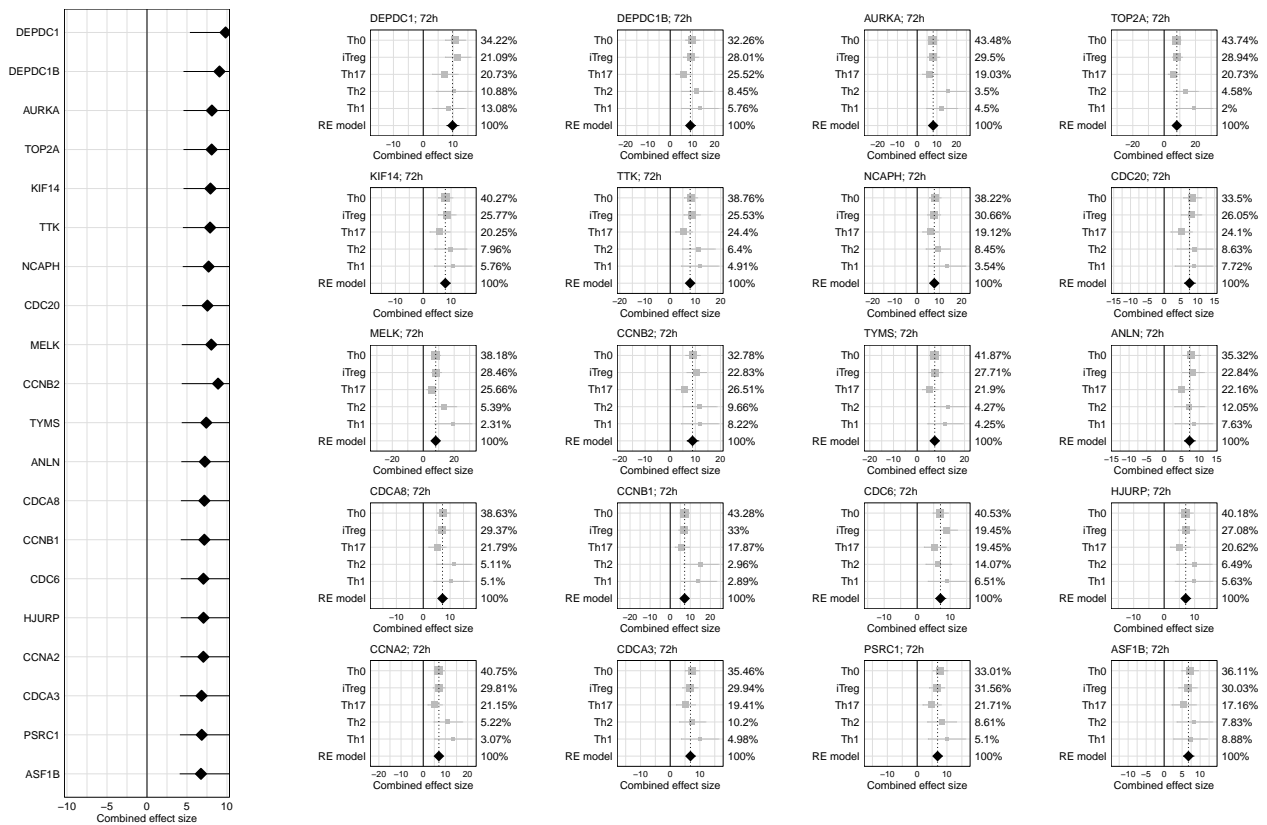
Fig. S11: *Discovery Set*: "Confect" values and forest plots of the 20 highest ranked genes from the **meta-analysis using a random-effect model. Left panel:** DE genes with a significant combined effect size (FDR <0.05) across the 4 avai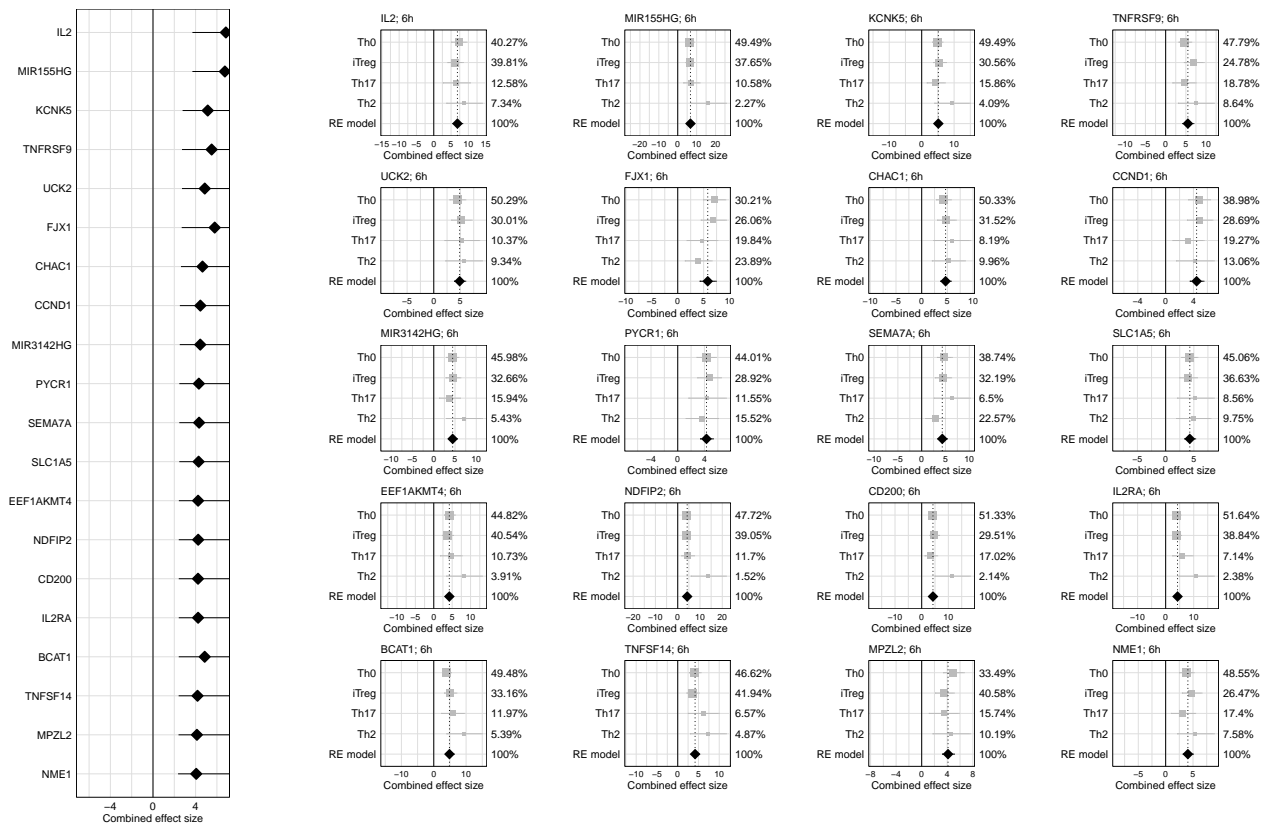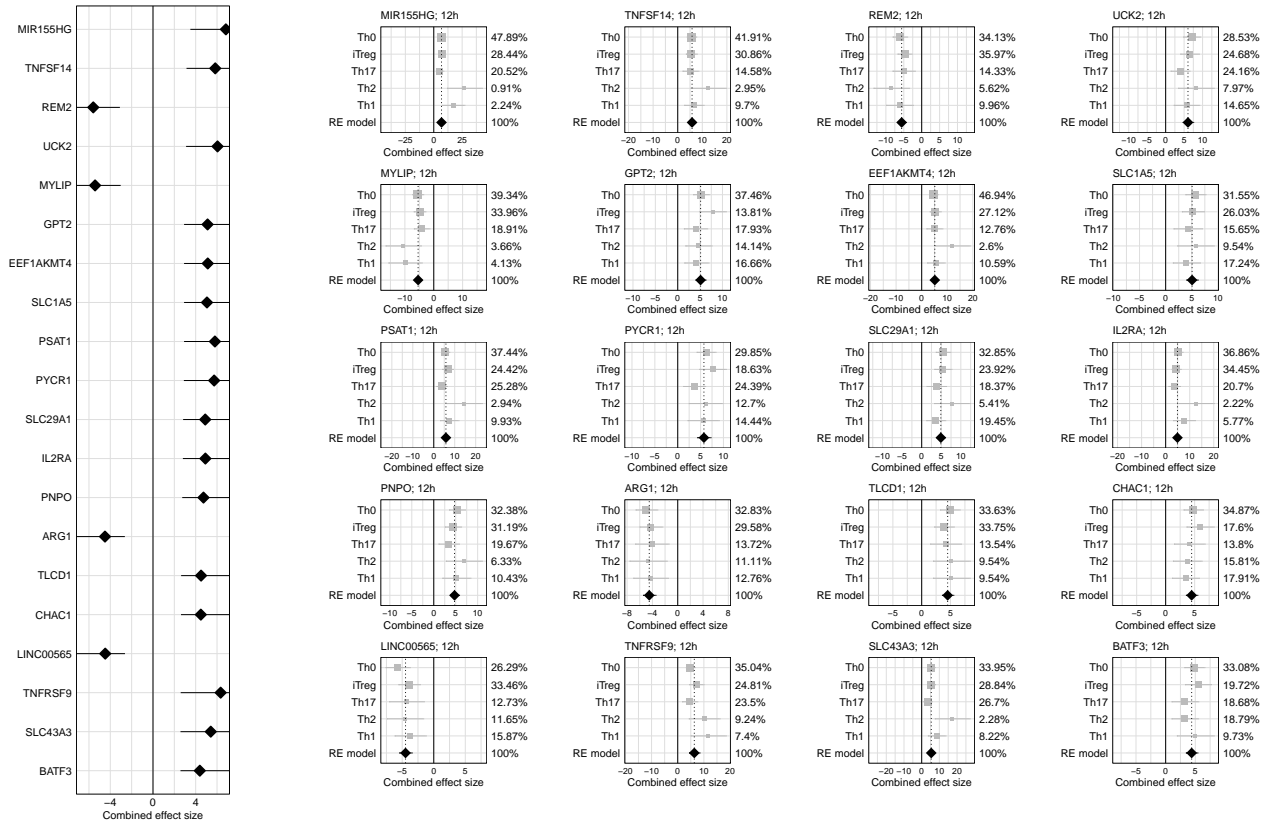lable populations after 0.5 to 6 hours and across all 5 populations after 12 to 72 hours of activation are shown. Genes are ranked by "confect" values. The diamonds show the estimated combined effect size from the meta-analysis and the inner end of the horizontal line shows the "confect" value (inner confidence bound). **Right panel:** Corresponding forest plots for 20 highest ranked genes. The confidence interval (CI) for each T-cell population is represented by a horizontal line, and the effect size (Hedges'g values) is represented by a square. The size of the square corresponds to the weight of the population in the meta-analysis. These weights are also shown as numbers on the right-hand side. The combined effect is represented by a diamond and the CI by the vertical line.

4.2 Enrichment analysis
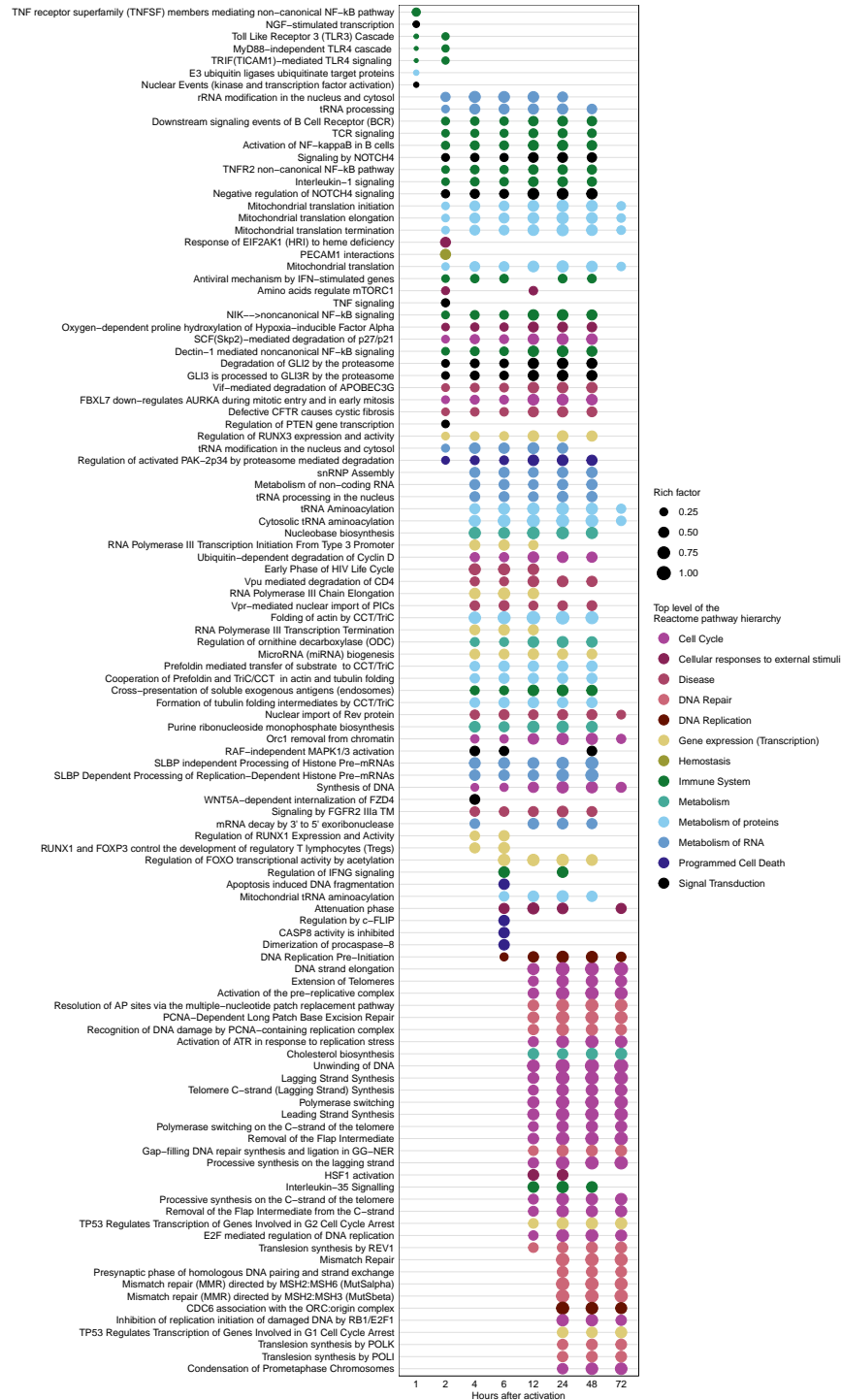
Fig. S12: *Discovery Set*: **Reactome pathway enrichment analysis of DE genes with a significant combined effect size for each analysis time point.** Dot plot depict the top 30 (sorted by rich factor) significantly enriched Reactome pathways (FDR <0.05) for DE genes with a significant combined effect size in 4 (0.5 to 6 hours) and 5 (12 to 72 hours) CD4[+] T-cell populations. The dot size indicates the rich factor which is the number of DE genes with a significant combined effect size in the pathway divided by the number of background genes in the pathway. Colors depicted the uppermost hierarchical level of Reactome database.

Fig. S13: **Discovery Set: GO term enrichment analysis of DE genes with a significant combined effect for each analysis time point.** Dot plot depict the top 30 (sorted by rich factor) significantly enriched GO term for biological processes (FDR <0.05) for DE genes with a significant combined effect size in 4 (0.5 to 6 hours) and 5 (12 to 72 hours) CD4$^+$ T-cell populations. The dot size indicates the rich factor which is the number of DE genes with a significant combined effect size in the GO term divided by the number of background genes in the GO term.
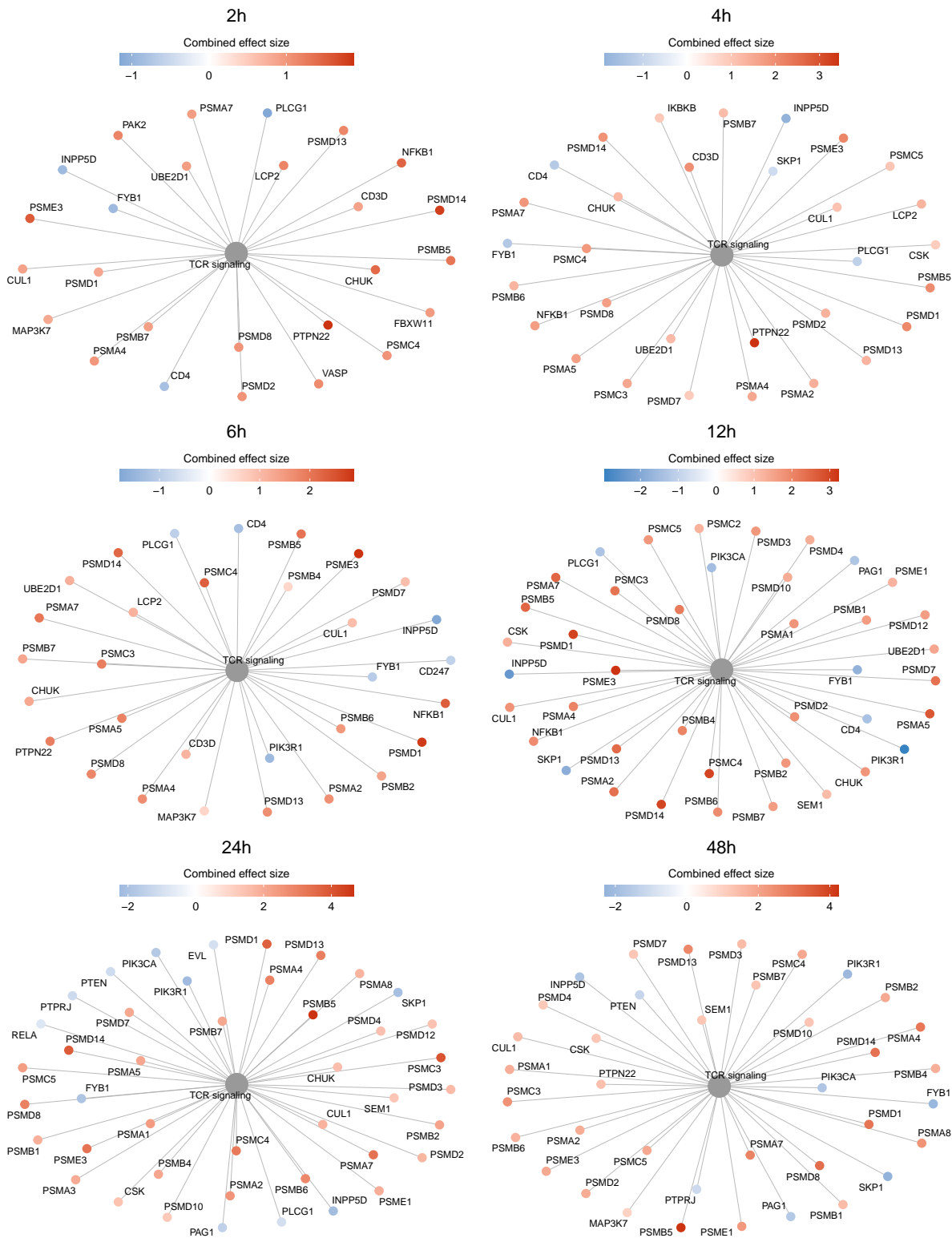
Fig. S14: *Discovery Set*: **T-cell receptor pathway**. Significantly (FDR <0.05) enriched Reactome T-cell receptor (TCR) pathway of DE genes with a significant combined effect size (FDR <0.05) from the meta-analysis. The linkages of genes and the TCR pathway are shown as a network. Genes with a significantly negative combined effect size are shown in blue, genes with a significantly positive combined effect are shown in red. Only those time points are shown in which the TCR signalling pathway is also significantly enriched.

5   Non-negative matrix factorization

Non-negative matrix factorization (NMF) is a form of dimension reduction where the $n \times m$ matrix X is approximated by:

$$\underset{\mathbb{R}^{n \times m}}{X} \approx \underset{\mathbb{R}^{n \times k}}{W} \times \underset{\mathbb{R}^{k \times m}}{H} \text{ such that } W, H \geq 0$$

where the rank $k \leq max(m,n)$ (k = number of metagenes). We want to find a basis W^* and coefficients H^* matrix which minimize the reconstruction error according to:

$$\underset{W,H \geq 0}{min} D(X, WH) \text{ w.r.t. } W, H$$

where $D(X, WH)$ is a loss function that measures the quality of the approximation. You can use any distance measure which evaluates how well the low-dimensional matrix WH approximates X. Here we used the Kullback-Leibler divergence

$$D(X, WH) = \sum_{i,j} (X_{ij}(\log \frac{X_{ij}}{(WH)_{ij}}) - X_{ij} + (WH)_{ij}).$$

The matrices W and H are found by maximizing the log-likelihood under an independent Poisson assumption of X. However, this is a non-convex optimization problem. The convexity is given for W or H, but is not given in respect to both variables together and therefore there is no unique solution. The optimization can be done with multiplicative rules [65] which update W and H according to the following rule:

$$H_{kj} \leftarrow H_{kj} \frac{\sum_{i=1}^{n} (W_{ik}X_{ij})/(WH)_{ij}}{\sum_{i=1}^{n} W_{ik}} \quad W_{ik} \leftarrow W_{ik} \frac{\sum_{j=1}^{m} (H_{kj}X_{ij})/(WH)_{ij}}{\sum_{j=1}^{m}}$$

The initialization of W and H was generated as random seed matrices drawn from a uniform distribution within the same range as the entries in the matrix X.

We used the model selection method proposed by Brunet et al. [65] to estimate the optimal number of ranks. We performed NMF on the normalized data (see main methods) by repeating the rank value in the interval 2–10. Due to the unavoidably converges to local minima, we ran the NMF algorithm 200 times for each rank. For each run, a connectivity matrix C of the size $m \times m$ was calculated. Matrix C has an entry of 1 if samples i and j cluster together (assigned to the same metagene) and 0 otherwise, where $ij = 1, \ldots, m$. The sample assignment here depends on the relative values in each column of H. The consensus matrix $\overline{C}$ is computed as the average connectivity matrix over the 200 runs. The entries of $\overline{C}$ reflect the probability that samples i and j belong to the same cluster. This yields the empirical probability for each pair of samples clustered together. Hierarchical clustering was performed to reorder the samples (rows and columns of the consensus matrix).

The visualization of the clustered matrix $\overline{C}$ can be used as a qualitative measurement to assess the stability of the clusters (see Figure S16A). As a quantitative measurement we can use the cophenetic correlation coefficient, which indicates the dispersion of the consensus matrix $\overline{C}$. It is computed as the Pearson correlation of two distance matrices. The first represent the sample distances induced by the consensus matrix. The second is the cophenetic distance obtained by hierarchical clustering [65] (see Figure S16B). If all entries in the matrix $\overline{C}$ are 1 or 0, the cophenetic correlation coefficient is 1. As another measure for cluster performance of the consensus matrix, we used the average silhouette width (see Figure S16C).

5.0.1 Defining of metagene associated genes
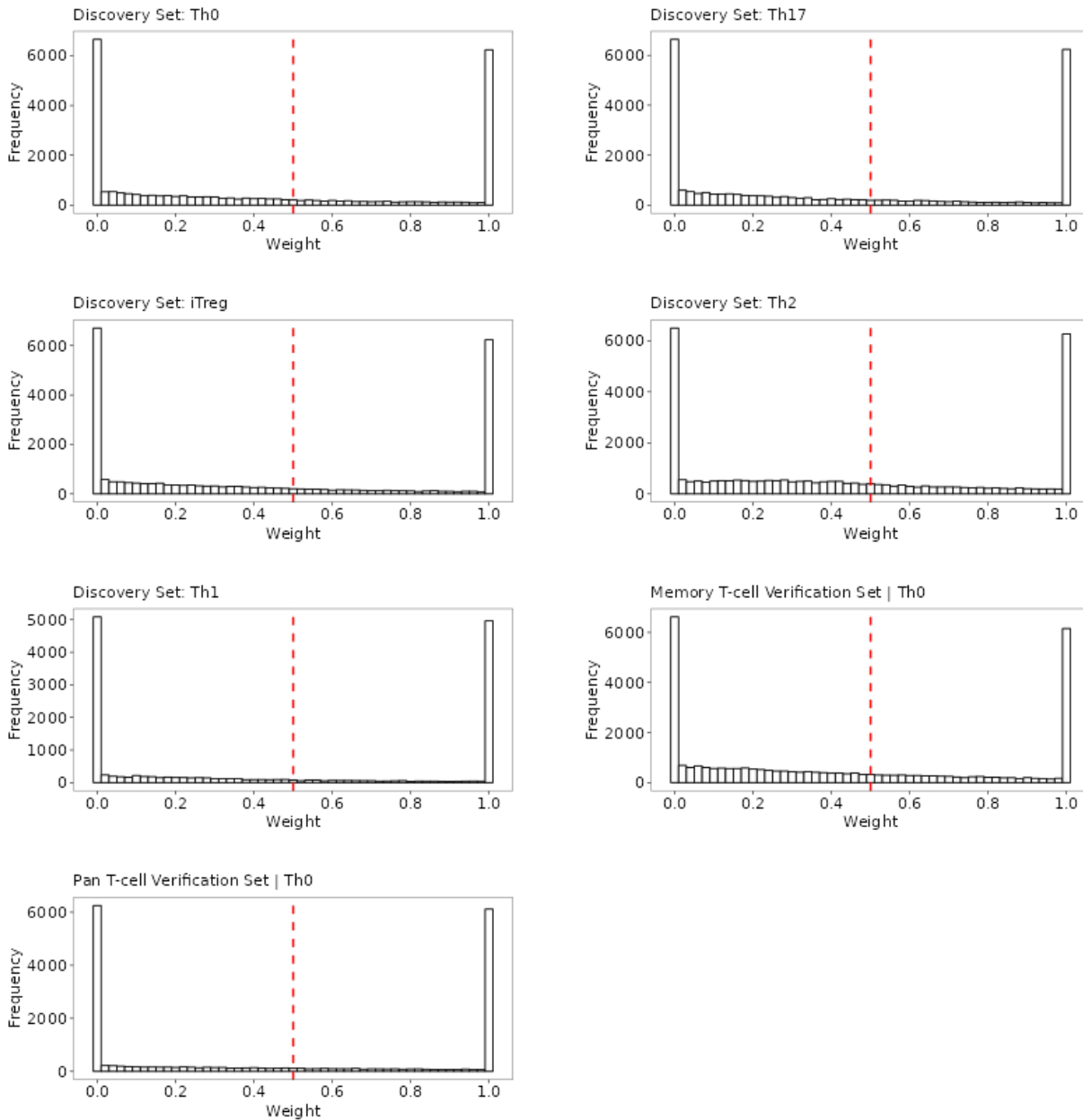
Fig. S15: *Discovery and Verification Sets*: **Distribution of normalized weights from the gene signature matrix** W **for each T-cell population**. Each row in the gene signature matrix W was min-max normalized. The red dashed line represents the cutoff (>0.5) at which gene i from the matrix W is assigned to metagene j.

5.0.2   NMF consensus clustering

Discovery Set | Th0

Discovery Set | Th2

A

k=2 — Cophenetic coef = 0.994
k=3 — Cophenetic coef = 0.996
k=4 — Cophenetic coef = 1
k=5 — Cophenetic coef = 0.997
k=6 — Cophenetic coef = 0.979
k=7 — Cophenetic coef = 0.967
k=8 — Cophenetic coef = 0.932
k=9 — Cophenetic coef = 0.939
k=10 — Cophenetic coef = 0.957
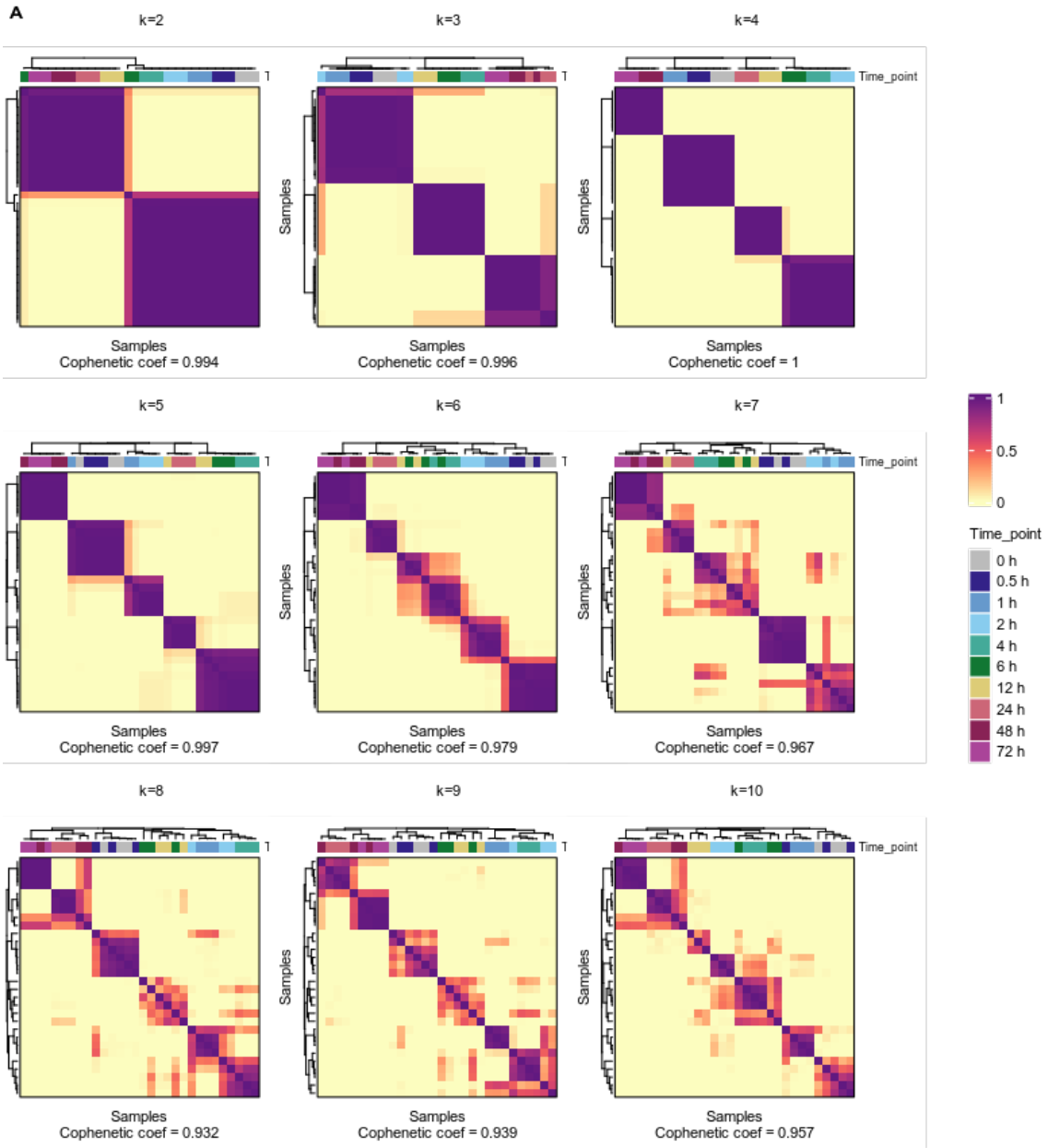
Time_point: 0 h, 0.5 h, 1 h, 2 h, 4 h, 6 h, 12 h, 24 h, 48 h, 72 h

B

Cophenetic coefficient — actual, randomized

C

Avg silhouette width — actual, randomized

Discovery Set | iTreg

Discovery Set | Th1

Discovery Set | Th17
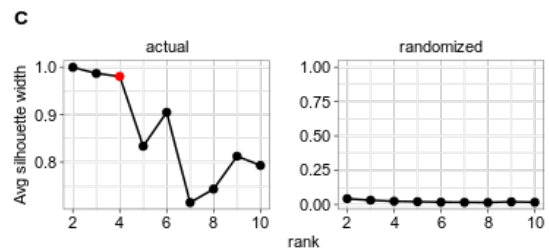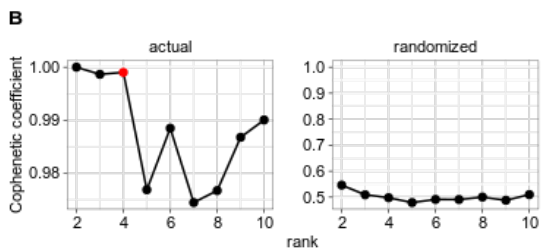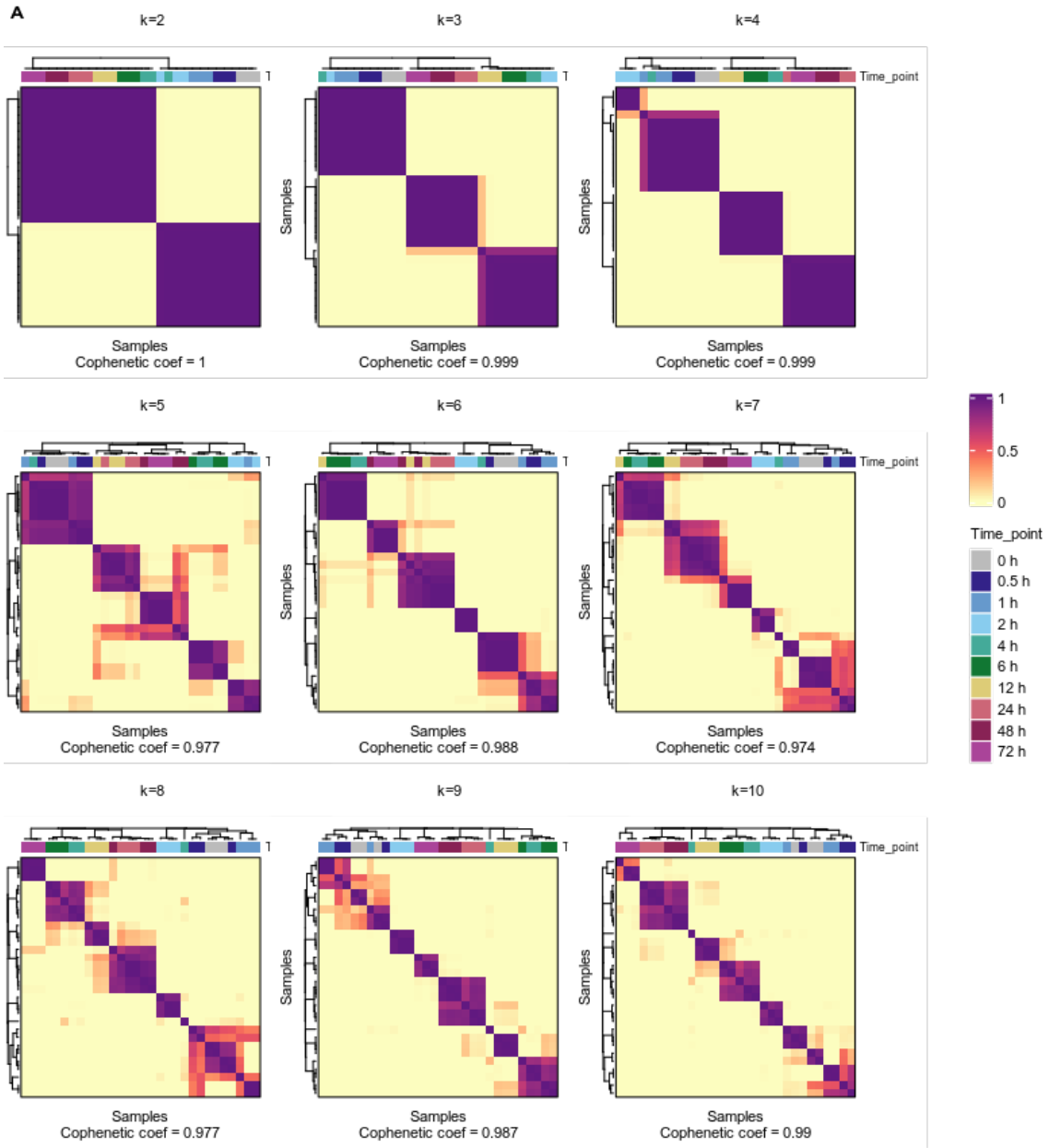
Memory T-cell Verification Set | Th0
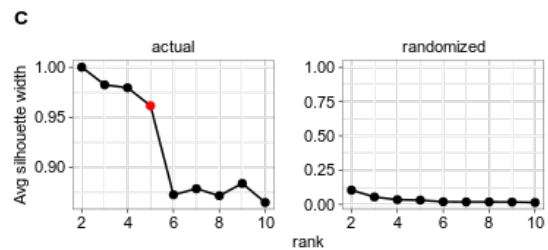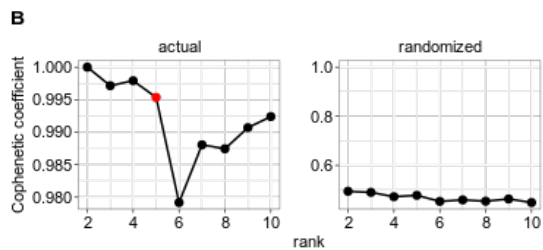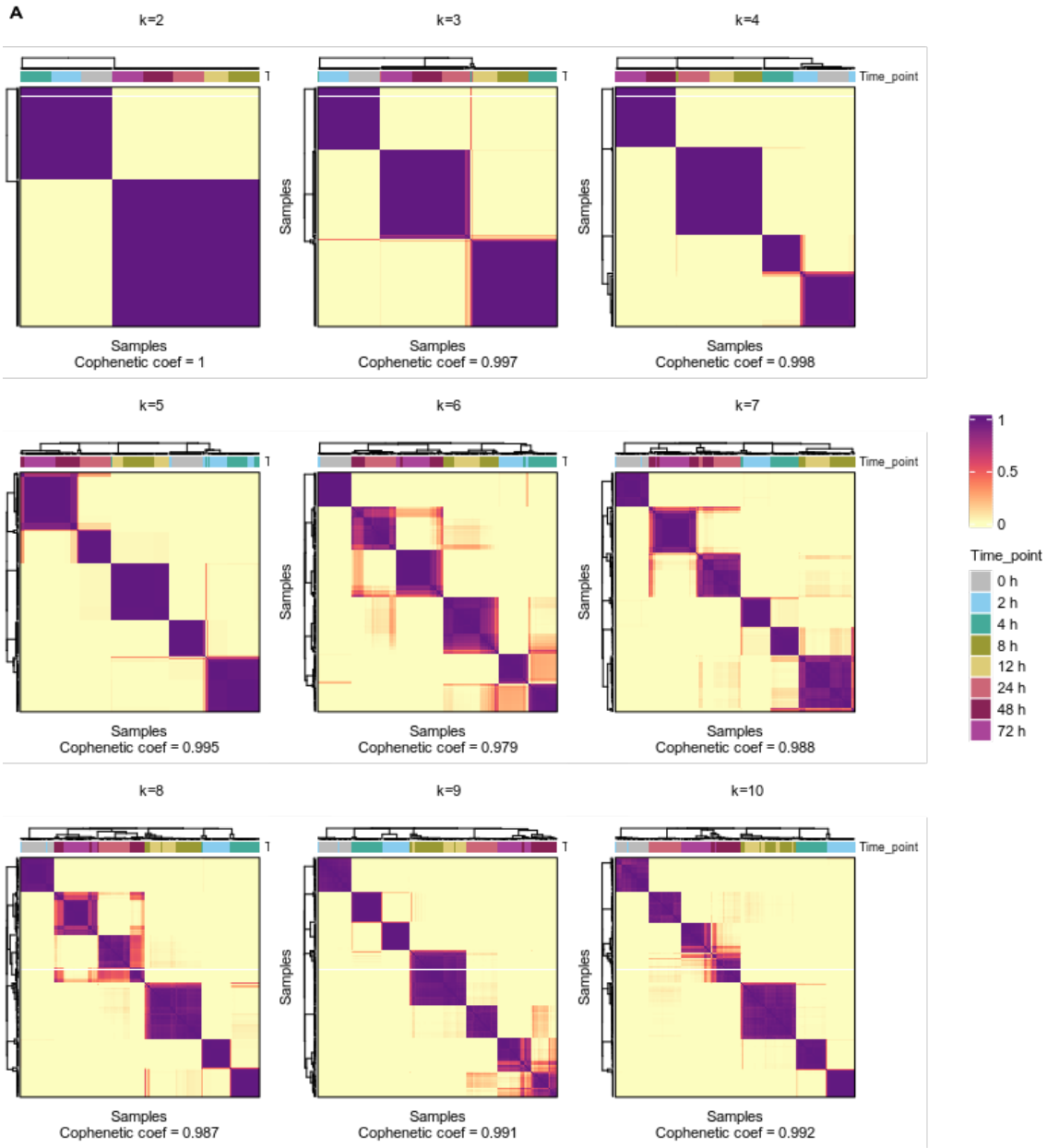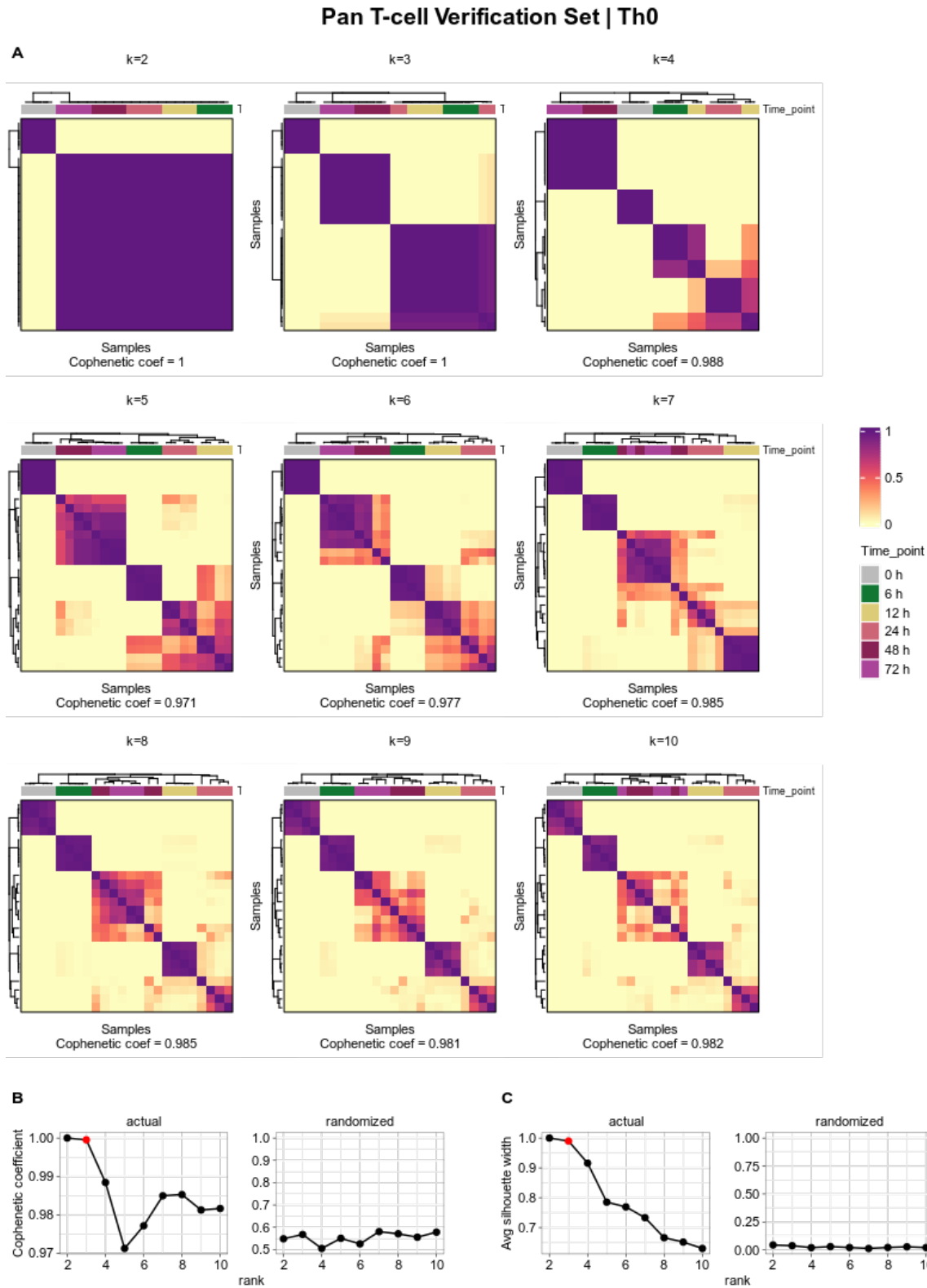
Fig. S16: **NMF consensus-clustering**. **(A)** Consensus matrices averaging 200 connectivity matrices and computed for factorization ranks k = 2–10 (number of metagenes). NMF was performed for each T-cell population separately. NMF computation and model selection were performed according to Brunet et al. [65] as described in the methods. Samples are hierarchically clustered by using distances derived from consensus clustering matrix entries, colored from 0 (samples are never in the same cluster) to 1 (samples are always in the same cluster). **(B-C)** Cophenetic correlation coefficient for the above hierarchically clustered matrices and average silhouette-width plots. NMF was also run with randomized data with same parameters as for the actual data. Red colored dots represent the chosen factorization rank for downstream analysis.

5.1 *Discovery Set*

5.1.1 25 highest ranked genes associated with metagenes
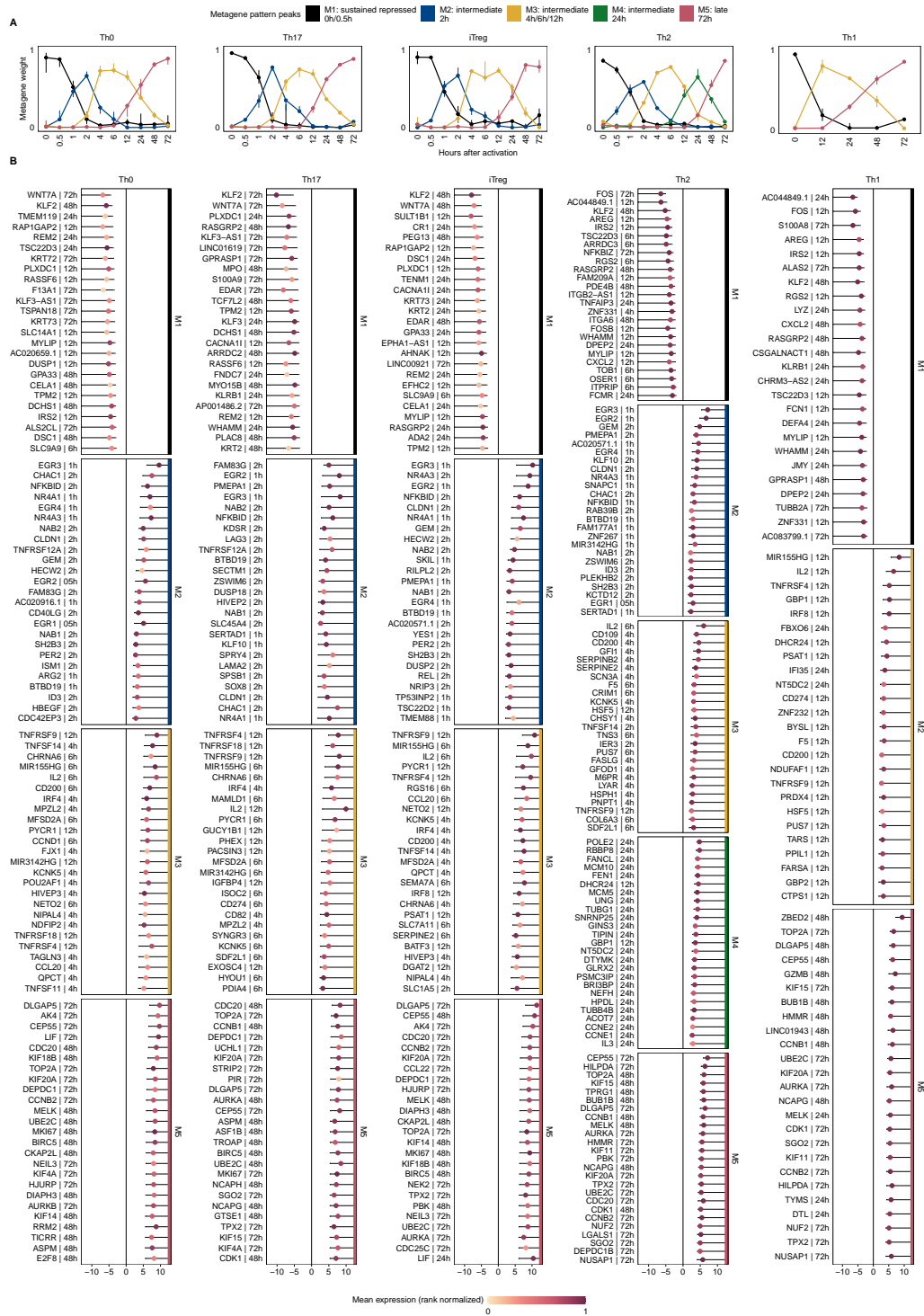
Fig. S17: ***Discovery Set*: 25 highest ranked genes associated with metagenes for each T-cell population**. **(A)** The pattern matrix for each T-cell population from the *Discovery Set* is shown as continuous profiles, with samples assigned to time points of activation. We scaled each column in the pattern matrix to sum up to one. Dots depict the median weights for all samples from identical analysis time points. Vertical lines represent interquartile ranges. We annotated and colored the metagenes based on their maximum median values across all analysis time points. The time point with maximum median value is depicted in the legend. **(B)** For each CD4[+] T-cell population and gene used for NMF, we used the highest absolute "confect" value estimated in the DGEA across all contrasts (e.g. 12h vs. 0h). Genes are ranked by "confect" values. Dots represent log2 fold changes for contrasts with the highest absolute "confect" value. The time point to the right of the gene represents the contrast with the highest absolute "confect" value. Color of dots correspond to rank normalized average expression values of the activation group in the contrast with the highest absolute "confect". The inner end of the horizontal line shows the "confect" value (inner confidence bound).

5.1.2 Enrichment analysis for gene sets associated with metagenes

Fig. S18: *Discovery Set*: **Enrichment analysis of metagene associated genes for each T-cell population.** The top 15 (sorted by rich factor) significantly (FDR <0.05) enriched in Reactome pathways of metagene-associated genes for each CD4+ T-cell population. The dot size indicates the rich factor, which is the number of metagene associated genes in the pathway divided by the number of background genes in that pathway. Colors depicted the uppermost hierarchical level of the Reactome database.
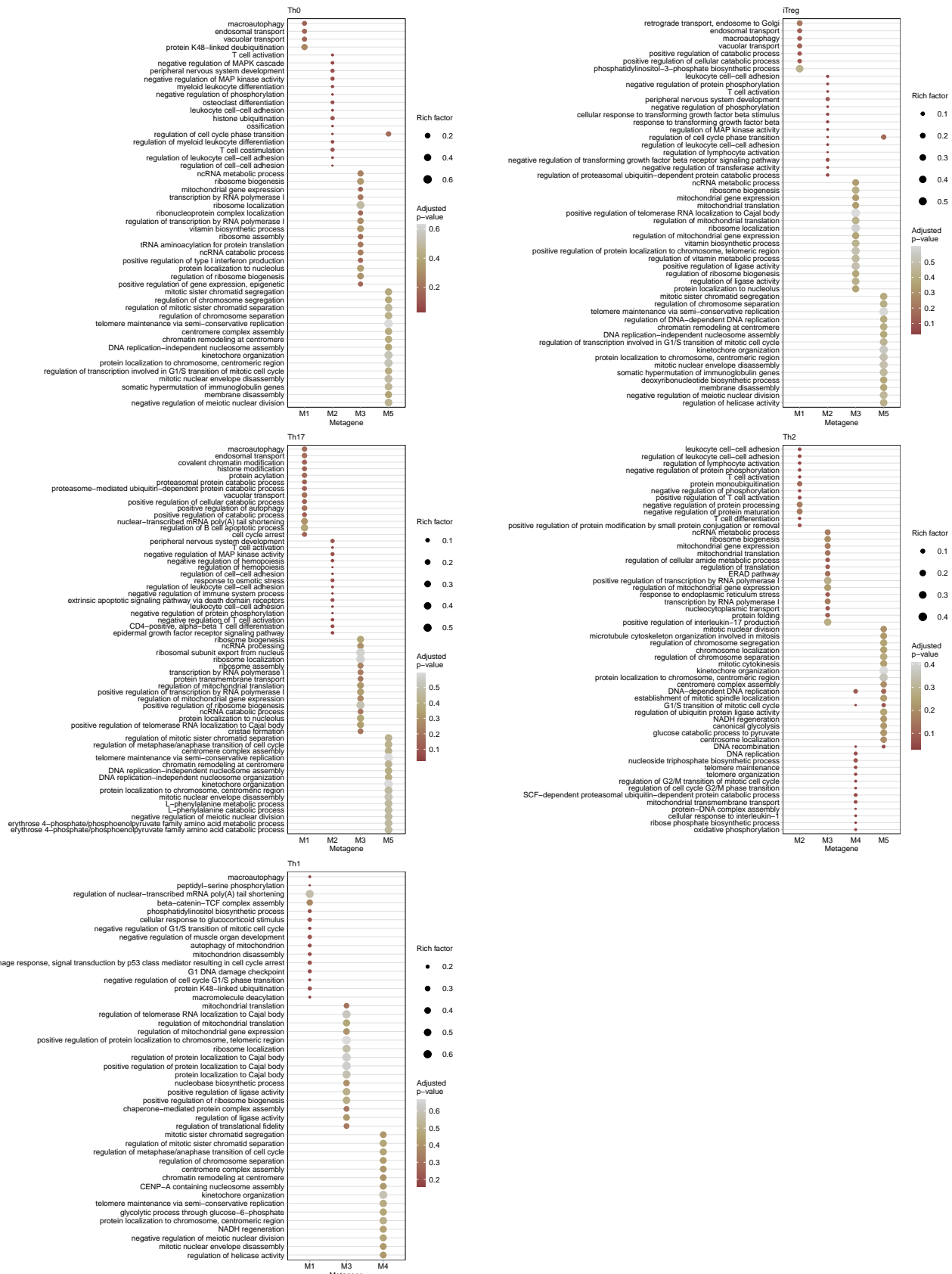
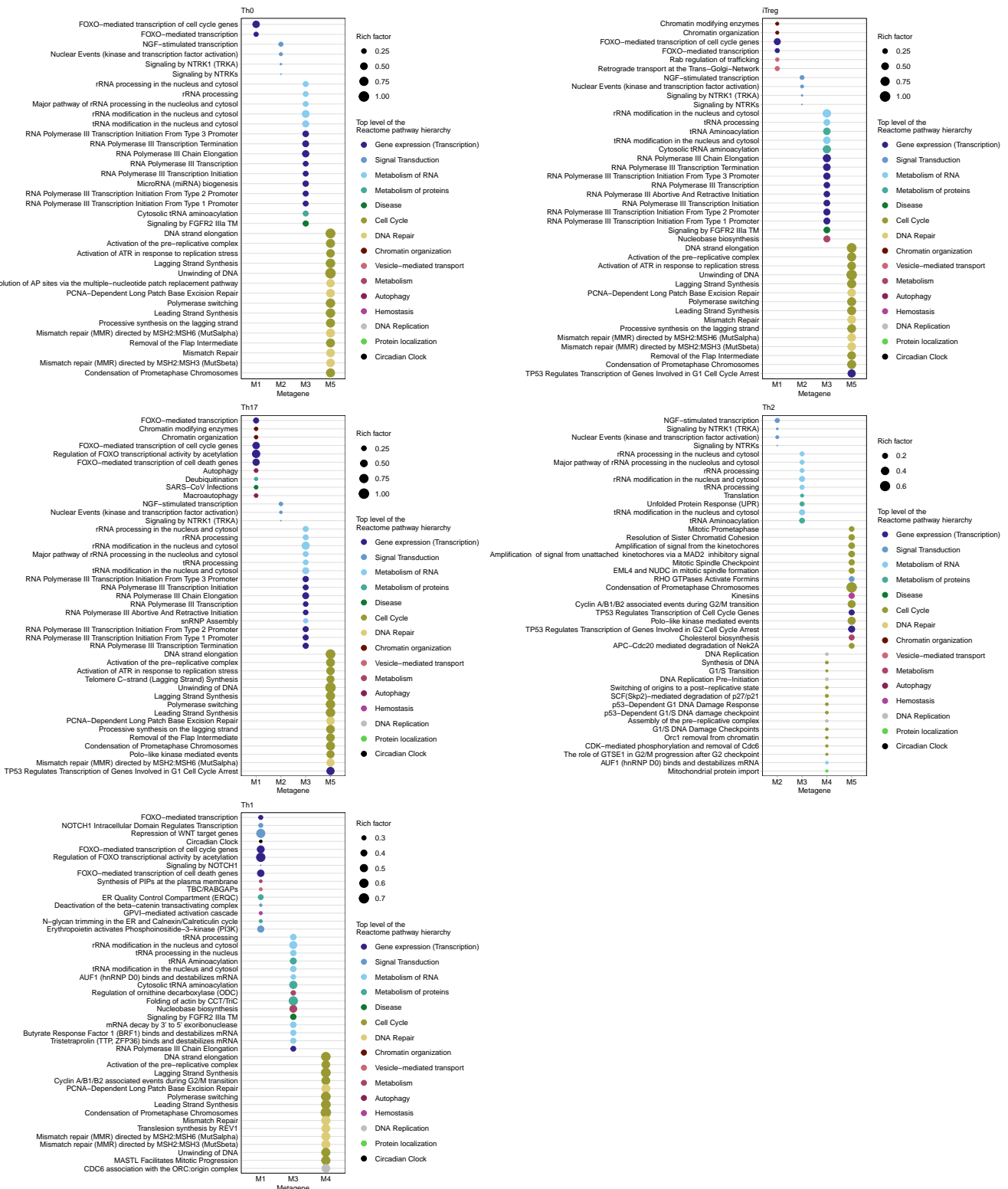Fig. S19: *Discovery Set*: **Enrichment analysis of metagene associated genes for each T-cell population.** The top 15 (sorted by rich factor) significantly (FDR <0.05) enriched in Gene-Ontology terms (bottom) of metagene associated genes for each CD4$^+$ T-cell population. The dot size indicates the rich factor, which is the number of metagene associated genes in the term divided by the number of background genes in that term.

5.1.3 Metagene landscape (combined pattern matrix)
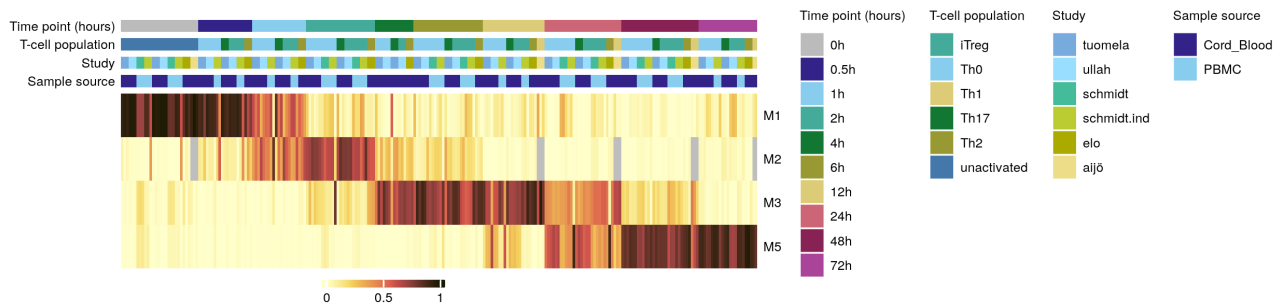
Fig. S20: ***Discovery Set*: Heatmap of the concatenated pattern matrix**. The pattern matrix H was obtained from the factorization that achieved the lowest approximation error across 200 runs of the algorithm from Brunet et al. [65]. Factorization and subsequent rank determination was performed for each T-cell population (see methods). The matrices H from each population were concatenated. For this, we used only temporally consistent metagenes across the T-cell populations. The columns in the combined matrix H were scaled to sum to one and ordered by analysis time points. Since metagene 2 (expression peak at 2 hours) is not available for the Th1 population (only the analysis time points 12 to 72 hours are available), the values at Th1 were set to NA for metagene 2. The gray values in the combined matrix H represented the NA values. The combined pattern matrix was used to develop the Metagene landscape (see main methods).

5.1.4 Enrichment analysis (Reactome) for gene sets from the consensus gene expression profiles
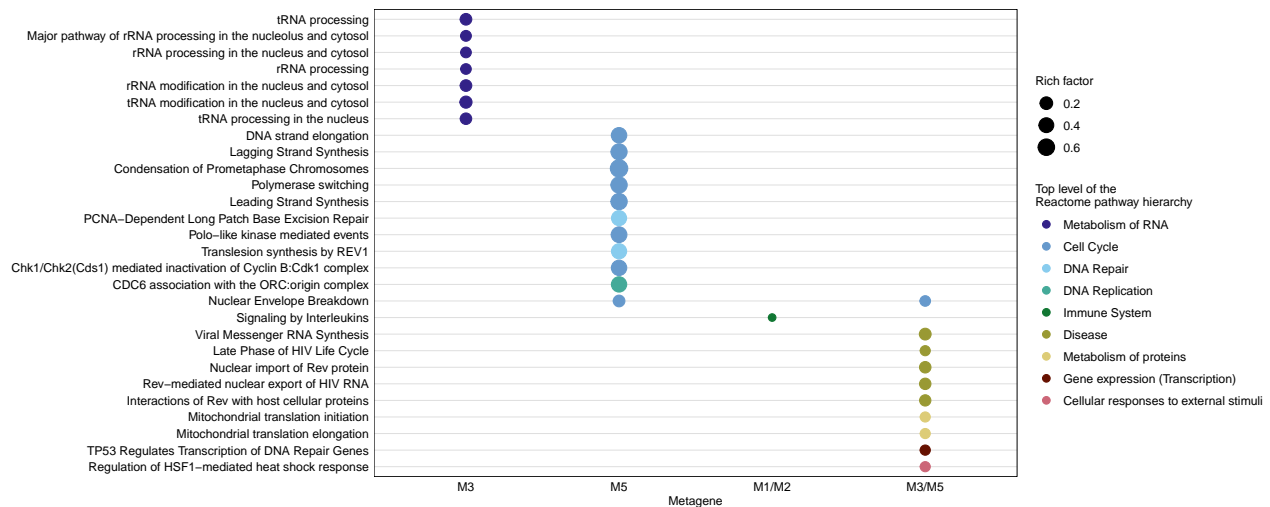
Fig. S21: ***Discovery Set*: Reactome enrichment analysis for genes sets from the consensus gene expression profiles.** Top 10 (sorted by rich factor) significantly enriched Reactome pathways (FDR <0.05) are shown. The dot size indicates the rich factor, which is the number of metagene associated genes in the pathway divided by the number of background genes in that pathway. Colors depicted the uppermost hierarchical level of Reactome database.

5.2 Verification Sets

5.2.1 Temporal profiles and highest ranked genes for metagenes
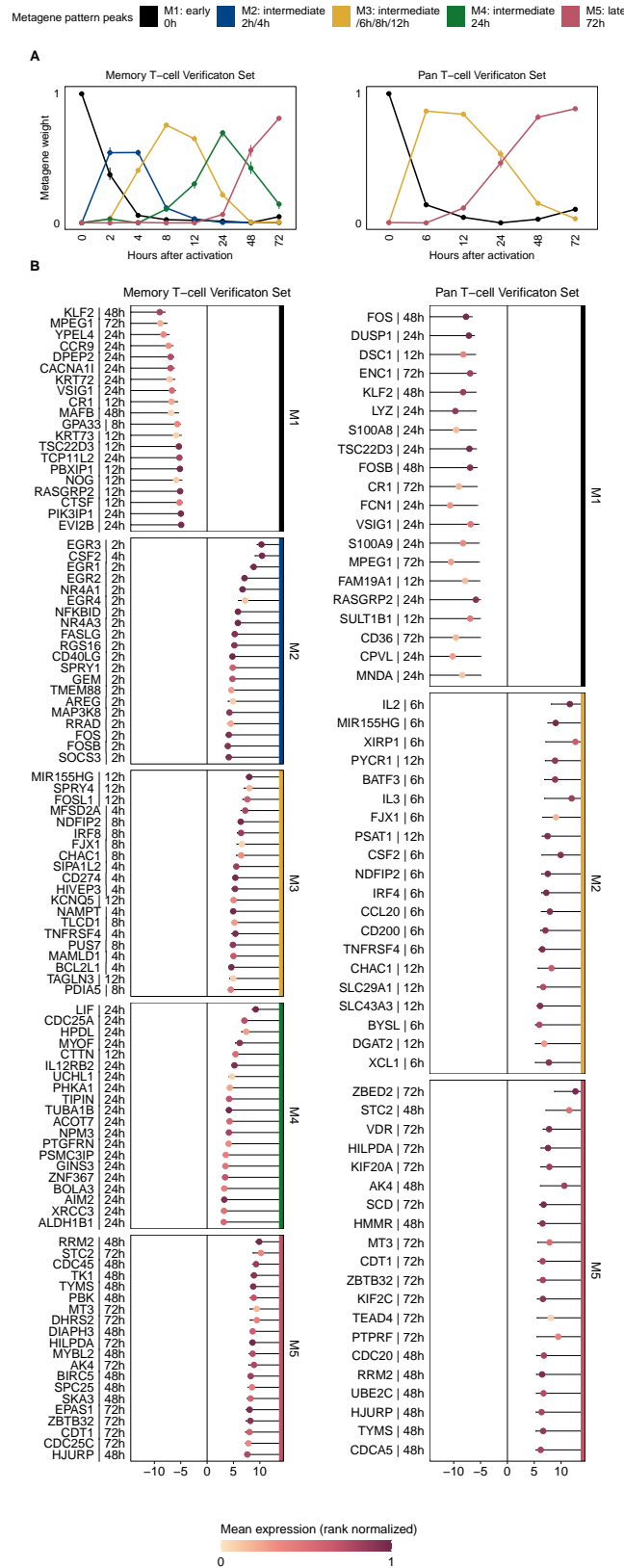
Fig. S22: ***Verification Sets*: Temporal profiles and 25 highest ranked genes for metagenes obtained from NMF. (A)** The pattern matrix for each T-cell population from the *Verification Sets* is shown as continuous profiles, with samples assigned to time points of activation. We scaled each column in the matrix to sum up to one. Dots depict median weights for all samples from identical analysis time points. Vertical lines represent interquartile ranges. We annotated and colored the metagenes based on their maximum median values across all analysis time points. The time point with maximum median value is depicted in the legend. **(B)** Top 25 genes associated with metagenes for each T-cell population. For each T-cell population and gene used for NMF, we used the highest absolute "confect" value estimated in the DGEA across all contrasts (e.g. 12h vs. 0h). Genes are ranked by "confect" values. Dots represent log2 fold changes for contrasts with the highest absolute "confect" value. The time point to the right of the gene represents the contrast with the highest absolute "confect" value. The Color of the dots correspond to rank normalized average expression values of the activation group in the contrast with the highest absolute "confect". The inner end of the horizontal line shows the "confect" value (inner confidence bound).

5.2.2 Summary of consensus gene expression profiles verified for temporal consistency by the *Memory T-cell Verification Set*
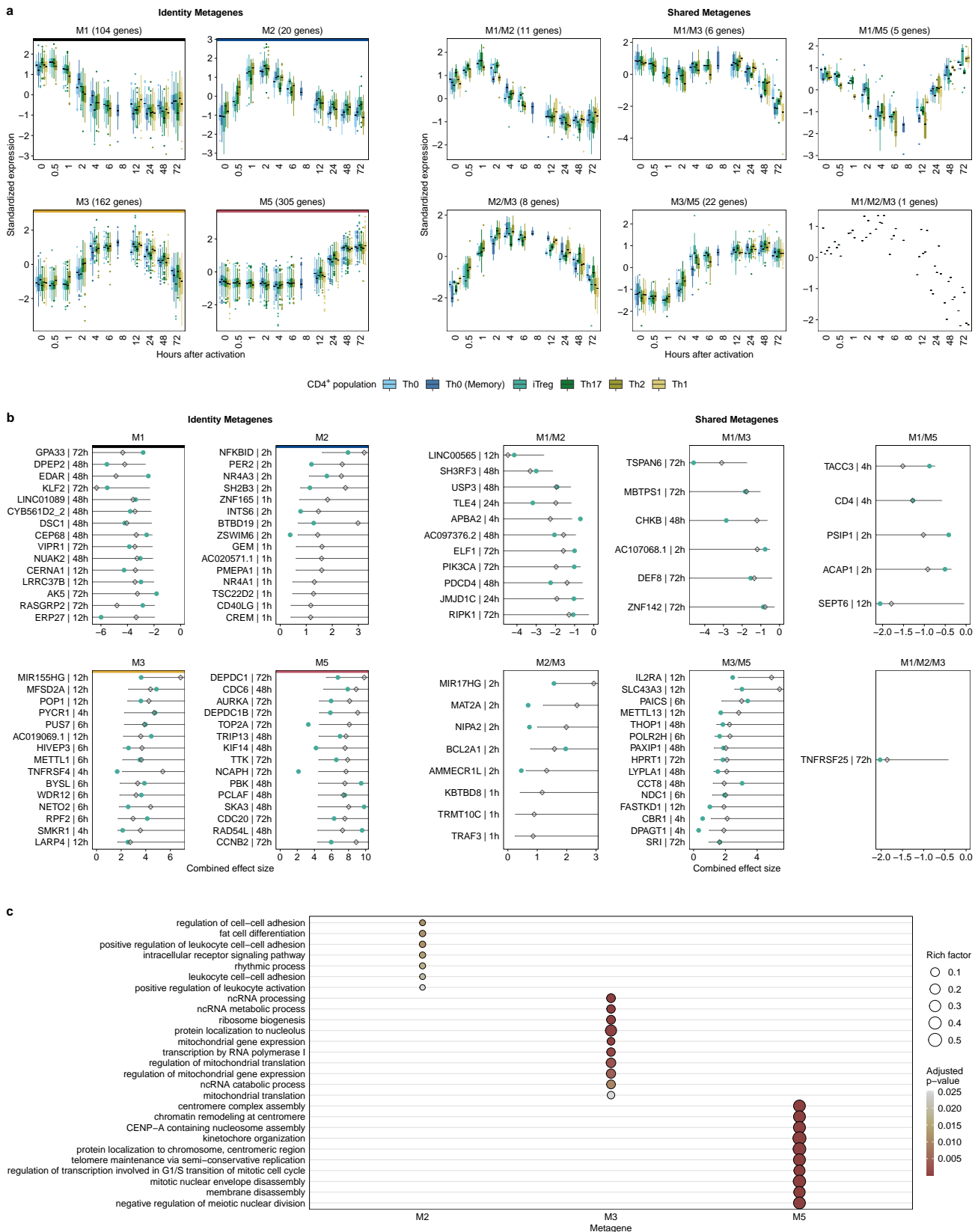
Fig. S23: **Consensus gene expression profiles verified for temporal consistency by the *Memory T-cell Verification Set*. (A)** We grouped the consensus expression profiles over the course by genes associated with identity and shared metagenes. Each boxplot represents one CD4$^+$ T-cell population from the *Discovery Set* and *Memory T-cell Verification Set*. The y-axis depicts standardized median expression of genes from samples with identical analysis time points. The number in parentheses represents the number of genes for the corresponding metagene. **(B)** Top 15 genes associated with identity and shared metagenes. For each gene belonging to the consistent metagenes, we used the highest absolute "confect" value estimated in the meta-analysis. Genes were ranked according to their absolute highest "confect" value. Diamonds represent combined effect size from the meta-analysis for the activation time point of the highest absolute "confect" value. The time point to the right of the gene represents the time point with the highest absolute "confect" value. The inner end the of horizontal line shows the "confect" value (inner confidence bound). The black dots represent the Hedges' g values of the genes from the *Memory T-cell Verification Set*. Since the analysis time point 6 hours is missing in the *Memory T-cell Verification Set*, the mean of the Hedge's g values from 4 hours and 8 hours was formed here. The time points of the largest "confect" value in the *Discovery Set* are not always present in the *Verification Set* and therefore cannot be mapped in the plot. **(C)** The top 10 (sorted by rich factor) significantly enriched GO terms of biological processes (FDR <0.05) of metagene associated genes identified by enrichment analysis. The dot size indicates the rich factor, which is the number of metagene associated genes in the GO term divided by the number of background genes of the term. Colors indicate adjusted p-values of significantly enriched GO terms For enrichment analysis of Reactome pathway see Figure S24.
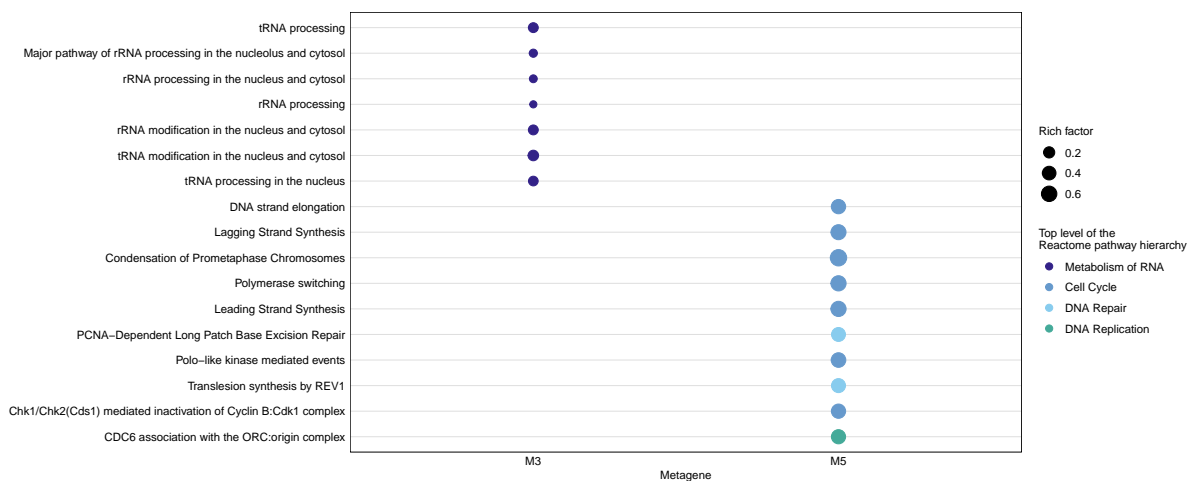
Fig. S24: **Reactome enrichment analysis for genes from the consensus gene expression profiles that passed the *Memory T-cell Verification Set*.** Top 10 (sorted by rich factor) significantly enriched Reactome pathways (FDR <0.05) are shown. The dot size indicates the rich factor, which is the number of metagene associated genes in the pathway divided by the number of background genes in that pathway. Colors depicted the uppermost hierarchical level of Reactome database.

5.2.3 *Pan T-cell Verification Set*: Enrichment Analysis

Fig. S25: **Reactome pathway enrichment analysis for DE genes from the *Pan T-cell Verification Set***. For each contrast (6 to 72 hours of activation vs 0h and 6 to 72 hours without activation vs. 0h), we performed a gene enrichment analysis of DE genes (FDR <0.05). Act.(up/down) = DE genes that were upregulated/downregulated under activated conditions but not DE under unactivated conditions. Neg.ctrl.(up/down) = DE genes that were upregulated/downregulated under unactivated conditions (Genes can also be DE under activated conditions). Shown are the 10 most significant (FDR <0.05) enriched Reactome pathways per group (sorted by rich factor). The dot size indicates the rich factor which is the number of DE genes in the pathway divided by the number of background genes in the pathway. Colors depicted the uppermost hierarchical level of Reactome database.
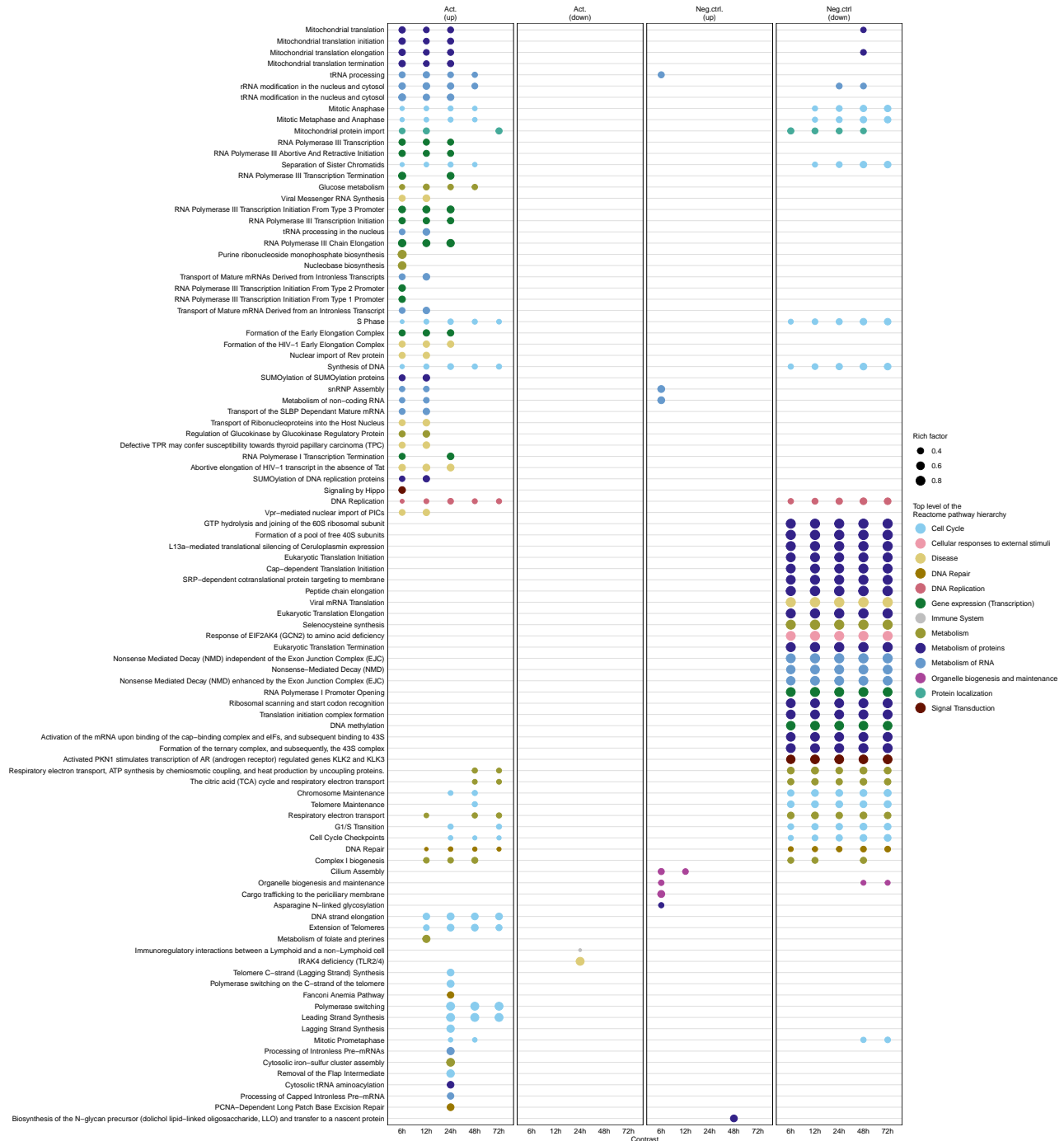
5.2.4 *Pan T-cell Verification Set*: DGEA

Fig. S26: **Hierarchical clustering of DE genes from the *Pan T-cell Verification Set*** in at least one contrast (6 to 72 hours vs. 0 hours) under **(A)** activation conditions, **(B)** unactivated conditions and **(C)** under both activated and unactivated conditions. Euclidean distance and Ward as clustering algorithm was applied to visualize similarity between samples. Each column represents a sample, each row represents a gene. CPM values for each gene were z-score standardized.

5.2.5 *Pan T-cell Verification Set*: Verifying of the consensus gene expression profiles with unactivation kinetics (negative controls)
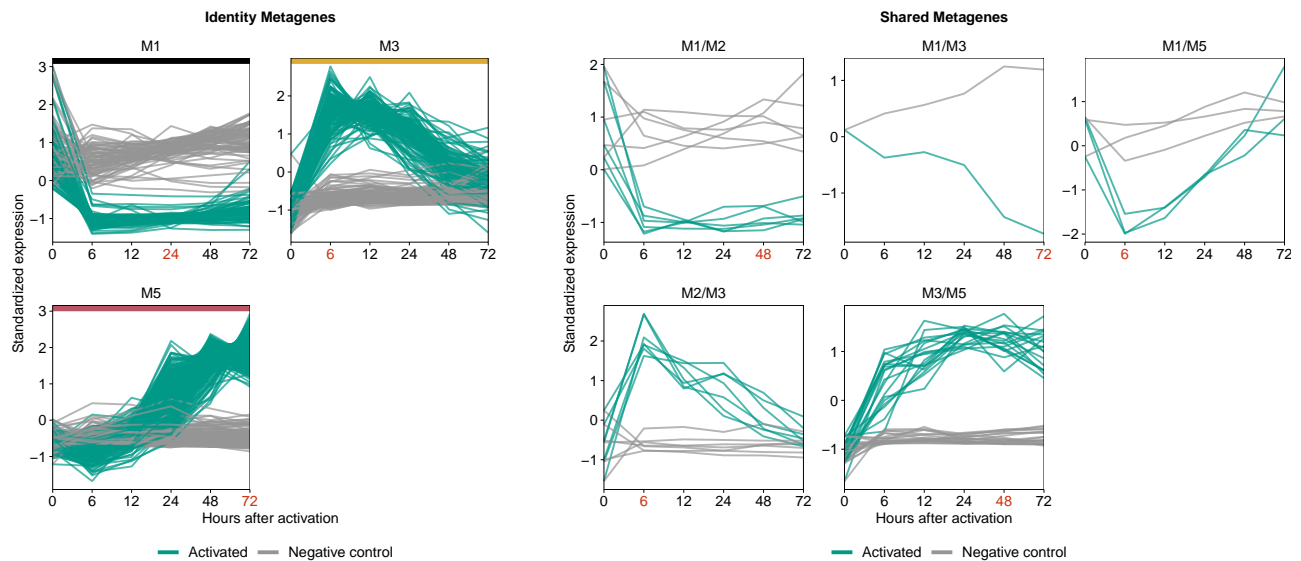

Fig. S27: *Pan T-cell Verification Set*: **Comparison of the expression profiles of the activation kinetics with the kinetics of the negative controls.** For the *Pan T-cell Verification Set*, the temporal expression pattern of genes from the consensus signatures that passed the 2 Verification Sets (activation and negative control kinetics) are shown. CPM values were z-score standardized. The red colored time points indicate the maximum centroid threshold (see main methods). Since only the analysis time points 6 to 72 hours are available, we could not make any conclusion about the time course of expression with respect to intermediate metagene 2 (expression peak after 2 hours). Therefore, we excluded metagene 2 from this analysis.
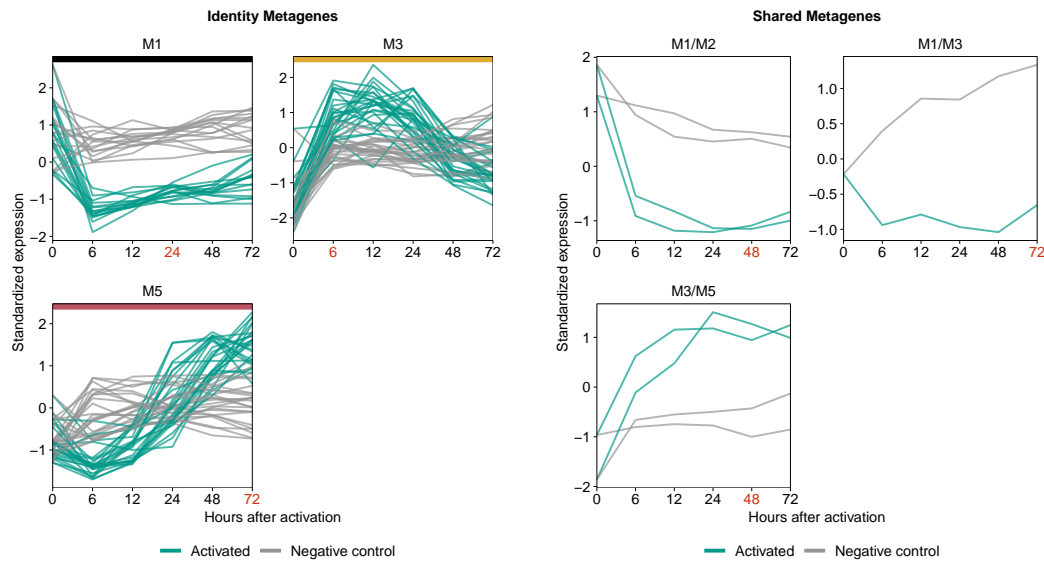
Fig. S28: For the *Pan T-cell Verification Set*, the temporal expression pattern of genes from the consensus signatures that passed the activation kinetics of the 2 *Verification sets* but **not** the kinetics of the negative controls is shown. CPM values were z-score standardized. Identity metagenes were depicted by a colored horizontal line

5.2.6 Summary of consensus gene expression profiles verified for temporal consistency by the *Memory T-cell and Pan T-cell Verification Set*

Fig. S29: **Consensus gene expression profiles verified for temporal consistency by both *Verification Sets* (activation and negative control kinetics).** **(A)** We grouped the consensus expression profiles over the course by genes associated with identity and sh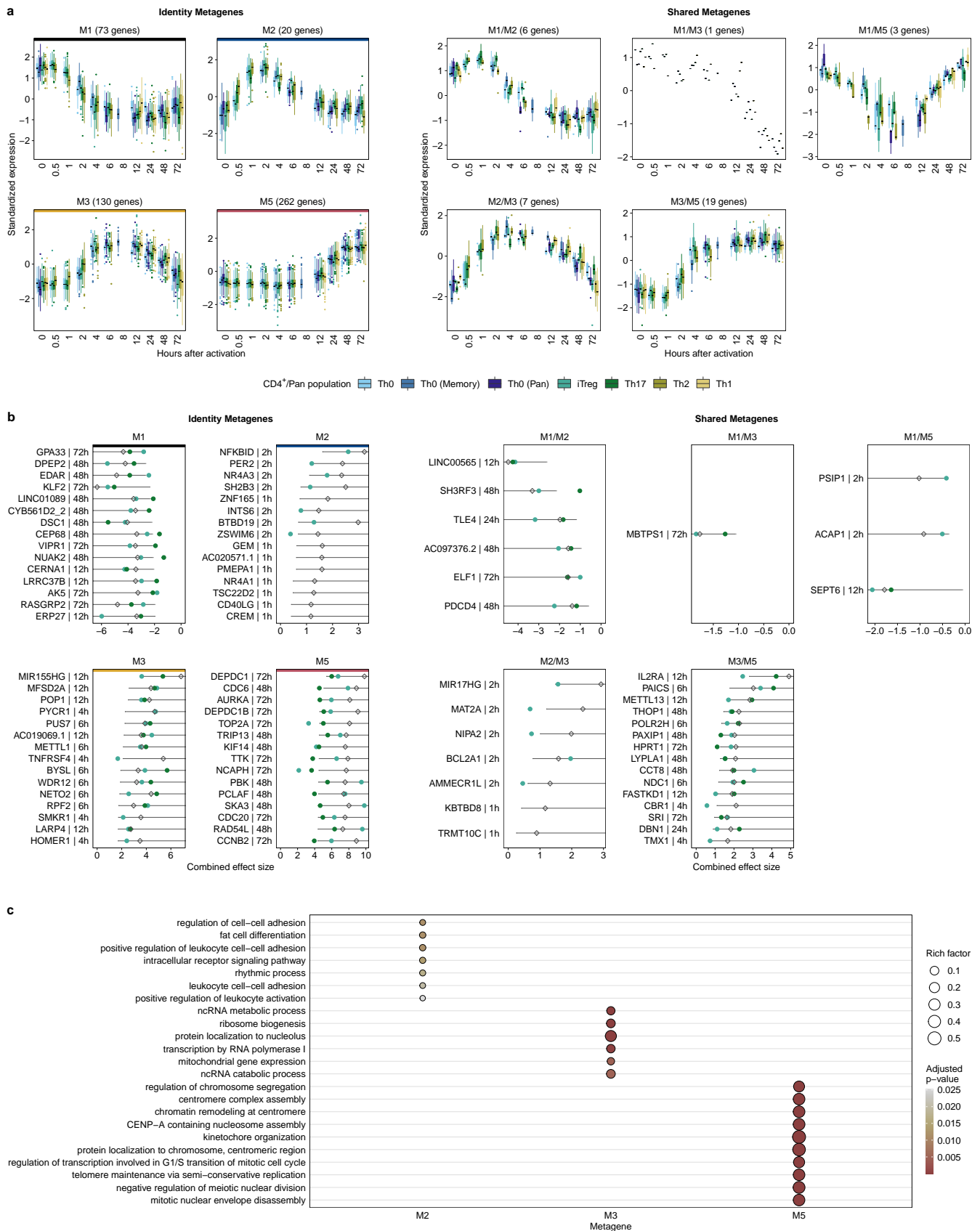ared metagenes. Each boxplot represents one T-cell population from the *Discovery Set*, *Memory T-cell Verification Set* and *Pan T-cell Verification Set*. The y-axis depicted the standardized median expression of genes from samples with identical analysis time points. The number in parentheses represents the number of genes for the corresponding metagene. **(B)** Top 15 genes associated with identity and shared metagenes. For each gene belonging to the consistent metagenes, we used the highest absolute "confect" value estimated in the meta-analysis. Genes were ranked according to their absolute highest "confect" value. Diamonds represent combined effect size from the meta-analysis for the activation time point with the highest absolute "confect" value. The time point to the right of the gene represents the time point with the highest absolute "confect" value. The inner end the of horizontal line shows the "confect" value (inner confidence bound). The green and dark green dots represent the Hedges' g values of the genes from the *Memory T-cell Verification Set* and *Pan T-cell Verification Set*, respectively. Since the analysis time point 6 hours is missing in the *Memory T-cell Verification Set*, the mean of the Hedge's g values from 4 hours and 8 hours was formed here. The time points of the largest "confect" value in the *Discovery Set* are not always present in the *Verification Set* and therefore cannot be mapped in the plot. **(C)** The top 7 (sorted by rich factor) significantly enriched GO terms of biological processes (FDR <0.05) of metagene associated genes identified by enrichment analysis. The dot size indicates the rich factor, which is the number of metagene associated genes in the GO term divided by the number of background genes of the term. Colors indicate adjusted p-values of significantly enriched GO terms For enrichment analysis of Reactome pathway see Figure S30.
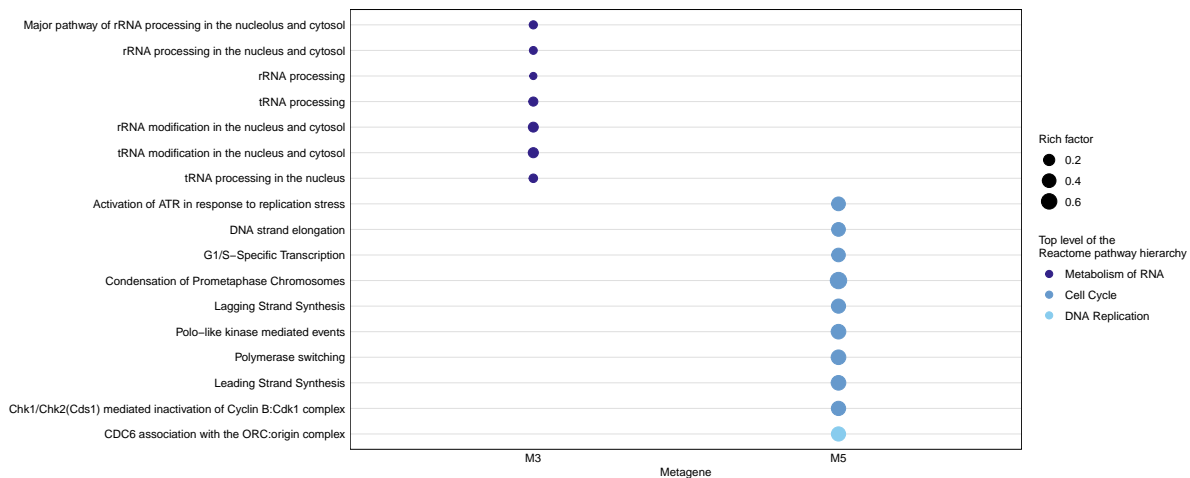
Fig. S30: **Reactome enrichment analysis for genes from the consensus gene expression profiles that passed both *Verification Sets*.** Top 10 (sorted by rich factor) significantly enriched Reactome pathways (FDR <0.05) are shown. The dot size indicates the rich factor, which is the number of metagene associated genes in the pathway divided by the number of background genes in that pathway. Colors depicted the uppermost hierarchical level of Reactome database.

6 Re-analysis of single-cell RNA sequencing data

6.1 Known T-cell state/molecular: Cluster expression
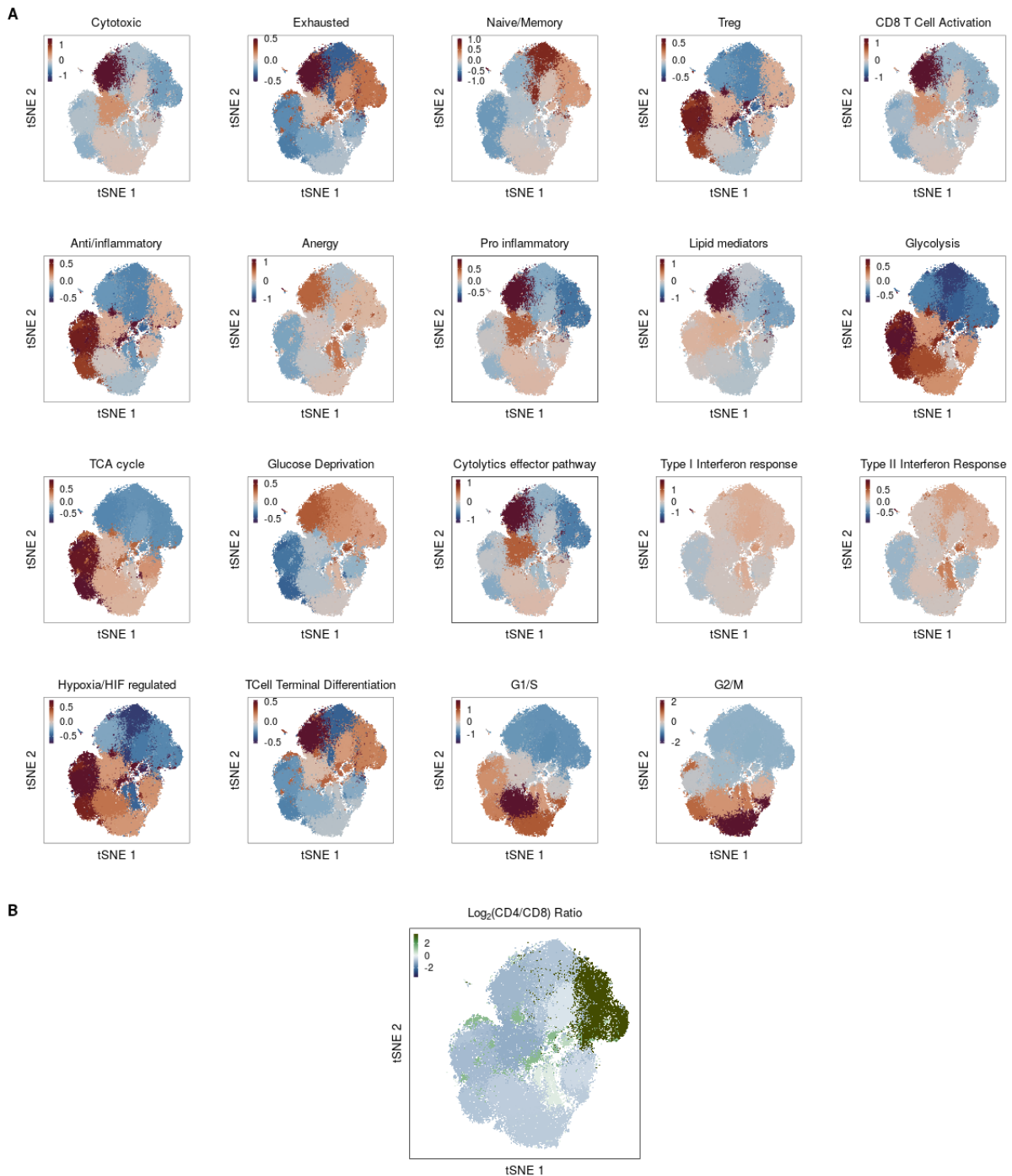

Fig. S31: **T-cell state/molecular mechanisms markers (A)** Standardized average expression of known T-cell state markers as well as markers for T-cell molecular mechanisms (see methods) for each cluster. **(B)** To obtain CD4/CD8 ratios for each cluster, we used the normalized expression values of $CD8^-CD4^+$ and $CD8^+CD4^-$ (average expression of CD8A and CD8B) in each cell. For both CD4 and CD8 we computed the average expression for each cluster. These values were then divided by the average for all cells. The log ratio of these values for CD4 and CD8 was then calculated. All cells in each cluster are labeled with the cluster's log-ratio.

6.2 Known T-cell state/molecular mechanisms' marker: Differences in aggregated expression between patients with low- and high-grade ICANS
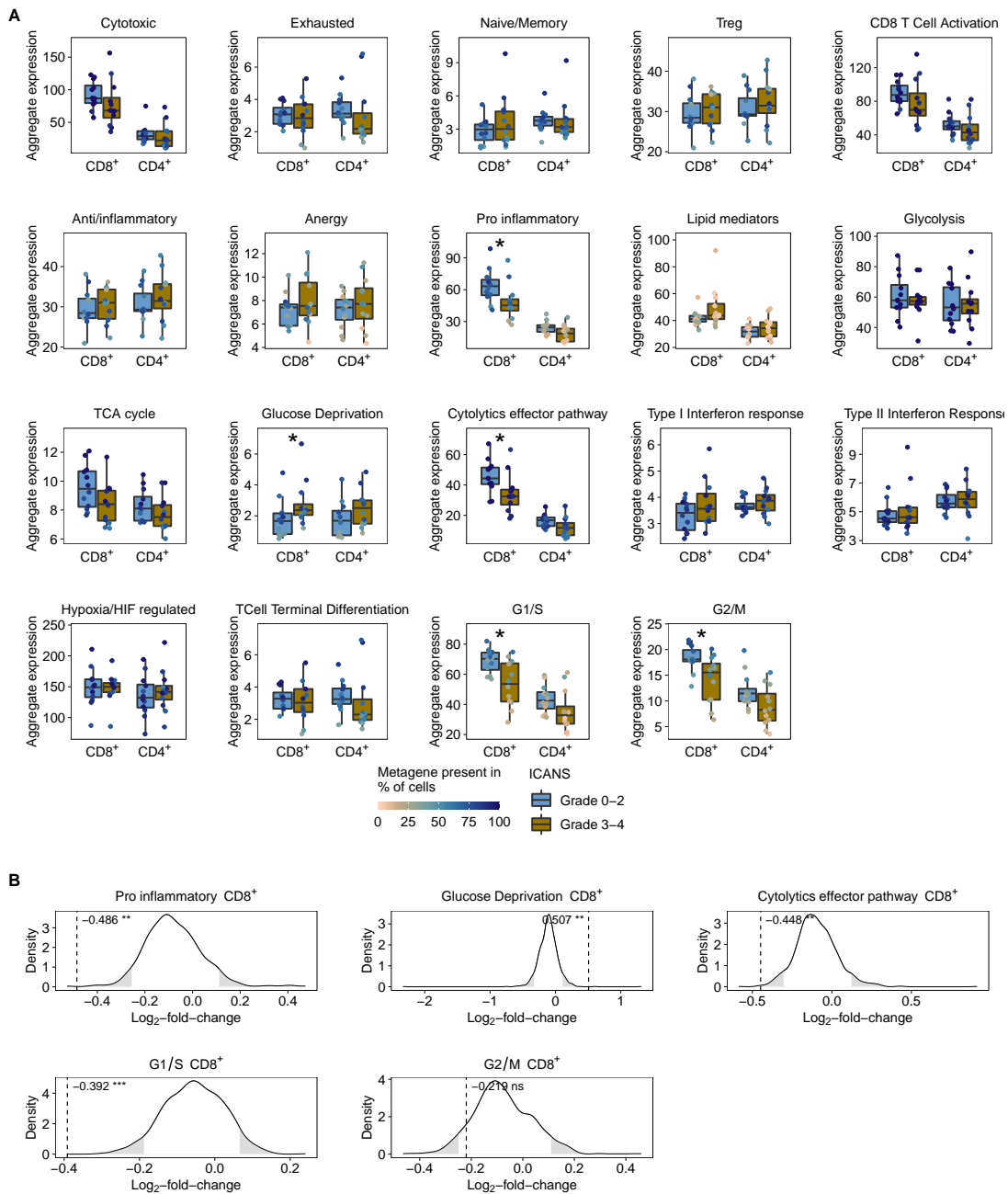

Fig. S32: **Differences in aggregate expression between patients with low- and high-grade ICANS using the Wilcoxon-rank-sum and permutation test for known T-cell state/molecular mechanisms markers.** **(A)** For each T-cell state/molecular mechanisms markers (denoted as gene sets) and patient, we calculated aggregated expression (summed average expression) for $CD8^+CD4^-$ and $CD8^-CD4^+$ cells. Differences in aggregated expression between patients with low- and high-grade ICANS were evaluated using the Wilcoxon rank-sum test. The colors of the dots in the boxplots indicate the percentage of cells for each patient in which the corresponding gene set is present. We considered a gene set as present in a cell if at least 25% of all associated genes had at least one UMI count. **(B)** For each gene set and T-cell population that was significant ($p < 0.05$) in the Wilcoxon rank-sum test, we generated a null distribution in order to confirm the results (see methods). Dashed vertical lines indicate median log2 fold change of aggregated expression between low- and high-grade ICANS patients of the gene set. Rejection regions (empirical p-value < 0.05) are highlighted in grey (* p <0.05, ** p <0.01, *** p <0.001).

6.3 Potential batch effects between aggregated expression and patient characteristics

Metagene	Tcell_Lineage	Sex (p-value)	Lymphoma (p-value)	Age (p-value)
M5	CD4-CD8+	0.264	0.850	0.211
M5	CD4+CD8-	0.417	0.689	0.182
M3	CD4-CD8+	0.192	0.643	0.045
M3	CD4+CD8-	0.172	0.367	0.242

Fig. S33: **Potential batch effects between aggregated expression and patient characteristics**. Only metagens with significant differences (p-value <0.05) in aggregated expression between patients with low-grade and high-grade ICANS were analysed. To determine the correlation between patient characteristics and metagenes/T cell lines, the following tests were performed: For sex, a Wilcoxon rank sum test was applied. The Kruskal-Wallis test was performed for the B-cell lymphoma subtypes (DLBCL=diffuse large B-cell lymphoma, PMBCL=primary mediastinal B-cell lymphoma, tFL= transformed follicular lymphoma). The significance for age was computed with Spearman's rank correlation using the `cor.test` R package.

6.4 Aggregate expression between patients with low- and high-grade CRS or clinical outcome
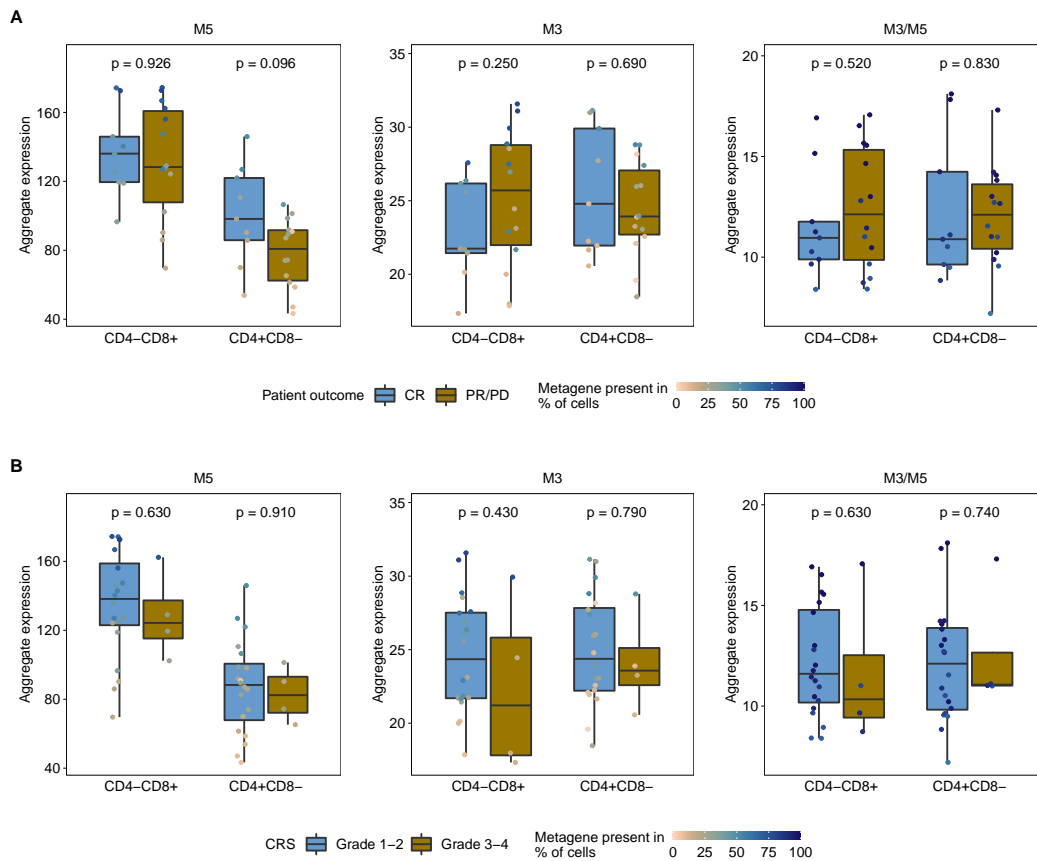

Fig. S34: **Test for differences in aggregate expression between patients with low- and high-grade CRS or clinical outcome (CR compared to PR/PD).** For each metagene and patient, we calculated aggregated expression (summed average expression) for $CD8^+CD4^-$ and $CD8^-CD4^+$ cells. Differences in aggregated expression were compared between patients with **(A)** complete response (CR) and partial response (PR)/progressive disease (PD), **(B)** low- and high-grade Cytokine Release Syndrome (CRS). Significance were calculated using the Wilcoxon rank sum test.