DiSCERN - Deep Single Cell Expression ReconstructioN for improved cell clustering and cell subtype and state detection.
- Supplementary data -

Fabian Hausmann[a,b,1], Can Ergen-Behr[a,1], Robin Khatri[a,b], Mohamed Marouf[a], Sonja Hänzelmann[a,b], Rajasree Menon[f], Matthias Kretzler[f], Nicola Gagliani[c,d,e], Samuel Huber[c,d], Pierre Machart[a,b,*], Stefan Bonn[a,b,*]

[a] *Institute of Medical Systems Biology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.*
[b] *Center for Biomedical AI, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.*
[c] *Section of Molecular Immunology and Gastroenterology, I. Department of Medicine, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany*
[d] *Hamburg Center for Translational Immunology (HCTI), University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany*
[e] *Department of General, Visceral and Thoracic Surgery, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany*
[f] *Department of Computational Medicine and Bioinformatics, University of Michigan, 500 S. State Street, Ann Arbor, 48109, Michigan, USA*

[*] Corresponding authors
Email addresses: pierre.machart@neclab.eu (Pierre Machart), sbonn@uke.de (Stefan Bonn)
[1] Authors contributed equally

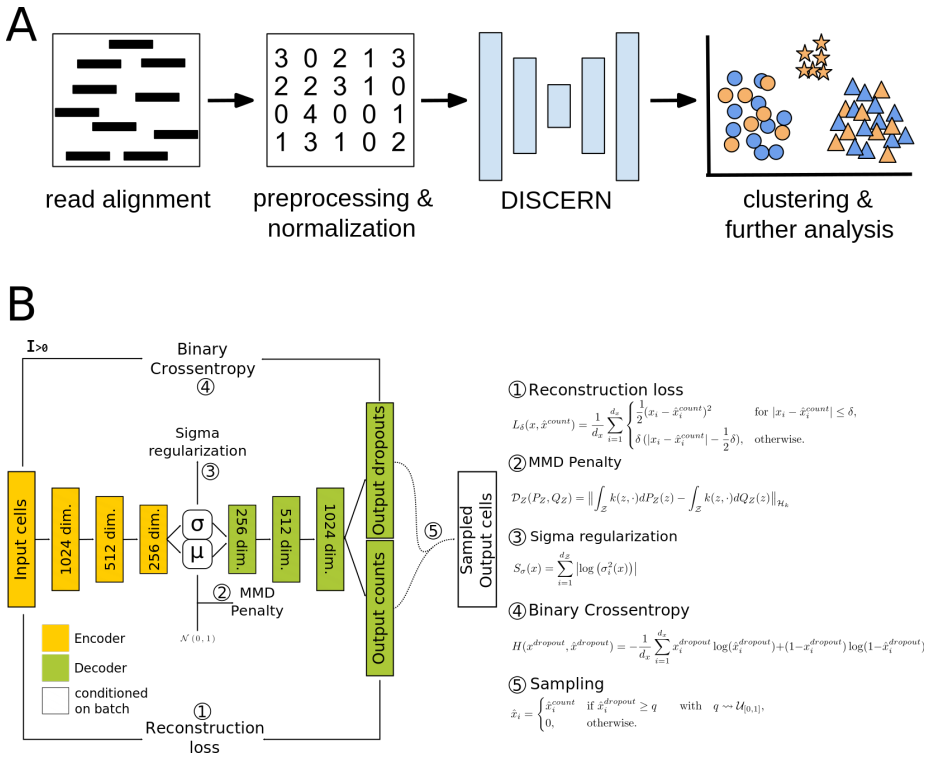


Figure S1: *DISCERN workflow and method details.* **A**: A standard scRNA-seq workflow starts by aligning the sequencing reads to a reference transcriptome to obtain a cell-by-gene count table, which is subsequently preprocessed, filtered and normalized. The normalized count matrices are then used for downstream analyses, such as clustering, differential expression between clusters, and marker gene identification. When combining multiple datasets an alignment or batch correction step is commonly performed to reduce differences between the datasets. DISCERN is used after the preprocessing and normalization steps to integrate the high quality and low quality datasets, reconstructing the gene counts of the low quality to that of the high quality dataset (or vice versa). DISCERN is able to correct for batch effects, provides a lower dimensional representation, and a corrected expression matrix. This corrected expression matrix can directly be used for downstream analysis or used with clustering algorithms. **B**: Overview of the DISCERN neural network architecture consisting of a random encoder (yellow) and a deterministic decoder (green) which can be conditioned on the batch information. DISCERN's loss function contains a (1) count fitting reconstruction loss, (2) a prior fitting MMD-penalty, (3) a sigma regulation term as to prevent the random encoder to collapse to a deterministic one, and (4) a binary cross-entropy term for learning the probability of a dropout event. The final output is generated by sampling from the estimated counts with the estimated dropout probabilities using formula (5).

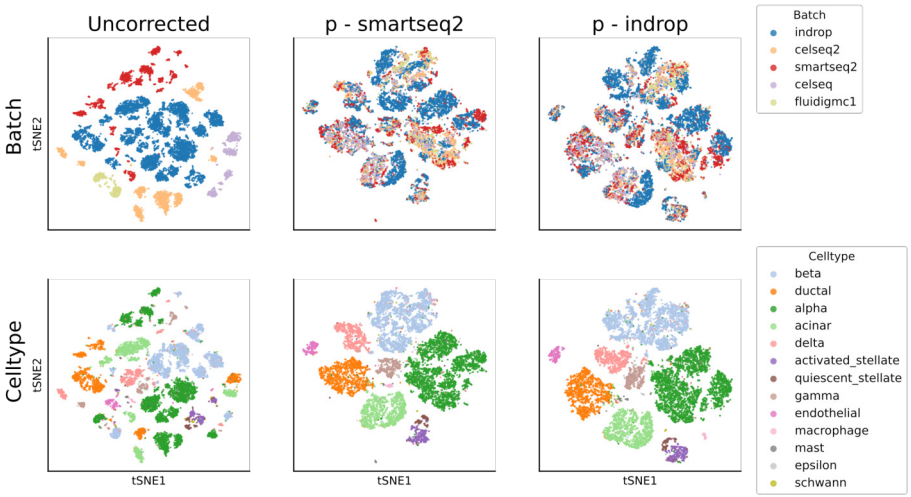


Figure S2: *t-SNE visualization of the pancreas dataset before reconstruction (Uncorrected) and after reconstruction with DISCERN.* The first row is colored by the origin of the dataset (batch) and the second is colored by the cell type annotations. Both batch and cell type annotations were taken from the published dataset. For DISCERN, two projections to the hq smartseq2 batch (second column) and to the lq indrop batch (third column) are shown.

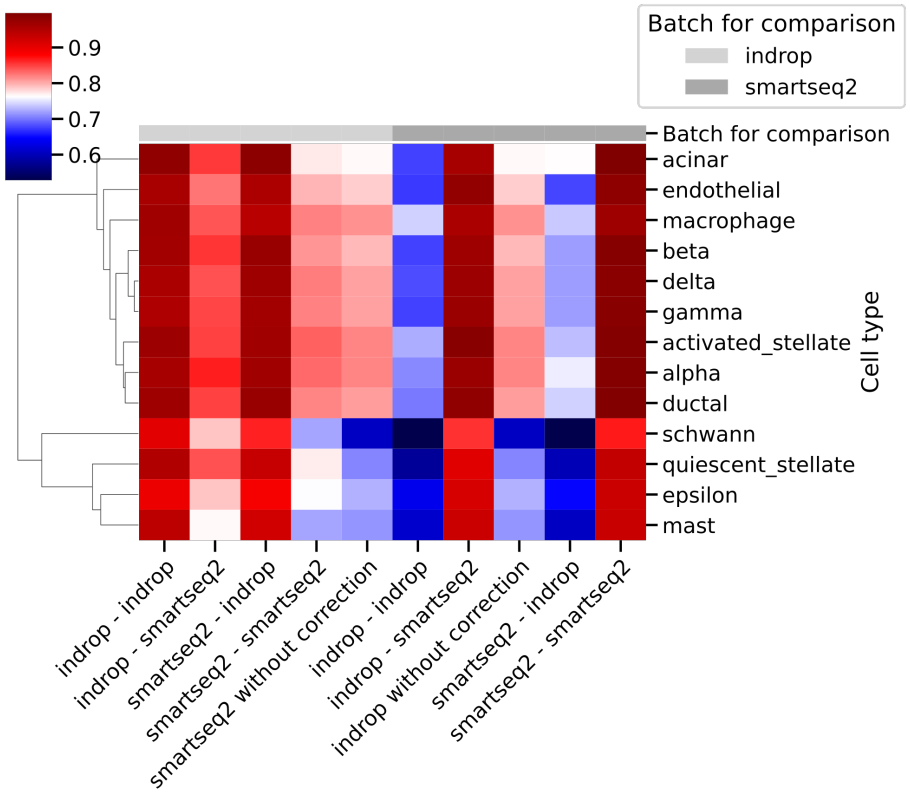


Figure S3: *Heatmap showing the Pearson correlation of the average gene expression per celltype (rows) for the pancreas dataset.* The starting datasets and target dataset for correction are listed on the x-axis. The second entry, for instance, signifies that an indrop dataset was projected to a smartseq2 dataset using DISCERN's expression reconstruction. The correlation is computed between the batch shown in the top row (light gray = indrop, dark gray = smartseq2) and the expression-reconstructed data as listed on the x-axis.

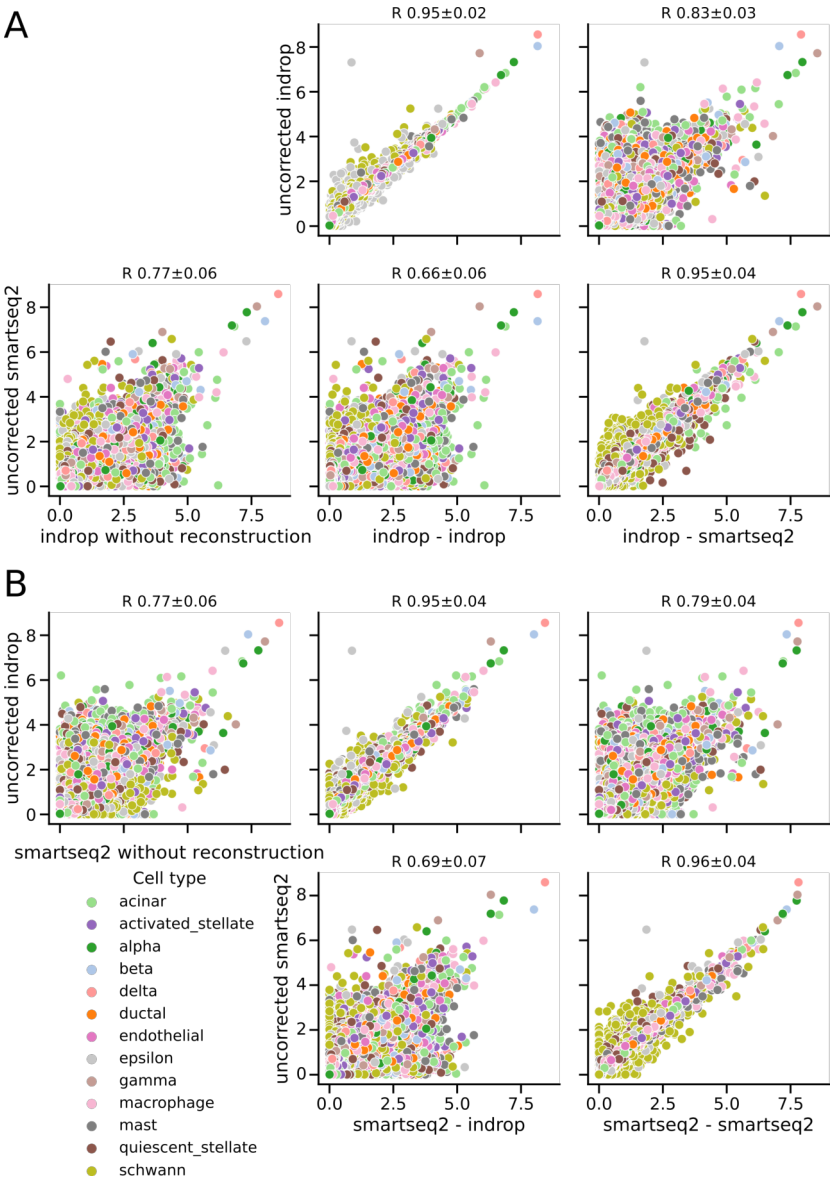


Figure S4: *Average gene expression of the pancreas dataset.* The figures are organized in three columns extended over A and B indicating before DISCERN reconstruction (first column) and after reconstruction using DISCERN (second and third column) stratified by cell type (color). The average expression is compared to the average gene expression of only the indrop (upper row) or the smartseq2 data (lower row). The dataset that is used for projection with DISCERN is shown at the x-axis of each plot after "-", e. g. "smartseq2 - indrop" means smartseq2 projected to indrop. Each colored dot represents one gene. The mean Pearson correlation with one standard deviation over all cell types is displayed in the figure title. **A:** Reconstruction of the indrop batch. **B:** Reconstruction of the smartseq2 batch.

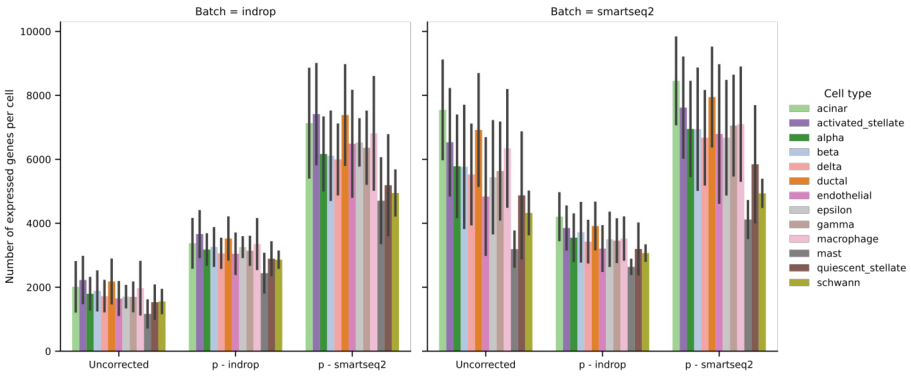


Figure S5: *Number of expressed genes in the indrop (left panel) and the smartseq2 data (right panel) of the pancreas dataset before (Uncorrected) and after projection using DISCERN stratified by cell type (color).* In the right panel, p-indrop displays the gene expression per cell after smartseq2 data was projected to indrop data using DISCERN. Bar heights indicate the average number of expressed genes per cell type and batch. Error bars indicate one standard deviation of the mean over cells in the corresponding batch and cell type.

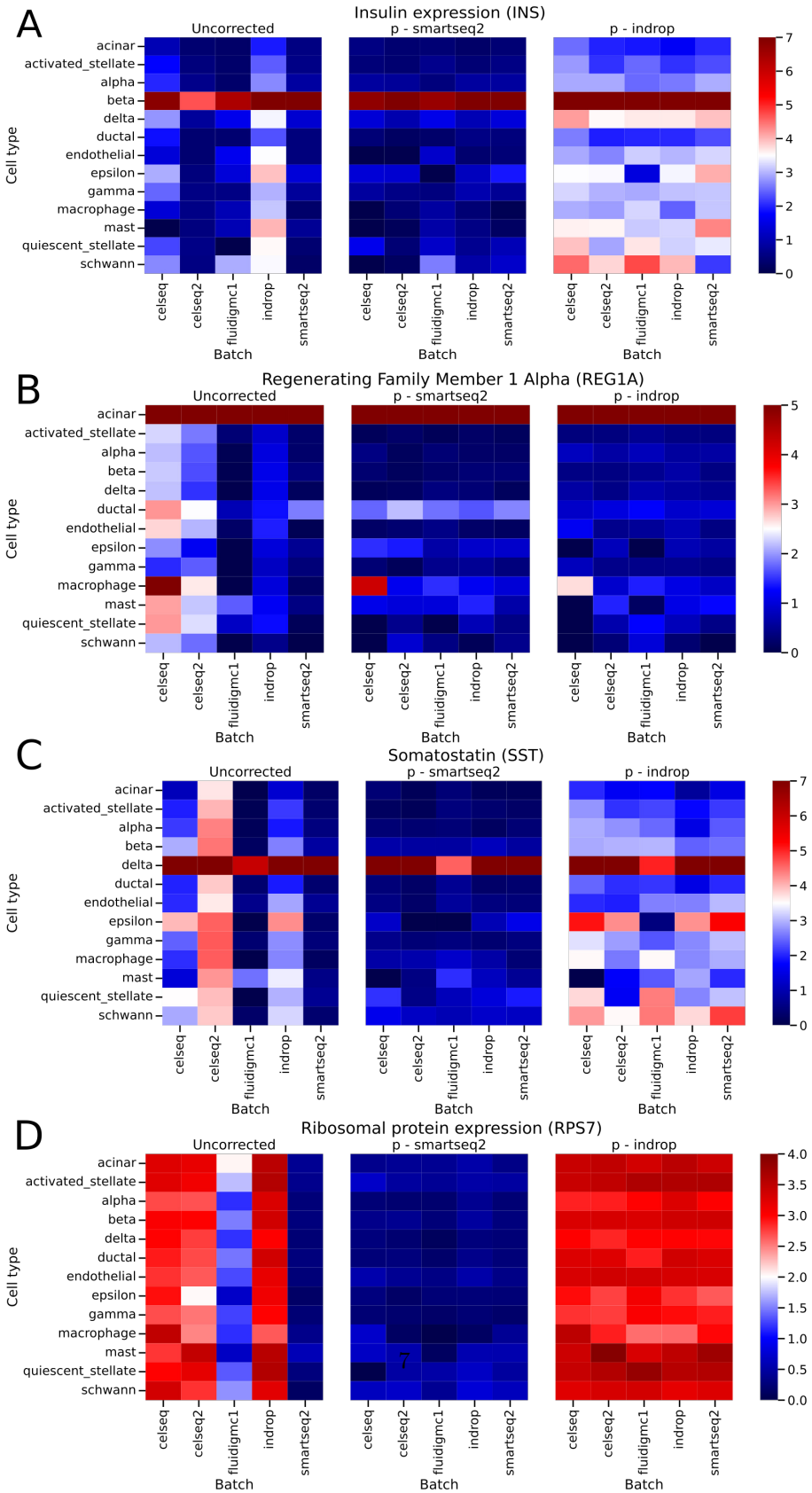

A
Insulin expression (INS)
Uncorrected
p - smartseq2
p - indrop
acinar
activated_stellate
alpha
beta
delta
ductal
endothelial
epsilon
gamma
macrophage
mast
quiescent_stellate
schwann
Cell type
celseq
celseq2
fluidigmc1
indrop
smartseq2
Batch
7
6
5
4
3
2
1
0
B
Regenerating Family Member 1 Alpha (REG1A)
Uncorrected
p - smartseq2
p - indrop
Cell type
Batch
5
4
3
2
1
0
C
Somatostatin (SST)
Uncorrected
p - smartseq2
p - indrop
Cell type
Batch
7
6
5
4
3
2
1
0
D
Ribosomal protein expression (RPS7)
Uncorrected
p - smartseq2
p - indrop
Cell type
Batch
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

Figure S6: *Average gene expression of Insulin (INS), a ribosomal gene (RPS7), REG1A, and Somatostatin (SST) by cell type (rows) and by batch (columns) in the pancreas dataset.* The first column shows the uncorrected datasets, while the second and third column show projections using DISCERN to the smartseq2 and the indrop dataset, respectively. **A:** *INS* was selected because it is a cell type-determining gene for beta cells and *RPS7* is known to be expressed in nearly all cells. While nearly all batches display exclusive *INS* expression in beta cells in uncorrected data, the indrop data shows a more dispersed expression of *INS* in several cell types. Projection to the smartseq2 batch results in a beta cell-specific expression in the corrected indrop data (second column). Projection to the indrop batch results in dispersed INS expression for all batches (third column). **B:** *REG1A* is a acinar cell specific gene, shown to be involved in acinar cell carcinoma [1]. For most pancreatic datasets, it is exclusively expressed in acinar cells in the uncorrected data. Only celseq shows a more dispersed expression across several cell types. After reconstruction to indtop or smartseq2 data the expression of *REG1A* is restricted to acinar cells and macrophages in the celseq batch. **C:** *SST* is known to be produced by delta cells in the pancreas [2], which can be observed for instance in the smartseq2 batch. After reconstruction to the smartseq2 batch delta cell-specific expression of *SST* is observed for all datasets. **D:** *RPS7* shows high expression in the indrop, celseq and the celseq2 batch, whereas smartseq2 and fluidigmc1 show low to no expression, as described previously [3]. This expression of *RPS7* can be removed by projecting to smartseq2 or reconstructed by projection to indrop data.

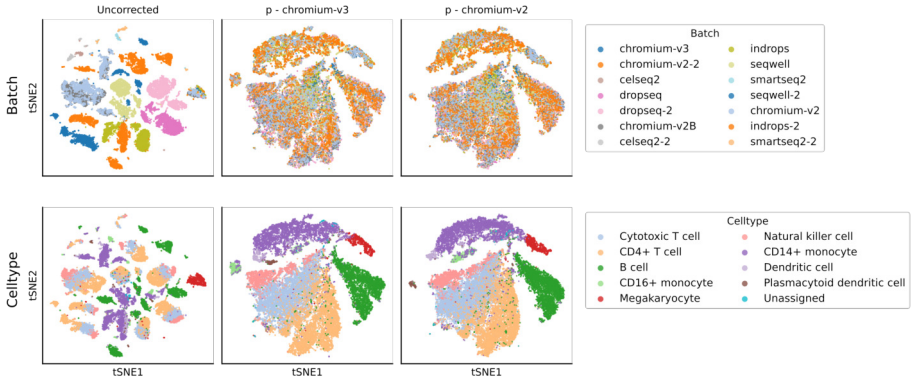


Figure S7: *t-SNE visualization of the difftec dataset before reconstruction (Uncorrected) and after reconstruction with DISCERN.* The first row shows the dataset of origin (batch) and the second row shows the cell type annotations which are available together with the dataset. For DISCERN two projections, one to the hq chromium-v3 batch and one to the lq chromium-v2 batch is shown.

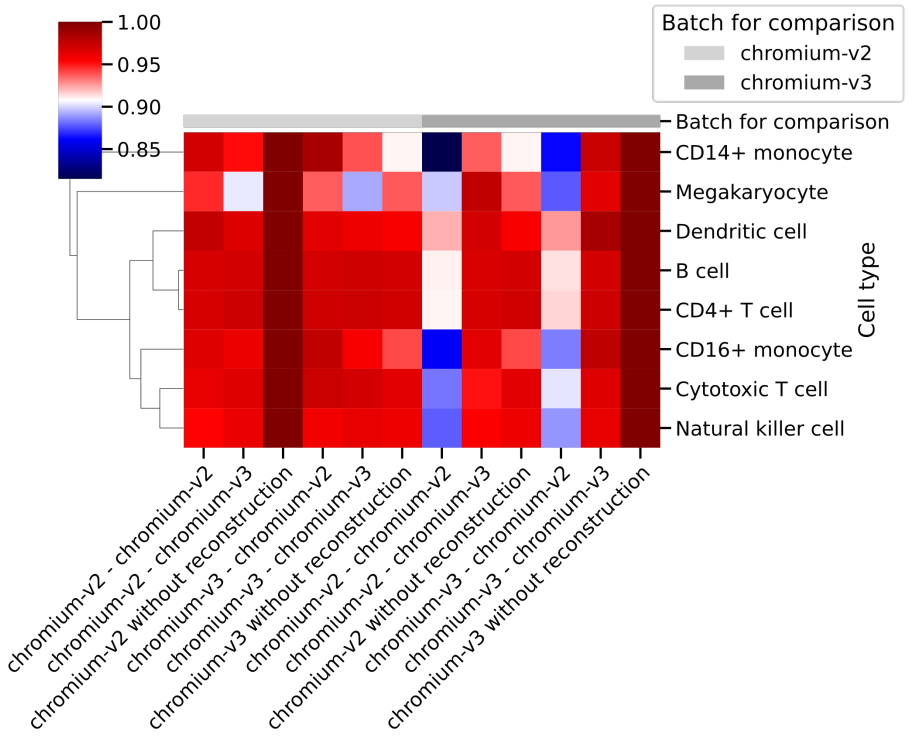


Figure S8: *Heatmap showing the Pearson correlation of the average gene expression per celltype (rows) for the difftec dataset.* The starting datasets and target dataset for correction are listed on the x-axis. The second entry, for instance, signifies that a chromium-v2 dataset was projected to a chromium-v3 batch using DISCERN's expression reconstruction. The correlation is computed between the batch shown in the top row (light gray = chromium-v2, dark gray = chromium-v3) and the expression-reconstructed data as listed on the x-axis.

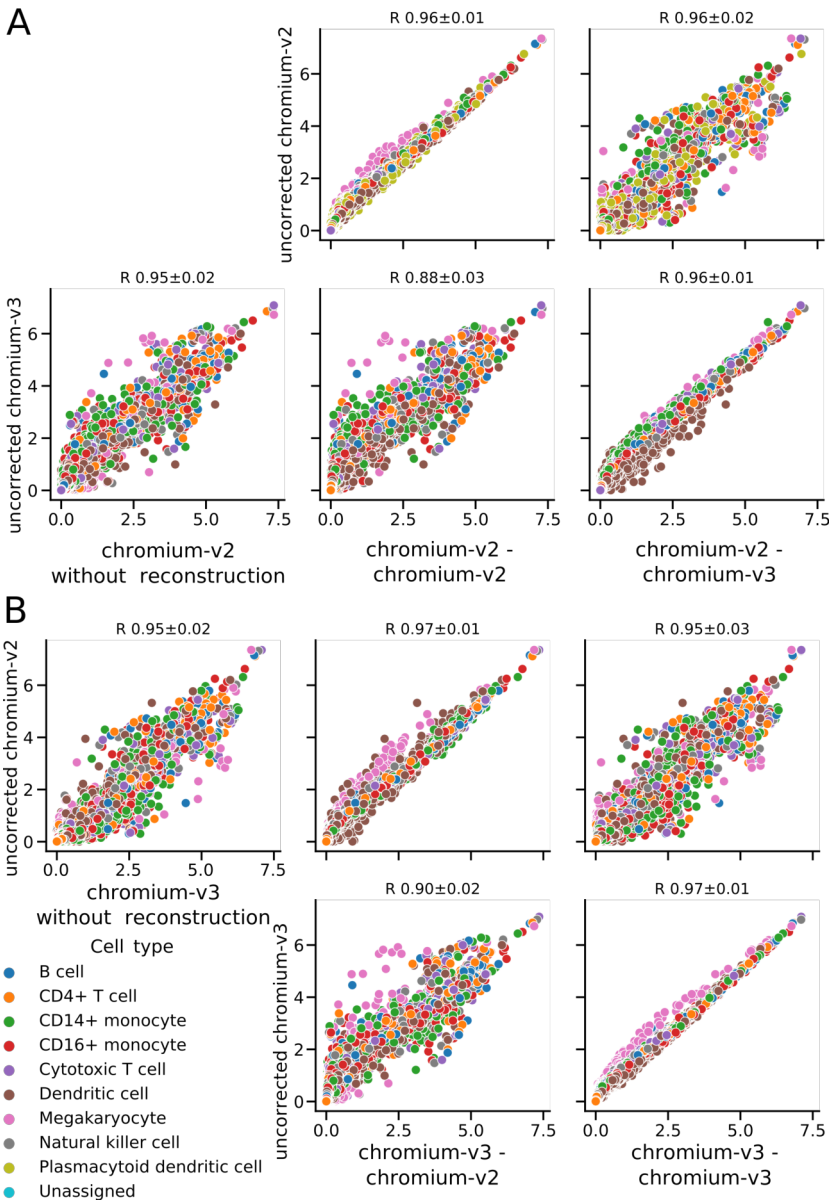


Figure S9: Average gene expression of the difftec dataset. The figures are organized in three columns extended over A and B indicating before DISCERN reconstruction (first column) and after reconstruction using DISCERN (second and third column) stratified by cell type (color). The average expression is compared to the average gene expression of only the chromium-v2 (upper row) or the chromium-v3 data (lower row). The dataset that is used for projection with DISCERN is shown at the x-axis of each plot after "-", e. g. "chromium-v2 - chromium-v3" signifies chromium-v2 data was projected to the chromium-v3 batch. Each colored dot represents one gene. Colors indicate the cell type identity. The mean Pearson correlation with one standard deviation over all cell types is displayed in the figure title. **A:** Reconstruction of the chromium-v2 batch. **B:** Reconstruction of the chromium-v3 batch.

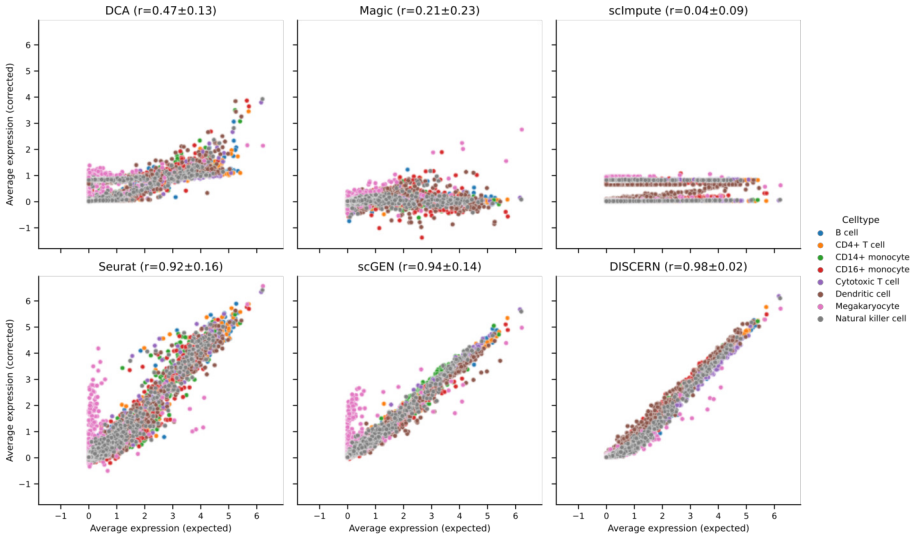


Figure S10: *Comparison of the average gene expression reconstruction performance for several methods for the difftec dataset.* Three imputation (DCA, Magic, scImpute), two batch correction methods (Seurat, scGEN), and DISCERN are compared. The dataset is based on the difftec dataset where the chromium-v3 batch was split into chromium-v3-lq and chromium-v3-hq and selected genes were removed (in silico gene drop out) from chromium-v3-lq. The corrected average gene expression (y-axis) is based on the reconstructed or imputed chromium-v3-lq data. For DISCERN and scGEN the projection onto the chromium-v3-hq reference is depicted. The expected average gene expression (x-axis) is based on the unmodified chromium-v3-lq batch. Mean Pearson correlation with one standard deviation over all cell types is displayed in parentheses of the figure title. Colors indicate the cell type identity.

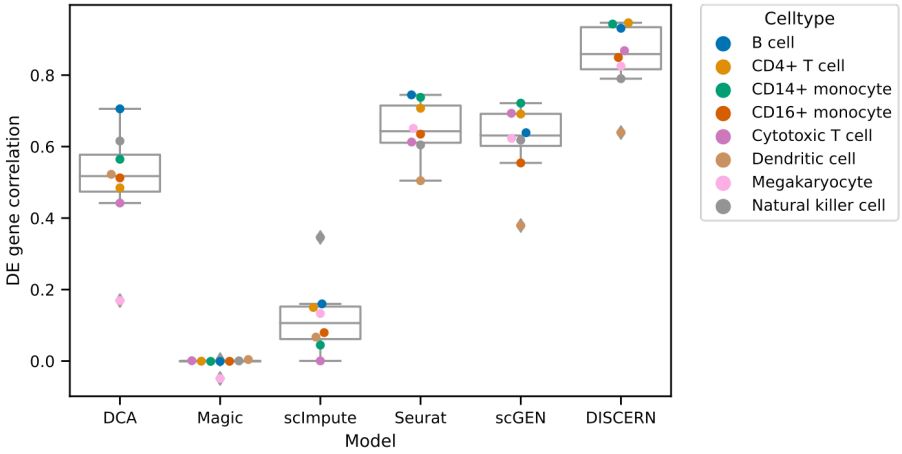


Figure S11: *Pearson correlation of DEG t-statistics for a one-vs-rest cell type comparison and in silico gene removal.* The dataset is based on the difftec dataset where the chromium-v3 batch was split into chromium-v3-lq and chromium-v3-hq and selected genes were removed from chromium-v3-lq data. The corrected average gene expression is based on reconstructed or imputed chromium-v3-lq only, while the expected average gene expression is based on the unmodified chromium-v3-lq batch. For DISCERN and scGEN the projection to chromium-v3-hq is shown. Boxplots represent median, quantiles, minimum, maximum, and potential outliers. Colors indicate the cell type identity.

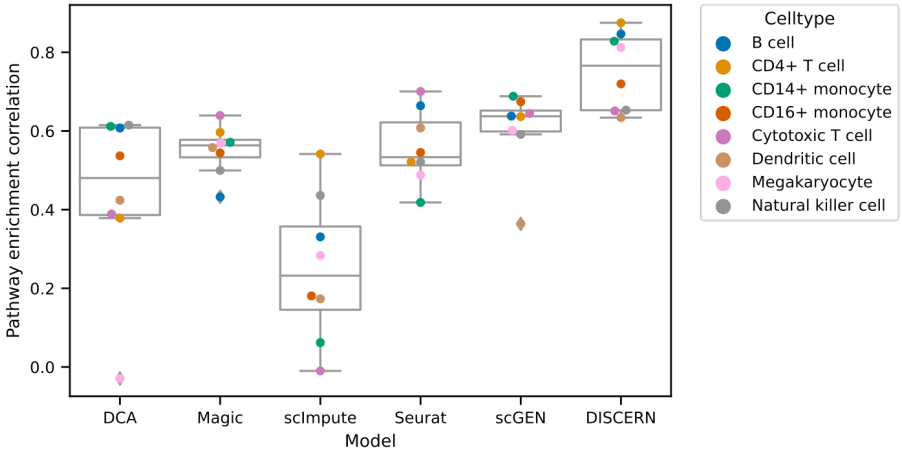


Figure S12: *Pearson correlation of KEGG gene set enrichment scores for a one-vs-rest cell type comparison and in silico gene removal.* The dataset is based on the difftec dataset where the chromium-v3 batch was split into chromium-v3-lq and chromium-v3-hq and selected genes were removed from chromium-v3-lq data. Instead of directly measuring DEG correlation as in Figure S11 a gene set enrichment analysis was performed for DEGs and correlated to the ground-truth 'expected' information. The corrected average gene expression is based on reconstructed or imputed chromium-v3-lq only, while the expected average gene expression is based on the unmodified chromium-v3-lq batch. For DISCERN and scGEN the projection to chromium-v3-hq is shown. Boxplots represent median, quantiles, minimum, maximum, and potential outliers. Colors indicate the cell type identity.

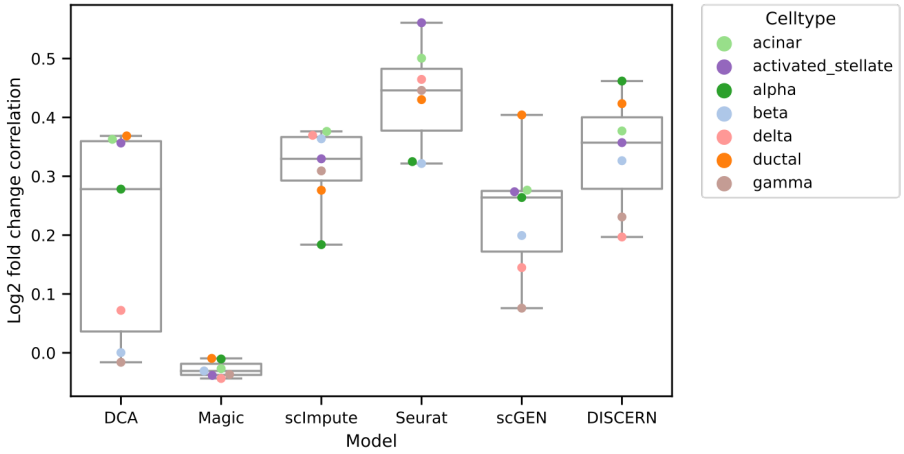


Figure S13: *Pearson correlation of the log2 fold-change (FC) per cell type for the reconstructed-hq and smartseq-hq pancreas data.* For each cell type, DEG and FC were calculated against all other cell types. For DISCERN and scGEN the projection of indrop-lq to smartseq2-hq data is shown, resulting in reconstructed-hq data. Boxplots represent median, quantiles, minimum, maximum, and potential outliers. Colors indicate the cell type identity.

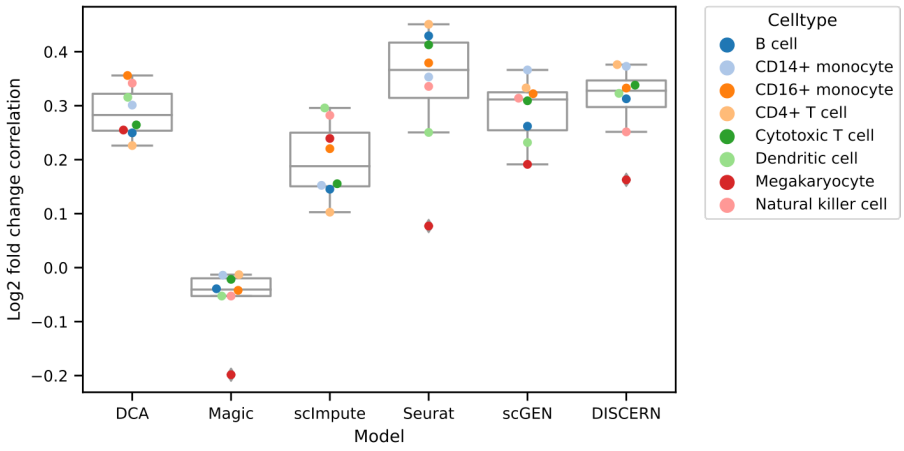


Figure S14: *Pearson correlation of the log2 fold-change (FC) per cell type for the reconstructed-hq and chromium-v3-hq difftec data.* For each cell type, DEG and fold-change were calculated against all other cell types. For DISCERN and scGEN the projection of chromium-v2-lq to chromium-v3-hq data is shown, resulting in reconstructed-hq data. Boxplots represent median, quantiles, minimum, maximum, and potential outliers. Colors indicate the cell type identity.

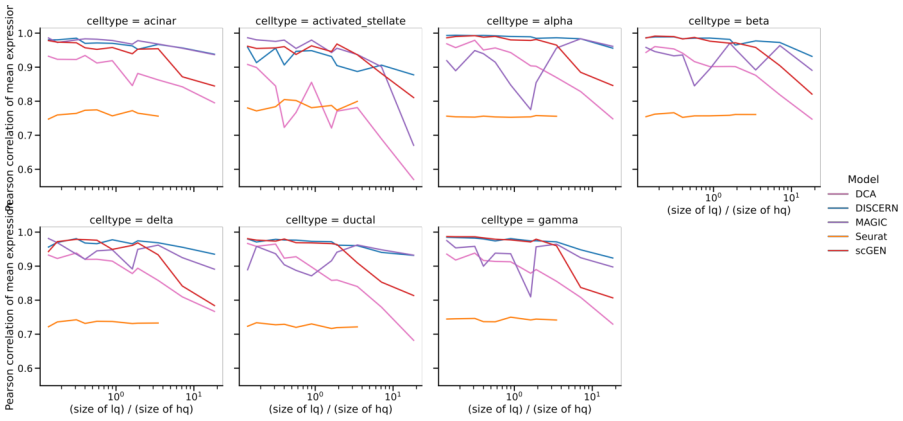


Figure S15: *Pearson correlation of the mean gene expression for the pancreas reconstructed-hq and smartseq2-hq data for different ratios of lq to hq training data.* The plot shows the dependency of the mean gene expression reconstruction on the ratio of lq to hq training data, showing increased performance for lower ratios and a marked decrease in performance for higher ratios, especially for scGen, while DISCERN remains relatively stable for all ratios tested. For DISCERN and scGEN the projection to smartseq2-hq is shown. Colors indicate different methods.

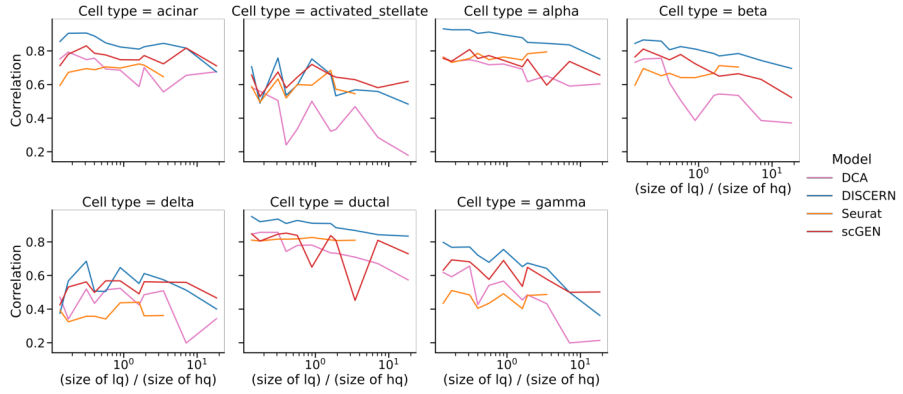


Figure S16: *Pearson correlation of DEG t-statistics for a one-vs-rest cell type comparison and in silico gene removal.* The dataset is based on the pancreas dataset where the smartseq2 batch was split into smartseq2-lq and smartseq2-hq and selected genes were removed from smartseq2-lq. The t-statistic is computed on removed genes after reconstruction of the smartseq2-lq batch to reconstructed-hq data and compared to the t-statistic of the unmodified smartseq2-lq batch. For DISCERN and scGEN the projection to smartseq2-hq is shown. Uncorrected and MAGIC corrected data have close to zero gene expression in the smartseq2-lq for the selected genes and thus cannot be shown. Colors indicate different methods.

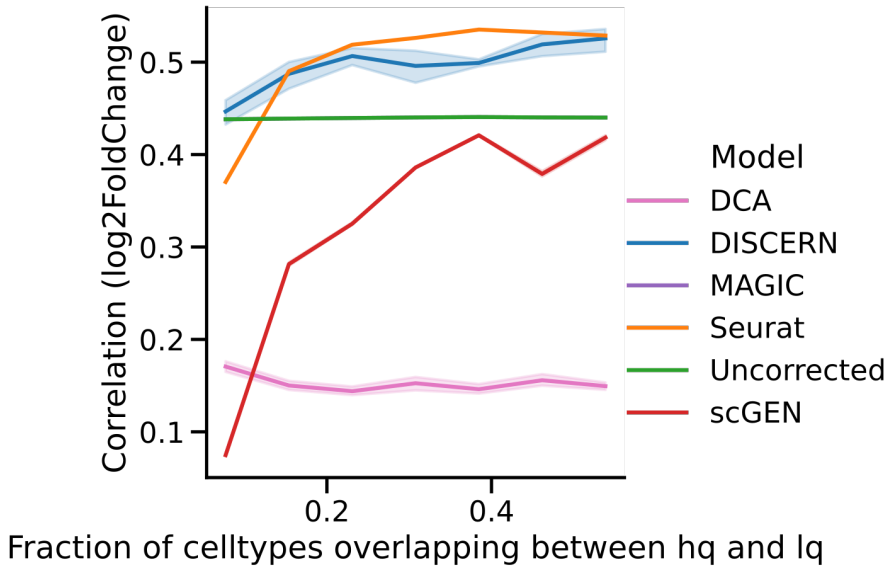


Figure S17: *Spearman correlation of the log2 fold-change (FC) of alpha cells that were reconstructed and ground-truth alpha cells that were excluded from training using pancreas data.* Different fractions of cell type overlap in the indrop-lq and smartseq2-hq training data were used to estimate the reconstruction performance when datasets become dissimilar. Alpha cells were only present in the indrop-lq data and smartseq2-hq alpha cells were extracted as ground truth information. X-axis shows fractions of cell types, which are non-alpha cells and overlap between lq and hq batches. Confidence intervals indicate the standard deviations from five independently trained models. The turquoise line for MAGIC is not visible as the correlation is the same as achieved for uncorrected data. Best performance is observed for DISCERN and Seurat.

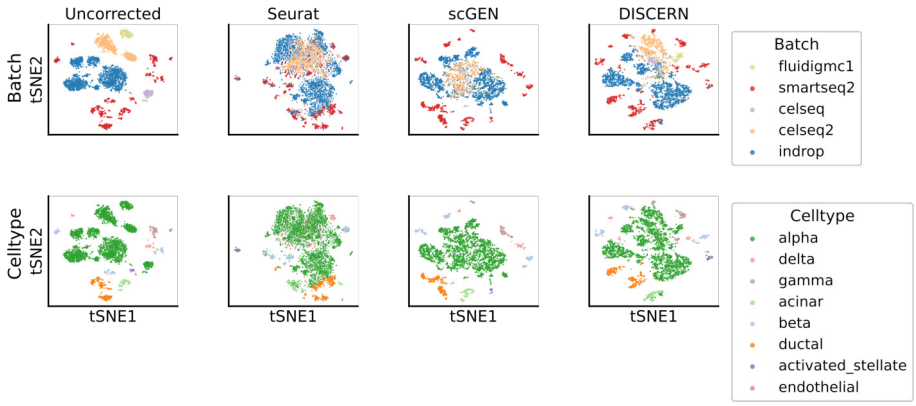


Figure S18: *t-SNE visualization of the pancreas dataset where alpha cells are removed from the smartseq2-hq batch and all cell types except alpha cells are removed from all other batches.* This means that there is no cell type overlap between the smartseq2 and the other batches. The first column shows the data before reconstruction (Uncorrected) and the other columns show the corrected dataset using Seurat (second column), scGEN (third column) and DISCERN (last column). The cells are colored by batch (upper row) or by cell type (lower row). The Seurat corrected dataset shows over integration, e.g. alpha and delta cells are mixed, whereas scGEN and DISCERN bring similar cells closer in t-SNE, but do not fully integrate alpha cells. For DISCERN the dataset was projected to the smartseq2 batch.

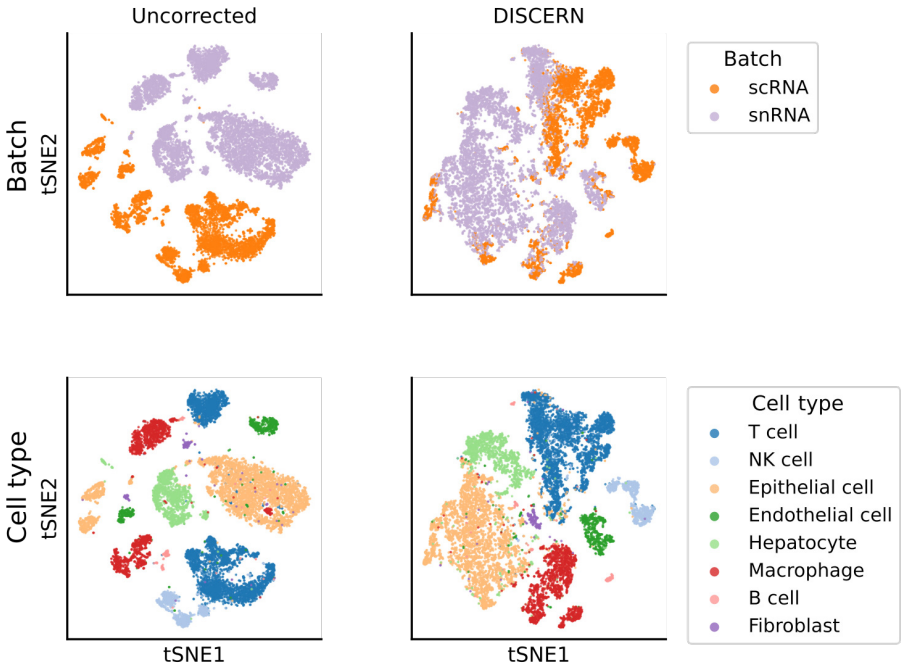


Figure S19: *t-SNE visualization of scRNA-seq and snRNA-seq data before (Uncorrected, first column), after reconstruction with DISCERN (second column) and Seurat (third column).* In this dataset the same sample from a metastatic liver biopsy was sequenced using scRNA-seq and snRNA-seq technology, yielding the sc-hq and sn-lq datasets. The sn-lq data was reconstructed using the sc-hq reference. The first row shows the color annotation by batch and the second row is colored by the different cell types found in the dataset. Annotation of the cell types was provided with the original data.

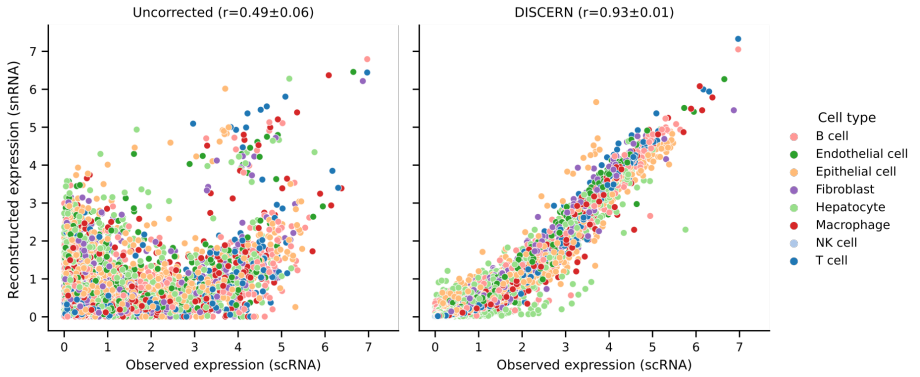


Figure S20: *Average gene expression of scRNA-seq and snRNA-seq data before (Uncorrected) and after reconstruction with DISCERN and Seurat.* In this dataset the same sample from a metastatic liver biopsy was sequenced using scRNA-seq and snRNA-seq technology. The sn-lq data was reconstructed using the sc-hq reference to yield reconstructed-hq data. Each colored dot represents one gene. Colors indicate the cell type identity. The mean Pearson correlation with one standard deviation over all cell types is displayed in the figure title.

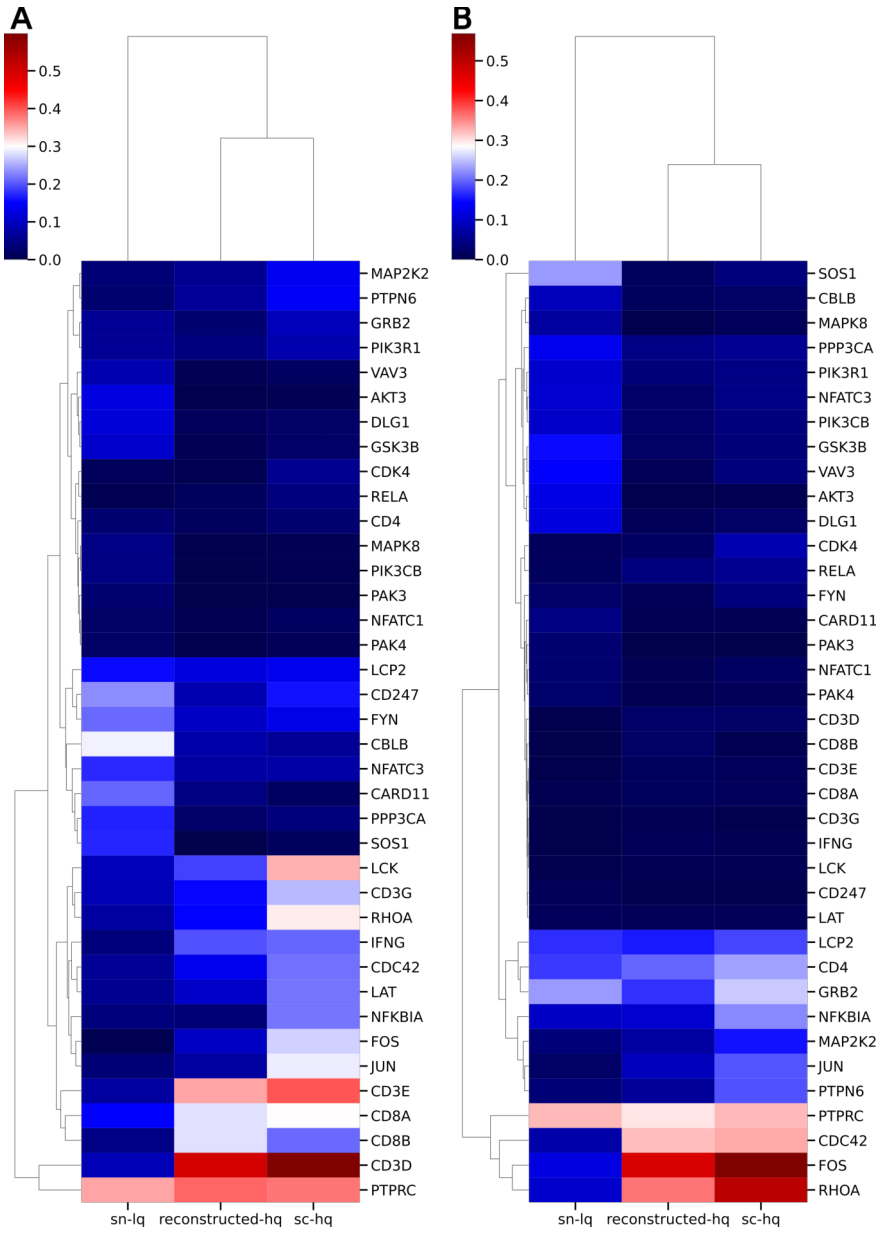


Figure S21: *Average gene expression of T cell receptor signaling genes in T cells (**A**) and Macrophages (**B**).* The columns show the data in the snRNA-seq (before reconstruction, sn-lq) and snRNA-seq dataset after reconstruction with DISCERN (reconstructed-hq), after reconstruction with Seurat (seurat-hq) and in the scRNA-seq data (sc-hq). The average expression was min-max scaled with adding a pseudocount of 1×10^{-3}. The reconstructed-hq shows high similarity with the expression in the sc-hq dataset. Only genes with a maximum expression greater than 0.2 are shown.

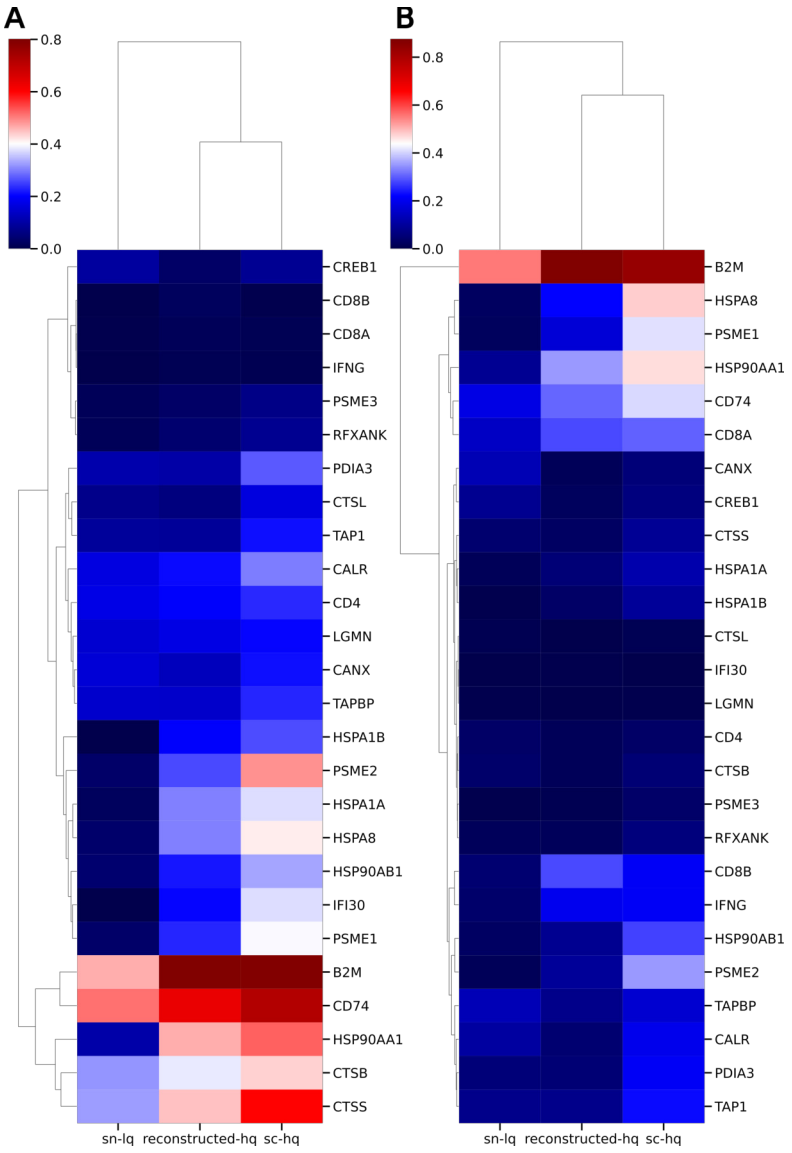


Figure S22: *Average gene expression of antigen presentation and processing genes in Macrophages (**A**) and T cells (**B**).* The columns show the data in the snRNA-seq (before reconstruction, sn-lq) and snRNA-seq dataset after reconstruction with DISCERN (reconstructed-hq), after reconstruction with Seurat (seurat-hq) and in the scRNA-seq data (sc-hq). The average expression was min-max scaled with adding a pseudocount of 1×10^{-3}. reconstructed-hq shows high similarity with the expression in the sc-hq dataset. Only genes with a maximum expression greater than 0.2 are shown. *CD4* and *CD8A* genes are part of the antigen presentation and processing pathway (`https://www.genome.jp/pathway/hsa04612`) but are naturally not expressed in Macrophages, thus no expression of these genes is expected in **A**.

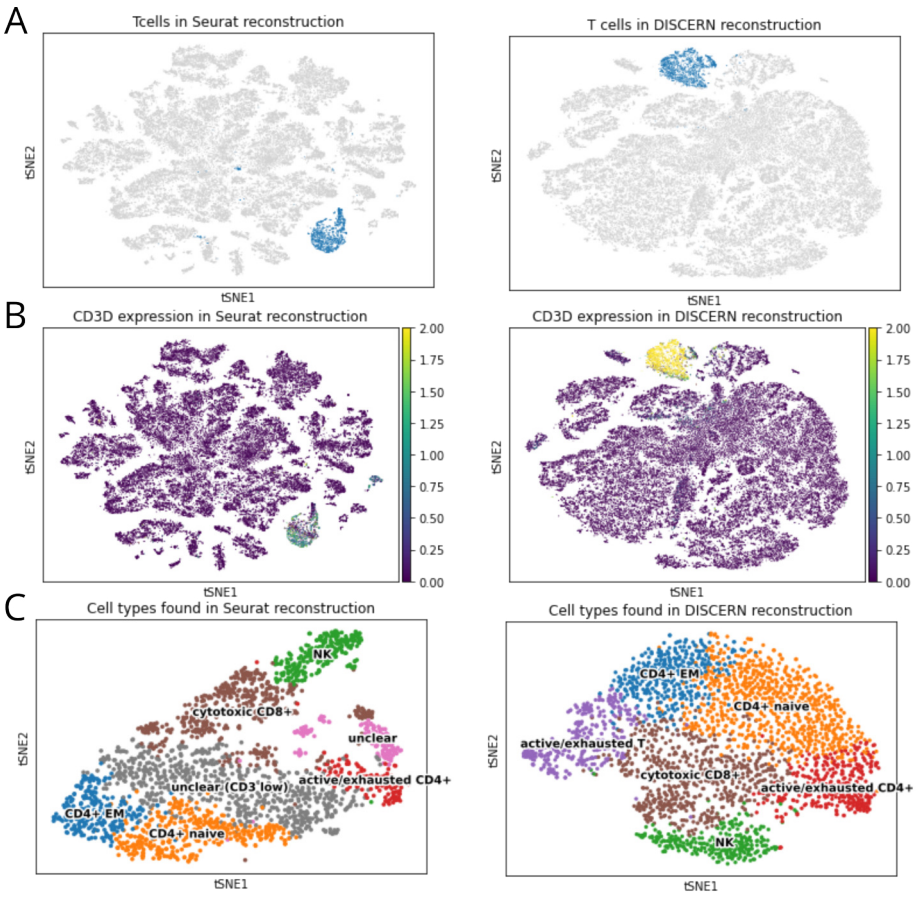


Figure S23: *T cell detection and sub-clustering in Kidney snRNA-seq (kidney-lq) and scRNA-seq (kidney-hq) data of patients with acute kidney injury.* **A**: tSNE representation of T cells found in Seurat (left) and DISCERN (right) reconstructed snRNA-seq and scRNA-seq data. **B**: tSNE representation of T cells found in Seurat (left) and DISCERN (right) reconstructed snRNA-seq and scRNA-seq data colored by *CD3D* expression as marker for T cells. **C**: tSNE representation of T cell subtypes found in Seurat (left) and DISCERN (right) reconstructed kidney-lq and kidney-hq data. A high number of cells in Seurat reconstruction could not be further classified due to absent or low expression of marker genes.

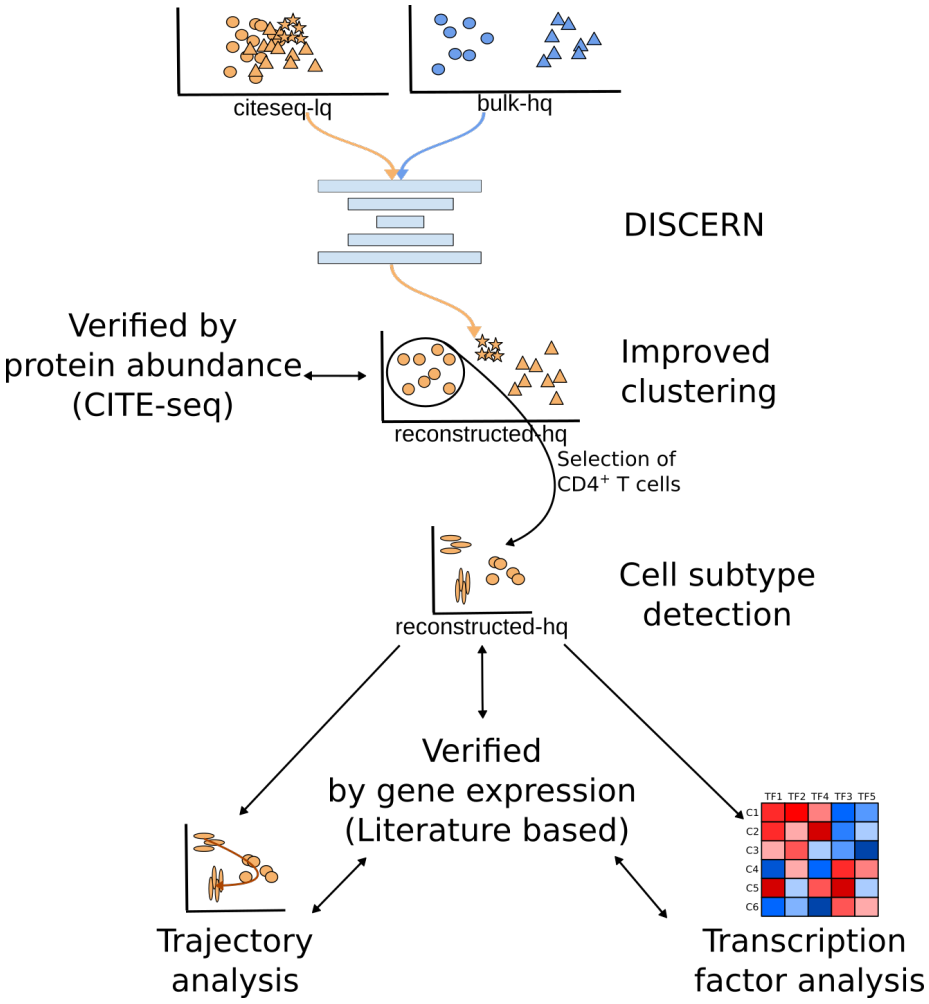


Figure S24: *Schematic representation of the experiments conducted with blood-based citeseq dataset.* To improve cite-lq data we reconstructed it with bulk-hq data to obtain reconstructed-hq data, which enabled improved clustering and CD4$^+$ T cell subtype detection. Additionally trajectory analysis and transcription factor analysis was performed on the CD4$^+$ T cell subset. Results were verified using protein abundance (CITE-seq) and literature information.

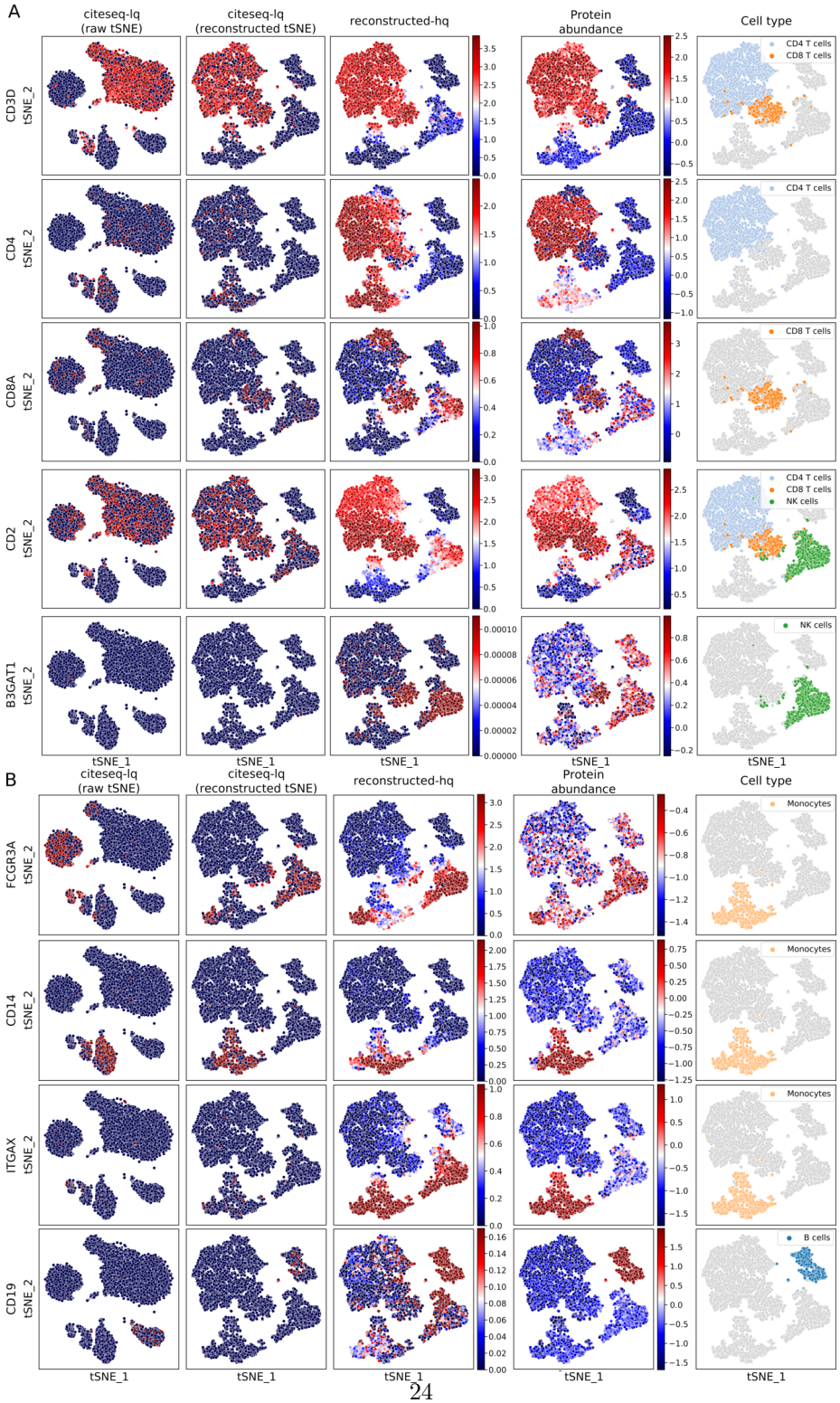

A
citeseq-lq (raw tSNE)
citeseq-lq (reconstructed tSNE)
reconstructed-hq
Protein abundance
Cell type
CD3D tSNE_2
CD4 tSNE_2
CD8A tSNE_2
CD2 tSNE_2
B3GAT1 tSNE_2
CD4 T cells
CD8 T cells
NK cells
tSNE_1
B
FCGR3A tSNE_2
CD14 tSNE_2
ITGAX tSNE_2
CD19 tSNE_2
Monocytes
B cells

Figure S25: *t-SNE representation of the gene expression and corresponding protein expression for the citeseq dataset before and after reconstruction.* **A** & **B**: Gene expression levels before reconstruction of the cite-lq data (first and second column) and after reconstruction to reconstructed-hq data using a bulk-hq reference (third column). Protein abundance measured using CITE-seq information is shown in the fourth column and the corresponding cell type in the fifth column. The first column shows tSNE representation computed on the uncorrected cite-lq data, while the others are computed on the reconstructed-hq data. Gene and protein expression levels are displayed in blue for low to red for high expression.

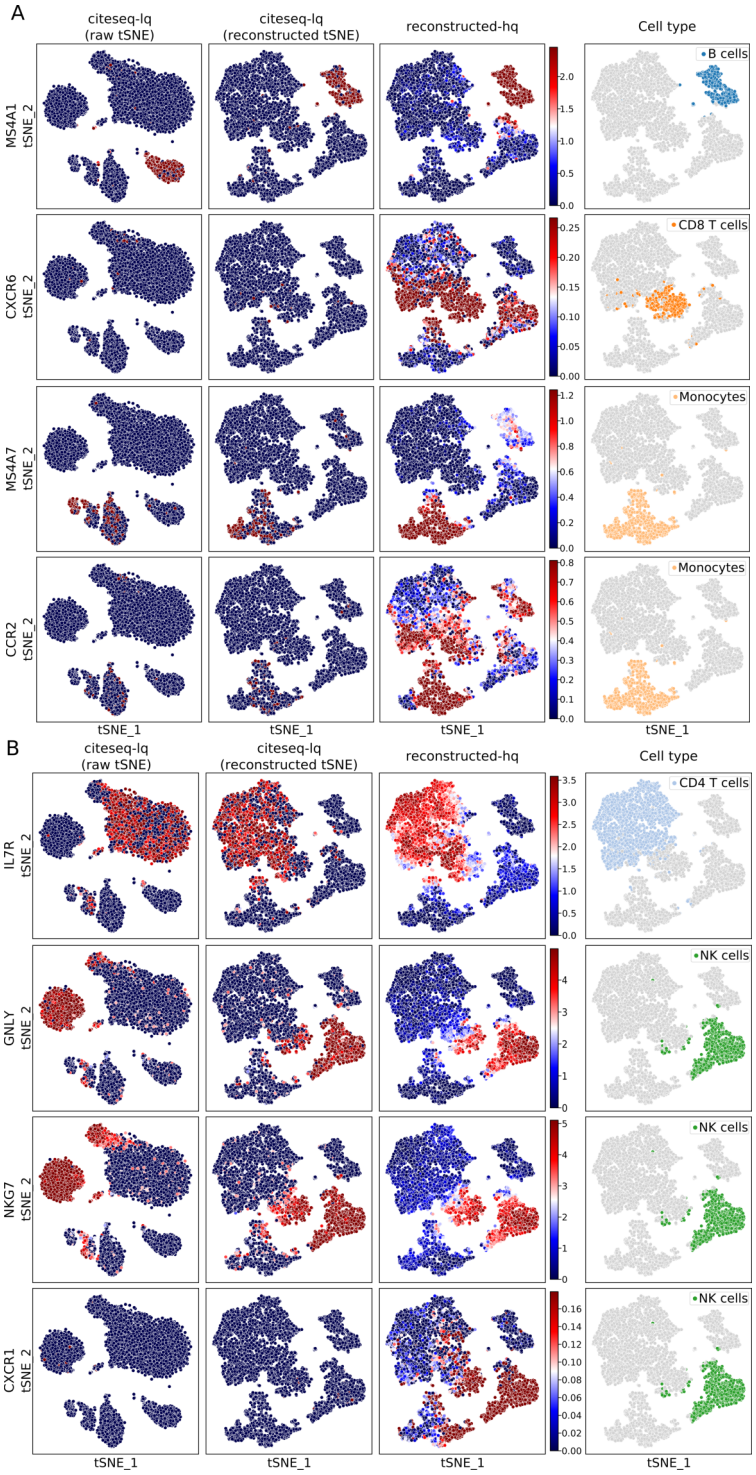

A
citeseq-lq (raw tSNE)
citeseq-lq (reconstructed tSNE)
reconstructed-hq
Cell type
MS4A1 tSNE_2
CXCR6 tSNE_2
MS4A7 tSNE_2
CCR2 tSNE_2
B cells
CD8 T cells
Monocytes
Monocytes
tSNE_1
B
IL7R tSNE_2
GNLY tSNE_2
NKG7 tSNE_2
CXCR1 tSNE_2
CD4 T cells
NK cells
NK cells
NK cells

Figure S26: *t-SNE representation of the gene expression of genes without CITE-seq information for the citeseq dataset before and after reconstruction.* **A** & **B**: Gene expression levels before reconstruction of the cite-lq data (first and second column) and after reconstruction to reconstructed-hq data using a bulk-hq reference (third column). The corresponding cell type is displayed in the fourth column. The first column shows tSNE representation computed on the uncorrected cite-lq data, while the others are computed on the reconstructed-hq data. Gene expression levels are displayed in blue for low to red for high expression.

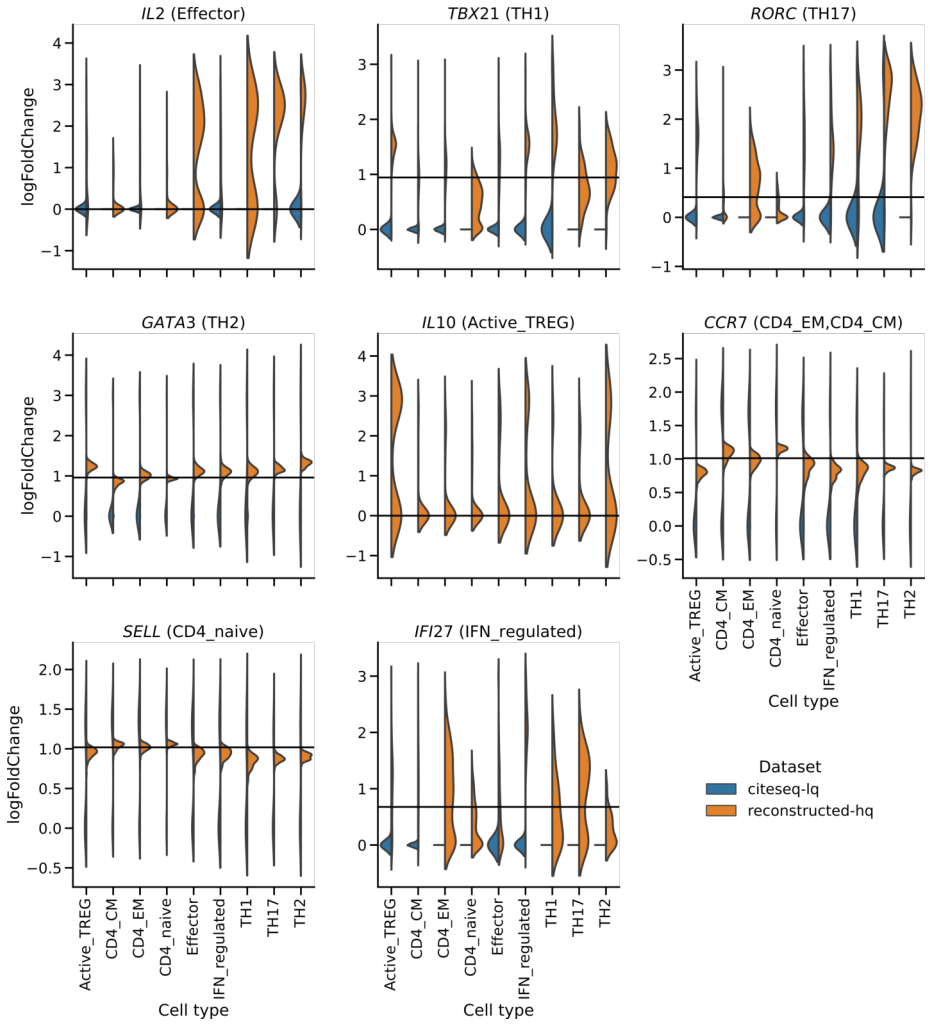


Figure S27: *Violin plots of cell type determining genes of CD4$^+$ T helper cell subtypes in the citeseq dataset. The expression is normalized by the mean over all cell types and log2-scaled.* Colors indicate whether they are shown for the uncorrected cite-lq data (blue) and the reconstructed-hq data (orange). The horizontal bars indicate median expression in the total citeseq dataset. The reconstructed gene expression is in many cases consistent with literature information. TH17 cells, for instance, are characterized by a high expression of *RORC* [4], TH2 cells express the transcriptional regulator *GATA3* [5], and TH1 cells the transcriptional regulator *TBX21* (*T-bet*) (fig. S28) [6]. *IL10* is produced by Foxp3 positive Treg cells (Active_TREG) [7]. *SELL* and *CCR7* are expressed in the CD4$^+$ T cell subtypes CD4_naive, CD4_EM, CD4_CM, but with a significantly lower expression of *CCR7* in CD4_EM cells [8], while CD4_naive cells show the highest expression of *SELL/CD62L* (fig. S29) [9].

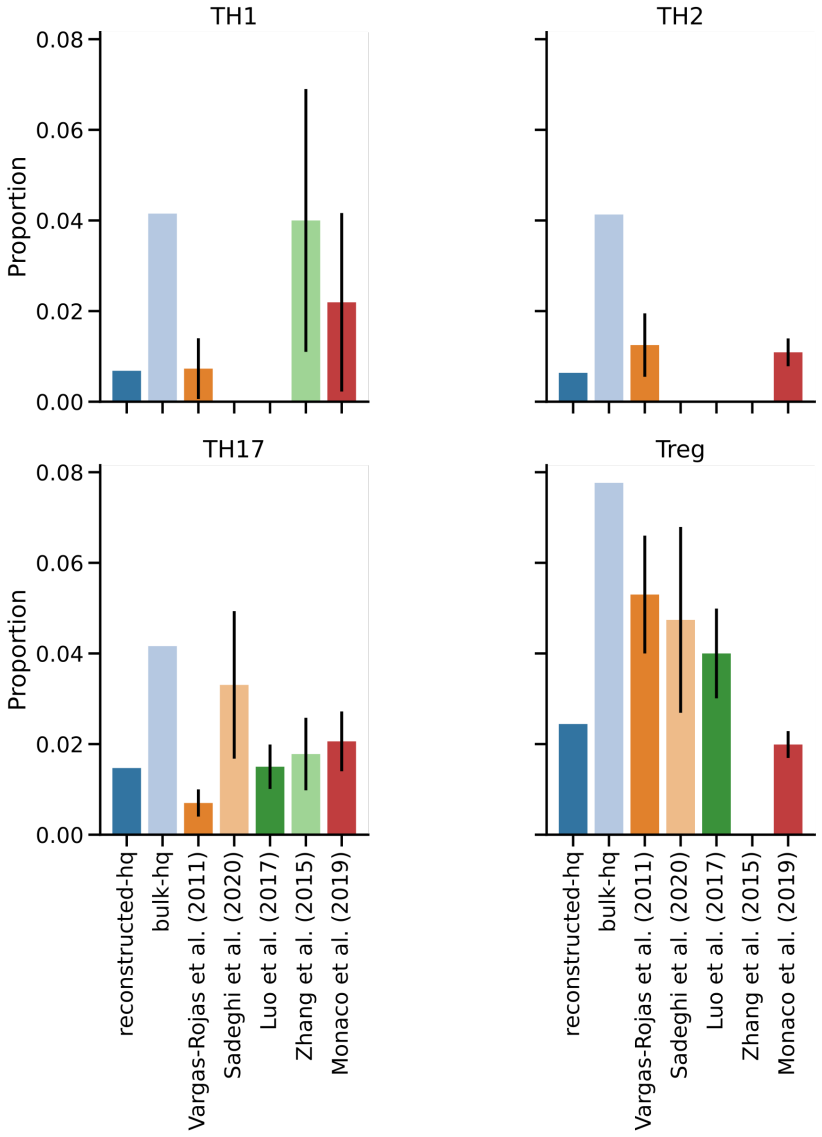


Figure S28: *Bar plot showing the proportions of CD4$^+$ T helper cell subtypes (TH1, TH2, TH17, and Treg) identified in the reconstructed-hq data, bulk-hq training data, and published ground-truth cell fractions in the citeseq data.* The proportions are calculated with respect to the total number of PBMCs. To compare the proportions in the reconstructed data with existing literature, five studies were considered (see also table S3). These studies estimate one or more of these subtypes using FACS and subsequent cell activation. For these references, bars represent means while error bars represent standard deviation. Missing bars indicate that the corresponding cell-type is not quantified in the referenced study.

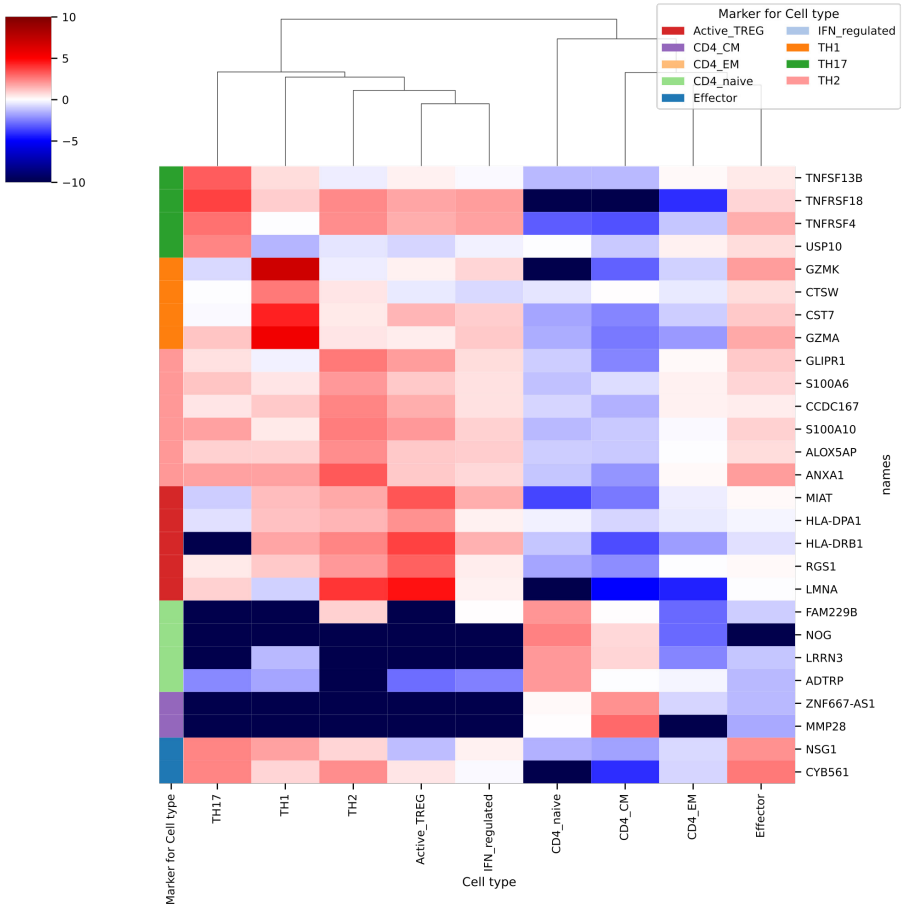


Figure S29: *Heatmap of log2 fold-change (FC) for the top CD4$^+$ T helper cell subtype marker genes in the citeseq dataset after reconstruction with DISCERN.* The cite-lq data was reconstructed with bulk-hq data to obtain reconstructed-hq data, which is displayed in the figure. Genes were filtered for an adjusted p-value ≤ 0.05 and a log2 FC ≥ 2. The top five genes with the lowest adjusted p-value were selected from the genes passing the threshold and duplicate gene entries were removed. For display reasons, the negative log2 FC was clipped at -10. The first column indicates cell type-specific expression according to the DEG analysis. FC magnitude is depicted with blue - low to red - high changes.

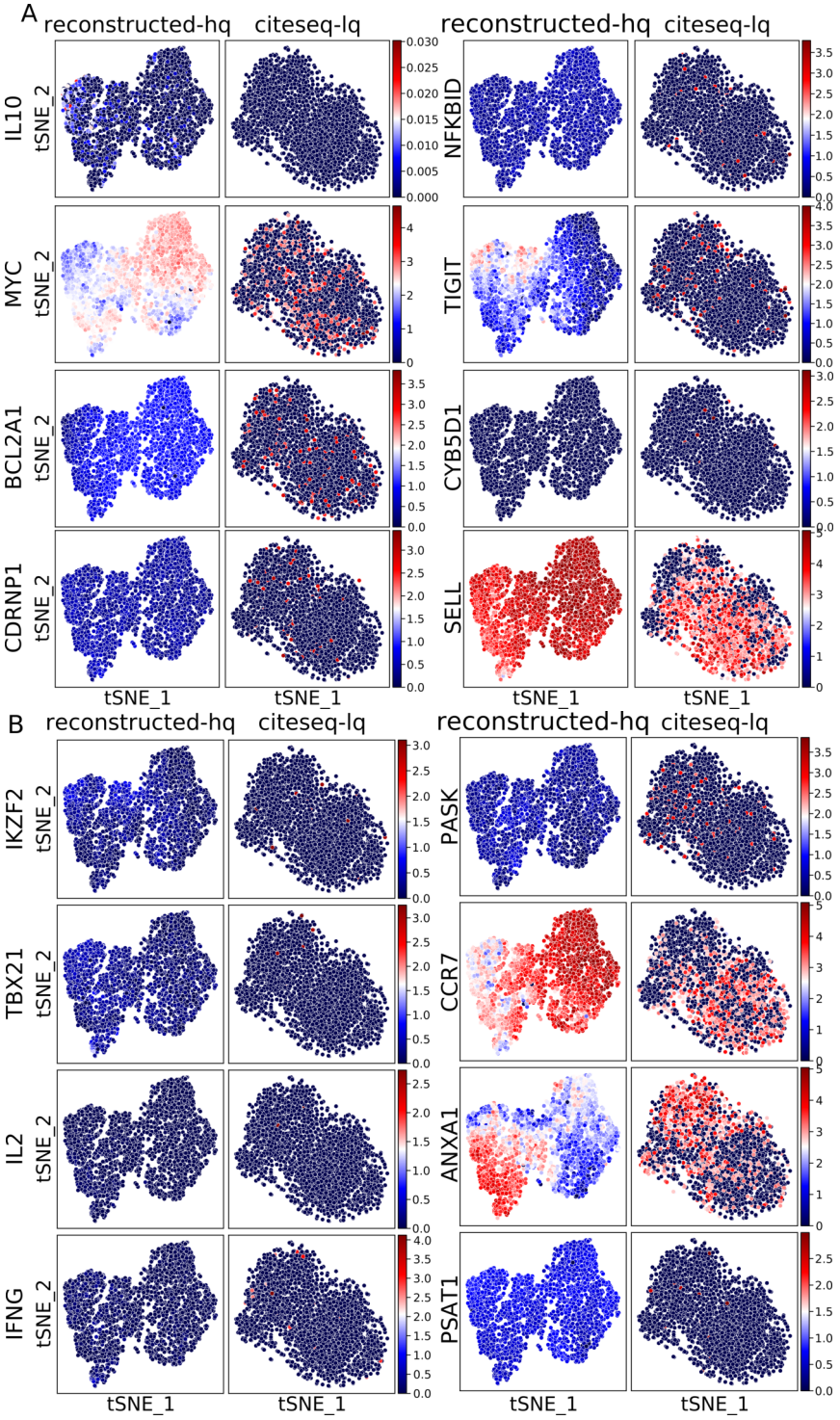

A
reconstructed-hq
citeseq-lq
IL10
NFKBID
MYC
TIGIT
BCL2A1
CYB5D1
CDRNP1
SELL
tSNE_2
tSNE_1
B
IKZF2
PASK
TBX21
CCR7
IL2
ANXA1
IFNG
PSAT1

Figure S30: *t-SNE representation of the gene expression of several established T cell marker genes in citeseq $CD4^+$ T cells before and after correction with DISCERN.* **A** & **B**: Gene expression levels before reconstruction of the cite-lq data (third and fourth column, t-SNE calculated on cite-lq data) and after reconstruction to reconstructed-hq data using a bulk-hq reference (first and third column, t-SNE calculated on reconstructed-lq data). Gene expression levels are displayed in blue for low to red for high expression. *MYC*, *NFKBID*, *BCL2A1*, *CYB5D1*, *CSRNP1*, *IL2*, and *PSAT1* are used as activation markers, *IL10*, *TBX21*, *ANXA1*, *IFNG* characterize TH1 cells, *TIGIT* and *PASK* TFH cells, *IKZF2* TREG, *CCR7* central memory T cells and *SELL* is a marker for naive T cells. In general, the cell type-specific expression of published marker genes in the reconstructed-hq data show good correspondence with the identified cell types.

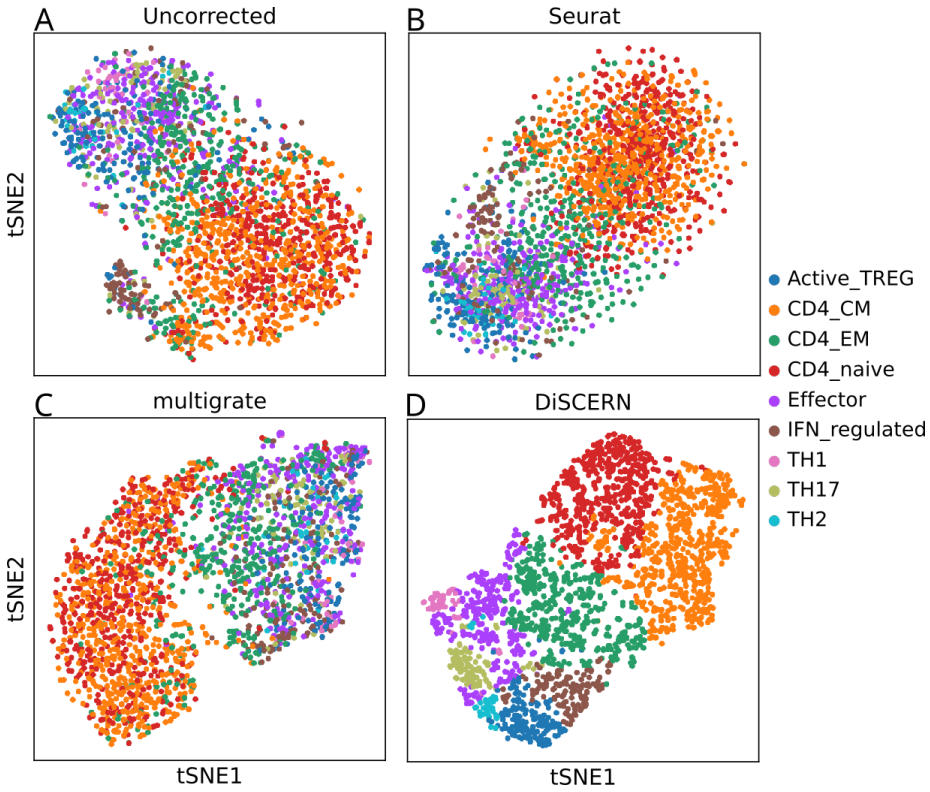


Figure S31: *tSNE representation of CD4$^+$ T cells in the citeseq dataset after annotation of the cell types found after expression reconstruction with DISCERN.* **A**: Uncorrected citeseq-lq CD4$^+$ T cells show some clustering of IFN_regulated, Active_TREG, CD4_CM and CD4_naive cells. **B**: Seurat reconstruction results in strong cluster and cell type mixing, a potential sign of overintegration. **C**: Multigrate imputed data shows strong mixing and splitting of clusters and cell types, for instance CD4$^+$ T cells are split into two clusters. **D**: DISCERN reconstructed cite-hq data provides a clear separation of functionally distinct cell types.

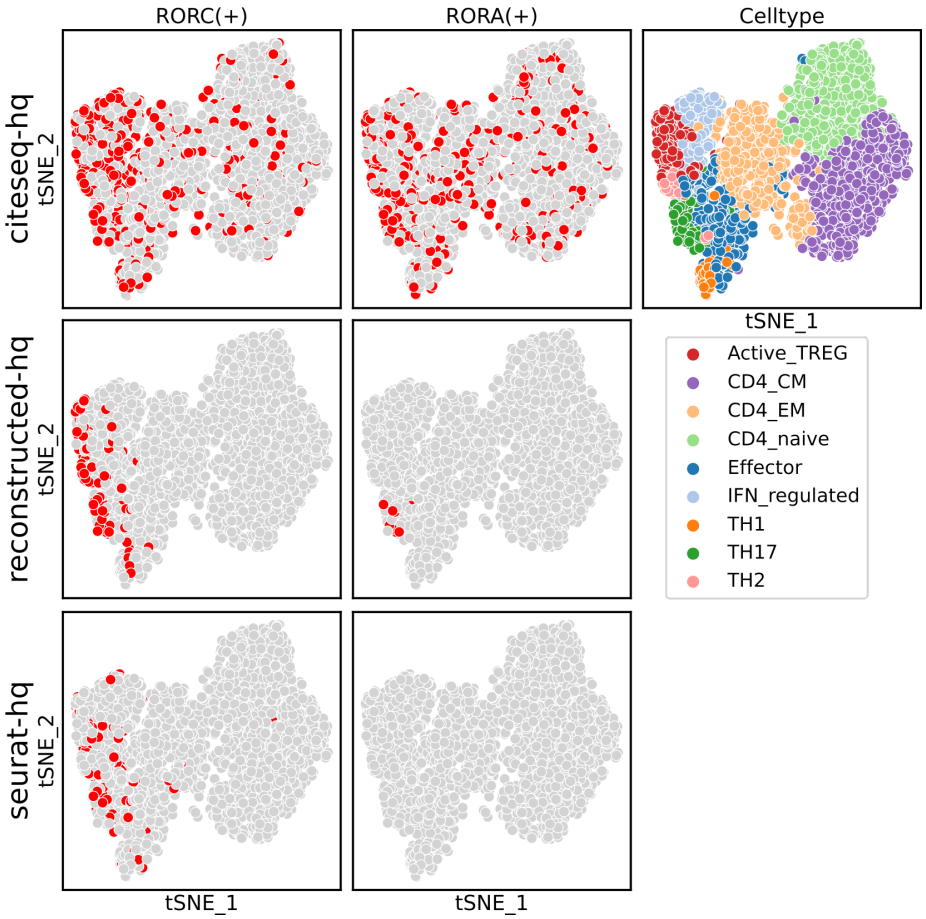
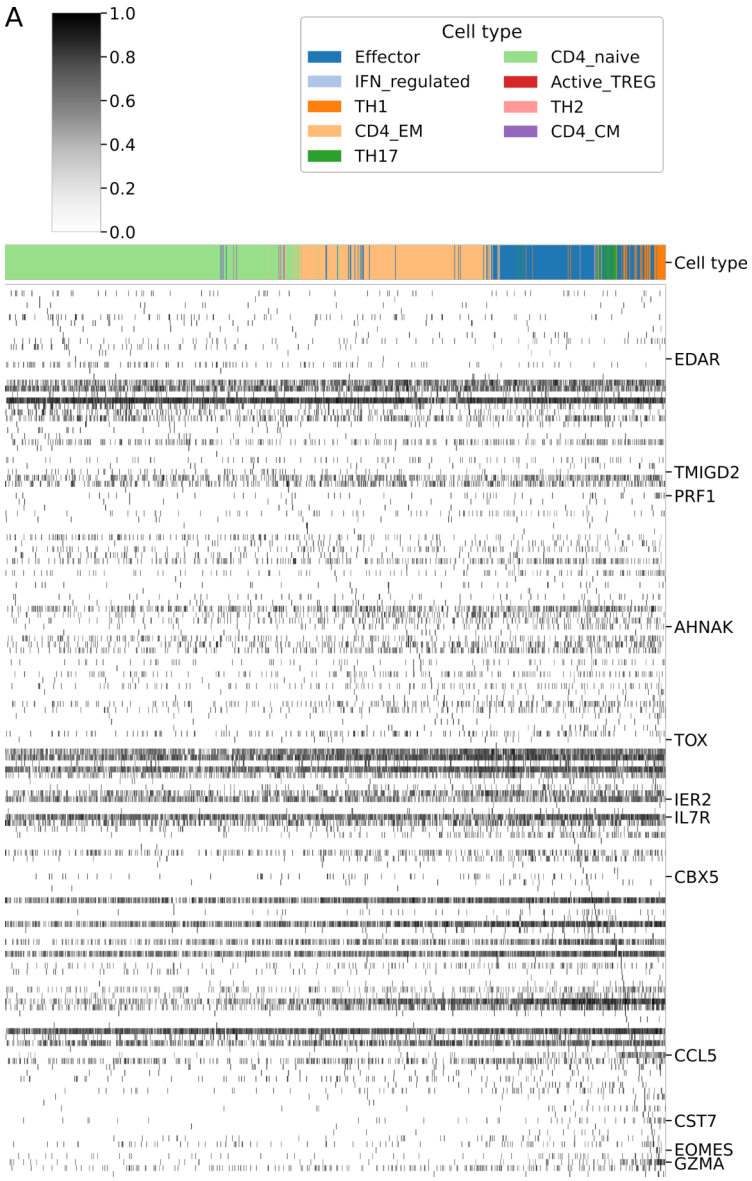
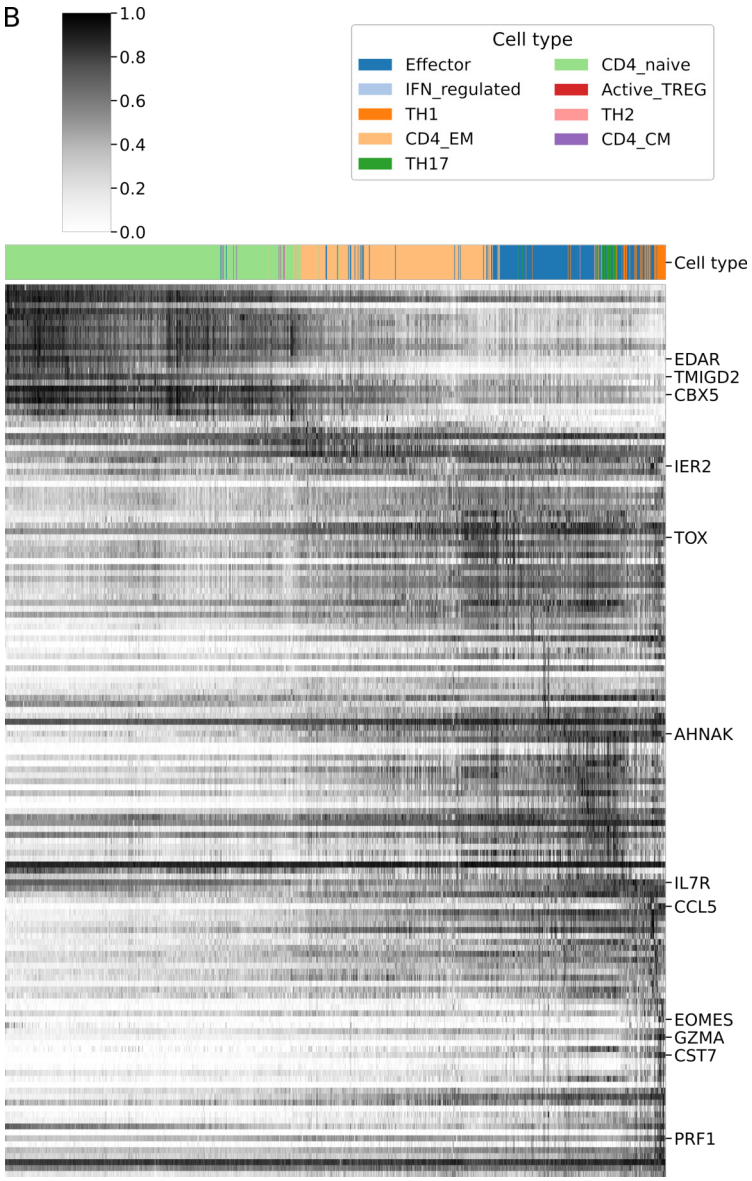


Figure S32: *t-SNE representation of a SCENIC transcriptional regulation analysis of citeseq T helper cells before and after correction with DISCERN and Seurat.* Gene expression levels before reconstruction of the cite-lq data (first row) and after reconstruction to reconstructed-hq data using a bulk-hq reference with DISCERN (second row) and Seurat (third row). The first column displays cells that express the RORC(+) regulon. The second column displays cells that express the RORA(+) regulon. Red color in the first two columns depends on the binarized AUCell score of the SCENIC discovered regulons. The third column displays the detected cell types in reconstructed-hq data. RORA(+) and RORC(+) regulons are expected to be specific for TH17 cells. The tSNE representation is calculated on DISCERN reconstructed data.


A
1.0
0.8
0.6
0.4
0.2
0.0
Cell type
Effector
IFN_regulated
TH1
CD4_EM
TH17
CD4_naive
Active_TREG
TH2
CD4_CM
Cell type
EDAR
TMIGD2
PRF1
AHNAK
TOX
IER2
IL7R
CBX5
CCL5
CST7
EOMES
GZMA


B
1.0
0.8
0.6
0.4
0.2
0.0
Cell type
Effector
IFN_regulated
TH1
CD4_EM
TH17
CD4_naive
Active_TREG
TH2
CD4_CM
Cell type
EDAR
TMIGD2
CBX5
IER2
TOX
AHNAK
IL7R
CCL5
EOMES
GZMA
CST7
PRF1

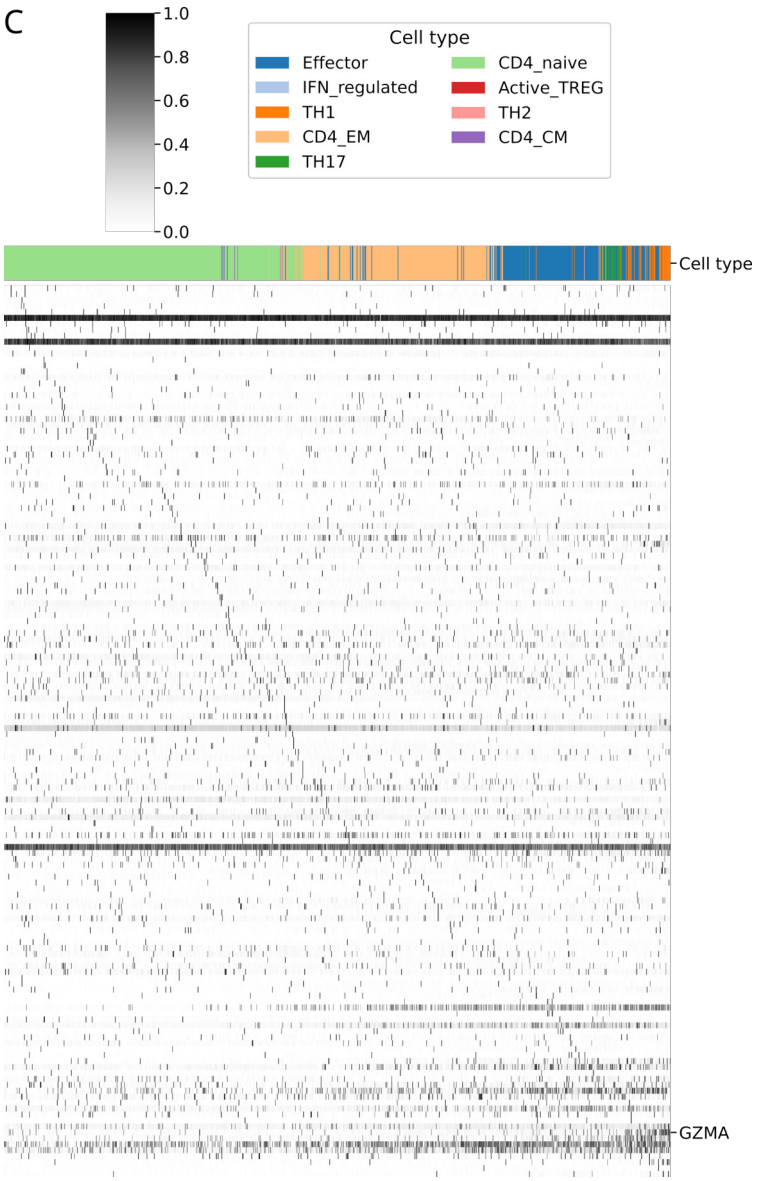
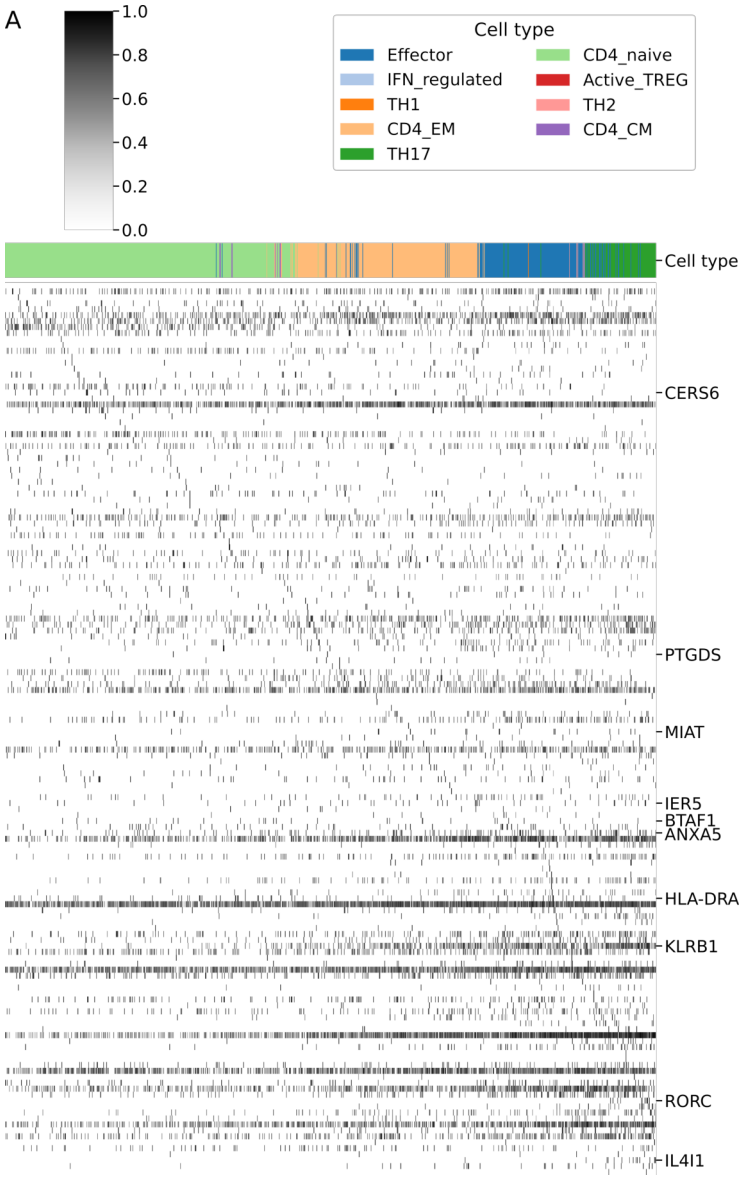
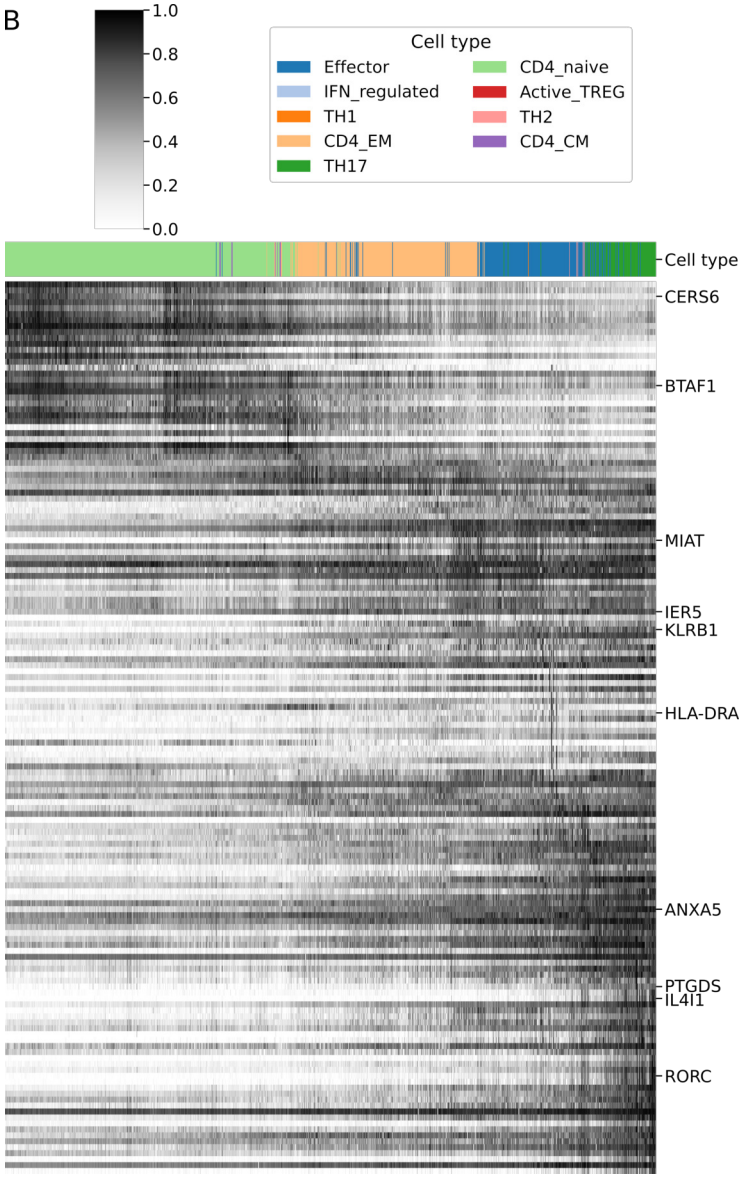


Figure S33: *Expression of differentially regulated genes of the CD4_naive to TH1 lineage (Lineage1) defined by a Slingshot trajectory analysis.* The top 150 genes by p-value are shown. Only a selection of T cell marker genes is shown by name. The cell types are color-coded and cells are sorted by pseudotime. **A**: Expression using the cite-lq data before reconstruction. **B**: Expression using the reconstructed-hq data that was reconstructed with DISCERN using a bulk-hq reference. In the reconstructed-hq data, Lineage1 shows a trajectory from *TMIGD2*,

EDAR and *CBX5* expressing CD4_naive cells [10, 11] to TH1 cells expressing cytotoxicity-related genes like *EOMES*, *CST7*, *GZMA*, *IL7R*, *CCL5* and *PRF1* [12, 13, 14, 15, 16, 17, 18]. The cells develop through CD4_EM cells to a (pre-) effector state (Effector cells) to the final TH1 subtype (fig. S34B). Effector and CD4_EM cells show higher expression of *IER2* [19], *AHNAK* [20] and *TOX* [21], as reported in the literature. In general, reconstructed data show cell trajectories that are biologically reasonable, while uncorrected data shows little structure. **C**: Expression using the Seurat to reconstruct the data (seurat-hq). While the heatmap shows a small trajectory line, most of the marker genes are not found along the trajectory.


A
1.0
0.8
0.6
0.4
0.2
0.0
Cell type
Effector
IFN_regulated
TH1
CD4_EM
TH17
CD4_naive
Active_TREG
TH2
CD4_CM
Cell type
CERS6
PTGDS
MIAT
IER5
BTAF1
ANXA5
HLA-DRA
KLRB1
RORC
IL4I1


B
1.0
0.8
0.6
0.4
0.2
0.0
Cell type
Effector
IFN_regulated
TH1
CD4_EM
TH17
CD4_naive
Active_TREG
TH2
CD4_CM
Cell type
CERS6
BTAF1
MIAT
IER5
KLRB1
HLA-DRA
ANXA5
PTGDS
IL4I1
RORC

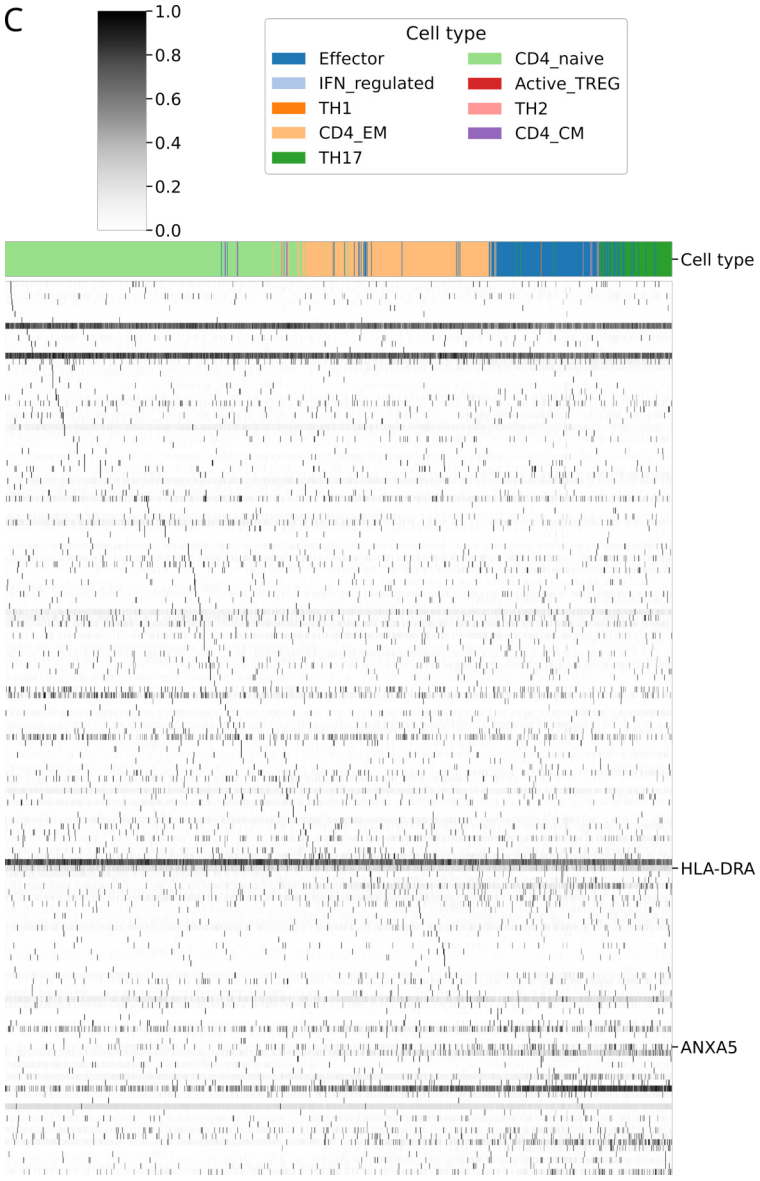


Figure S34: *Expression of differentially regulated genes of the CD4_naive to TH17 lineage (Lineage2) defined by a Slingshot trajectory analysis.* The top 150 genes by p-value are shown. Only a selection of T cell marker genes is shown by name. The cell types are color-coded and cells are sorted by pseudotime. **A:** Expression using the cite-lq data before reconstruction. **B:** Expression using the reconstructed-hq data that was reconstructed with DISCERN using a bulk-hq reference. In the reconstructed-hq data, the CD4_naive cells show the expected

BTAF1 and *CERS6* expression [11, 22], whereas effector cells express activation markers as MIAT [23], *HLA-DRA* [24], *IER5* [25] and *KLRB1* [26]. Finally the trajectory terminates in TH17 cells, expressing *ANXA5* [11], *RORC* [4], *IL4I1* [27] and *PTGDS* [28], showing that the found trajectory (Fig. 3C) is in line with known expression patterns. **C**: Expression using the Seurat to reconstruct the data (seurat-hq). While the heatmap shows a small trajectory line, most of the marker genes are not found along the trajectory.

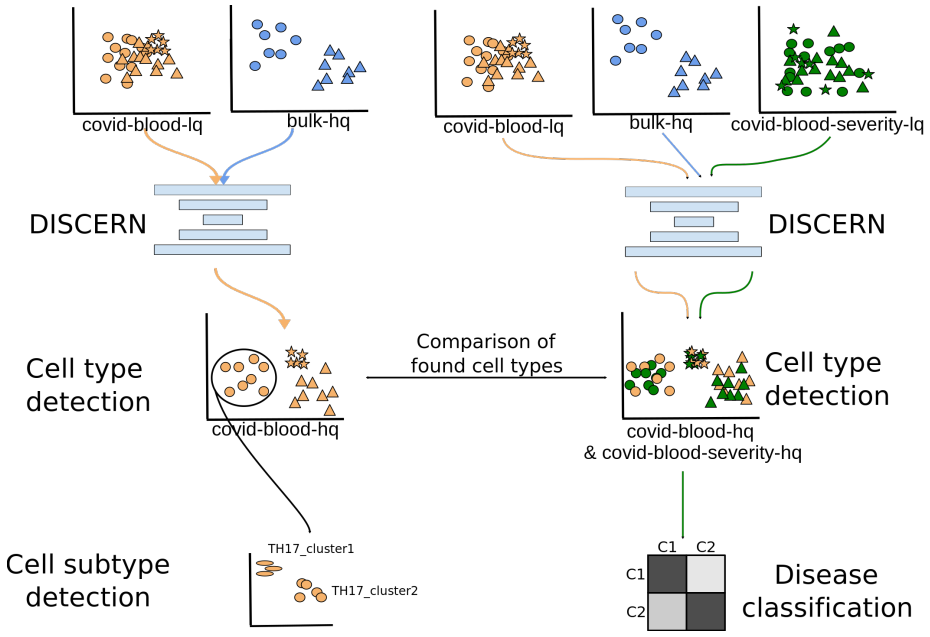


Figure S35: *Schematic representation of the experiments conducted with the covid-blood and covid-blood-severity datasets.* To improve covid-blood-lq and covid-blood-severity-lq data we reconstructed it using bulk-hq data to obtain covid-blood-hq and covid-blood-severity-hq data, respectively. We then performed cell type detection using the reconstructed data. We investigated T helper cell subtypes in great detail in the covid-blood-hq data and compared them to the ones found in the covid-blood-severity-hq data. Finally, we used the covid-blood-severity dataset and its disease severity information for COVID-19 patients to classify mild and severe cases using a GBM. TH17 cell subtypes could be detected in the covid-blood-hq data and linked to cells found in the covid–lung dataset using T cell receptor clonal information.

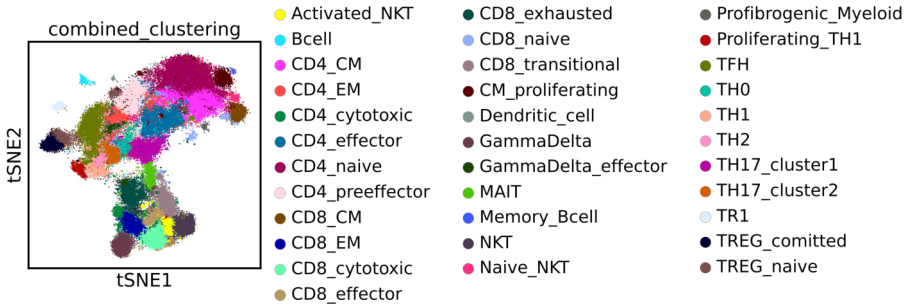
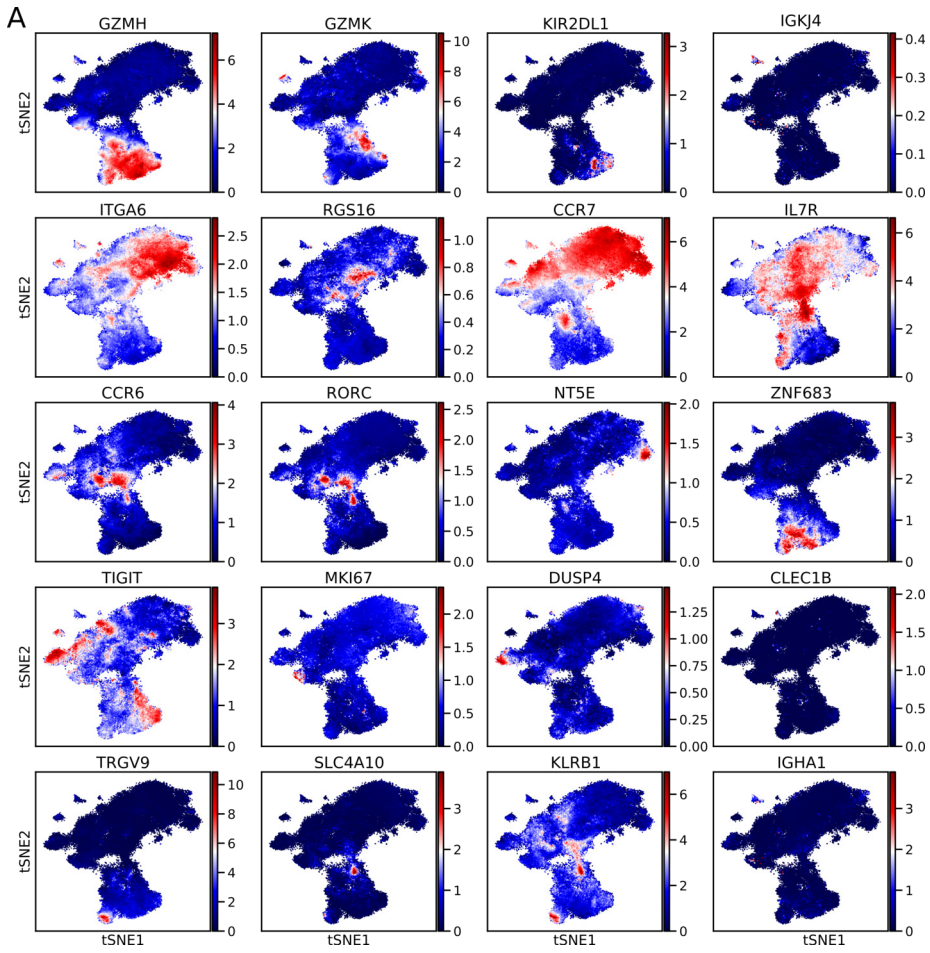


Figure S36: *t-SNE representation of cell types found in the covid-blood-hq data.* The covid-blood-lq data was reconstructed using bulk-hq data to obtain covid-blood-hq data. Colors indicate the annotated cell types. Especially T cell subtypes could not be annotated before reconstruction with DISCERN. It is especially interesting that TH17 subtypes can be detected, which are usually observed in FACS PBMC data only after stimulation.


A
GZMH
GZMK
KIR2DL1
IGKJ4
ITGA6
RGS16
CCR7
IL7R
CCR6
RORC
NT5E
ZNF683
TIGIT
MKI67
DUSP4
CLEC1B
TRGV9
SLC4A10
KLRB1
IGHA1
tSNE1
tSNE2

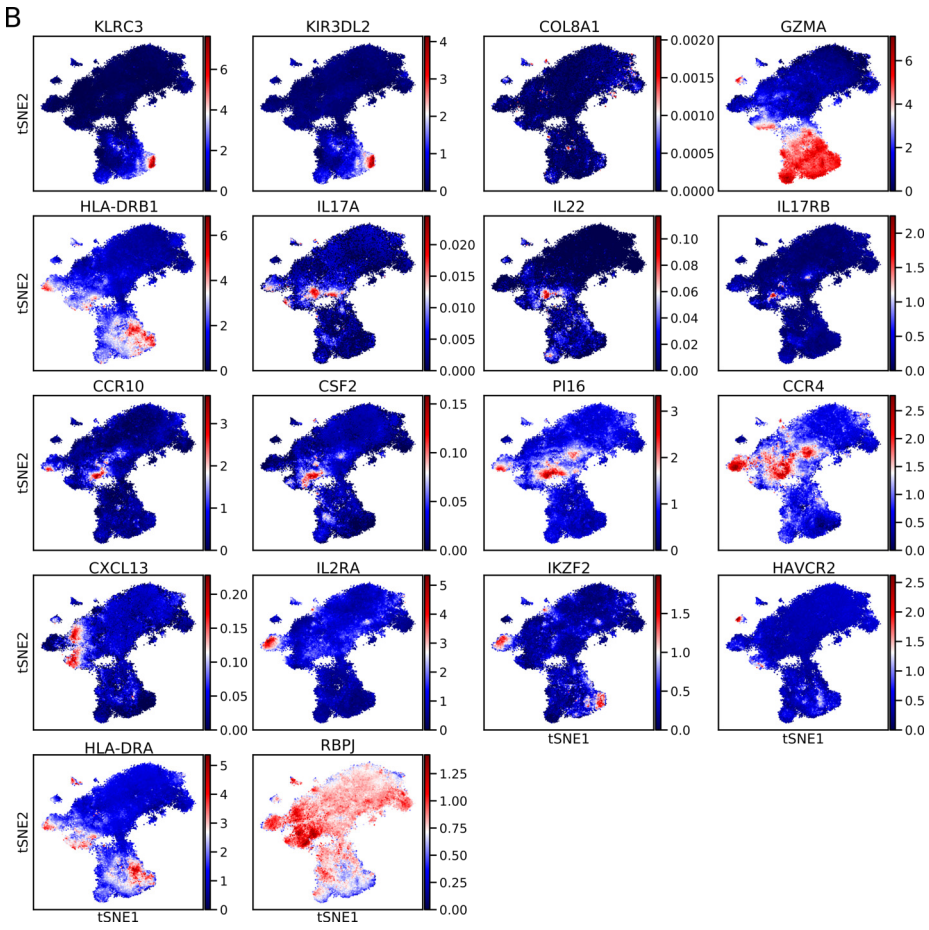
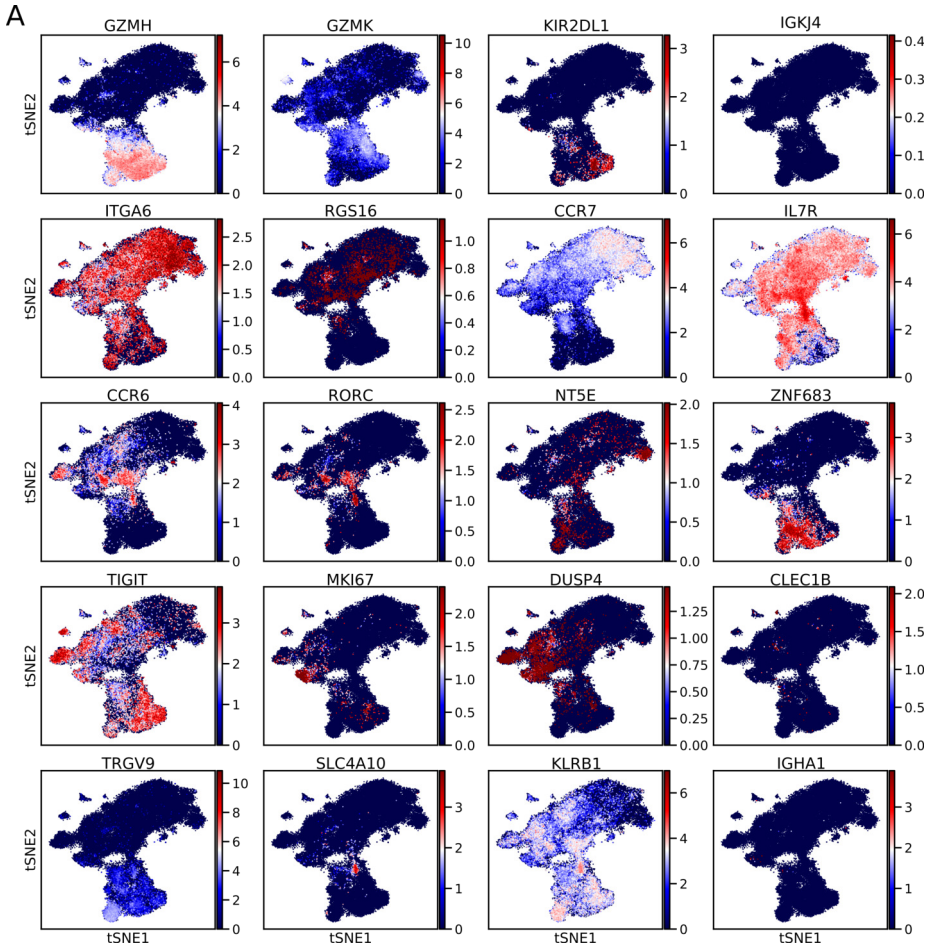


Figure S37: *t-SNE representation of the gene expression of several established marker genes in covid-blood-hq data.* **A** & **B**: The covid-blood-lq dataset after reconstruction with DISCERN using bulk-hq data to obtain covid-blood-hq data. The t-SNE representation was computed on the covid-blood-hq data. Gene expression levels are displayed in blue for low to red for high expression. In general, the cell type-specific expression of published marker genes in the covid-blood-hq data show good correspondence with the identified cell types.


A
GZMH
GZMK
KIR2DL1
IGKJ4
ITGA6
RGS16
CCR7
IL7R
CCR6
RORC
NT5E
ZNF683
TIGIT
MKI67
DUSP4
CLEC1B
TRGV9
SLC4A10
KLRB1
IGHA1
tSNE1
tSNE2

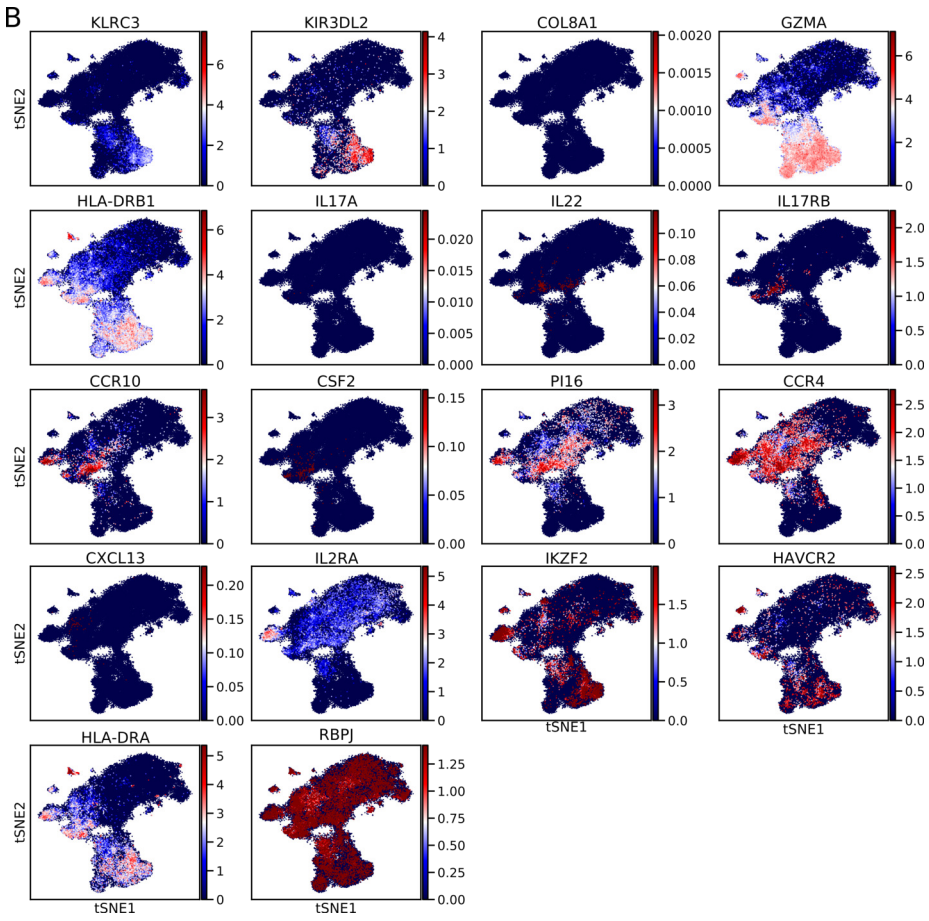


Figure S38: t-SNE representation of the gene expression of several established marker genes in covid-blood-lq data. **A** & **B**: The t-SNE representation was computed on the covid-blood-lq data without reconstruction. Gene expression levels are displayed in blue for low to red for high expression. In general, the cell type-specific expression of published marker genes in the covid-blood-lq data shows worse correspondence with the identified cell types as compared to the covid-blood-hq data in fig. S37.

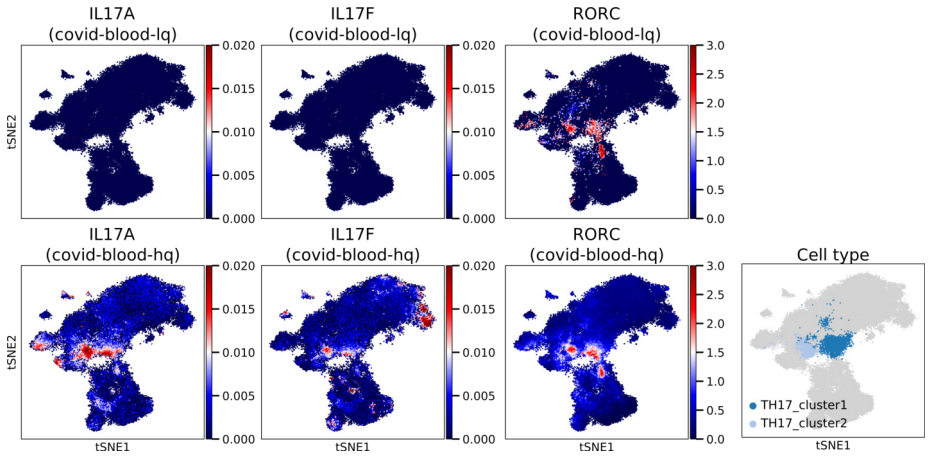


Figure S39: *t-SNE representation of TH17 marker genes in two TH17 subtypes detected in COVID-19 patient blood.* t-SNEs were calculated for $CD4^+$ T cells on covid-blood-hq data. The first row shows the expression of marker genes for uncorrected covid-blood-lq data. The second row displays the expression of the same marker genes for reconstructed covid-blood-hq data. The covid-blood-lq data was reconstructed using the bulk-hq reference to obtain covid-blood-hq data. The TH17 cell subclusters were found by louvain clustering after reconstruction. Colors represent the expression levels of genes as mentioned in the plot titles (*IL17A*, *IL17F*, *RORC*; from left to right). Expression levels of TH17 marker gene expression is barely visible for *IL17A/F* before reconstruction but can be detected after reconstruction with DISCERN. *RORC*, as transcription factor for TH17 cells, confirms the correct annotation of TH17 cells.

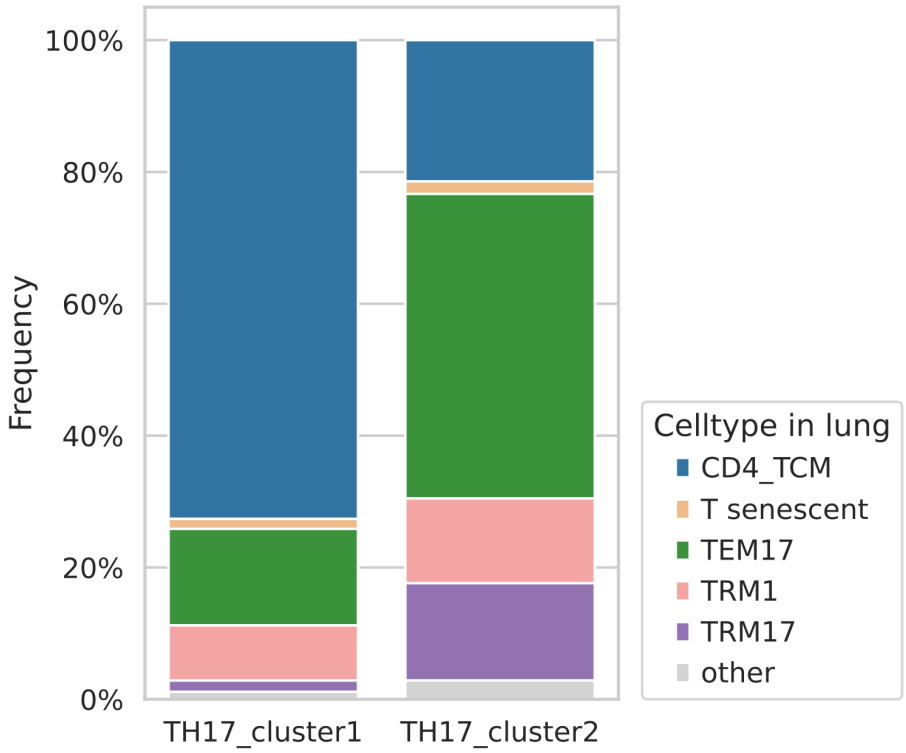


Figure S40: *Fraction of TH17 cells sharing the T cell receptor clonotype in covid-blood-hq and covid-lung data.* Cell type annotations of lung data were used as provided in the original publication. Cell types with an overlap < 1 % in both TH17 clusters were labeled as other. TH17_cluster1, detected in covid-blood-hq data, shares T cell receptor clones with CD4_TCM cells in the covid-lung data. TH17_cluster2, detected in covid-blood-hq data, shares most T cell receptor clones with TEM17 cells in covid-lung data. This corroborates the definition of the two TH17 subtypes detected in covid-blood-hq data and raises the question if these cells stay peripheral or re-enter tissues to promote inflammation.

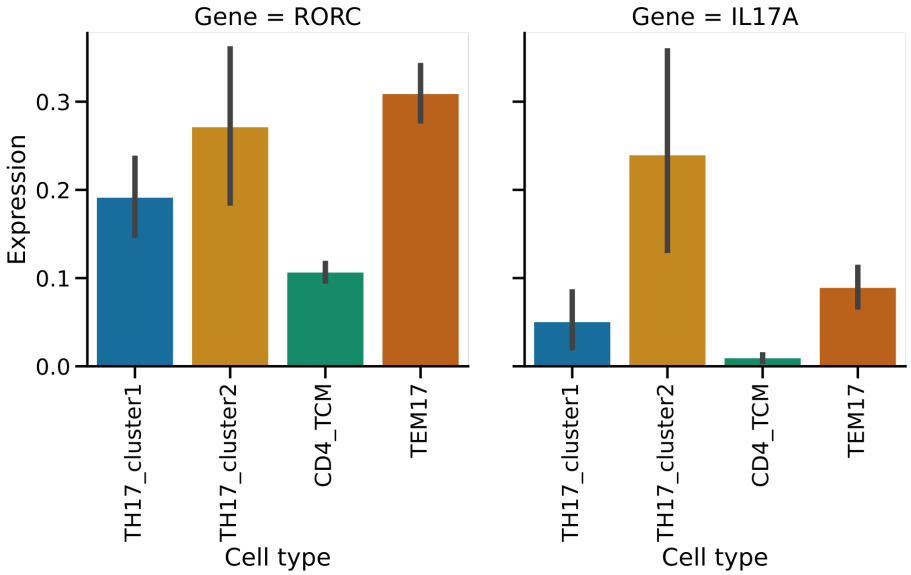


Figure S41: *Mean expression of RORC and IL17-A of covid-lung cells sharing a clonotype with TH17 cells of the covid-blood-hq data.* TH17_cluster1 and TH17_cluster2 are determined using the TCR clonotype information of reconstructed covid-blood-hq data and CD4_TCM or TEM17 covid-lung cells were annotated as in the original publication (see also fig. S40). A single cell can contribute to more than one bar, e.g. by being annotated as TEM and having a shared clonotype with TH17_cluster2 cell in covid-blood. Cell types sharing a clonotype with TH17_cluster1 and TH17_cluster2 cells from covid-blood have on average a higher or similar expression of the TH17 marker genes (*RORC* and *IL17A*) than cells in CD4_TCM or TEM17 cells in lung. This shows that CD4_TCM and TEM17 can most likely be further subdivided into clusters matching TH17_cluster1 and TH17_cluster2 in covid-blood and thus giving more evidence that these cell subtypes have a biological role in blood and lung.

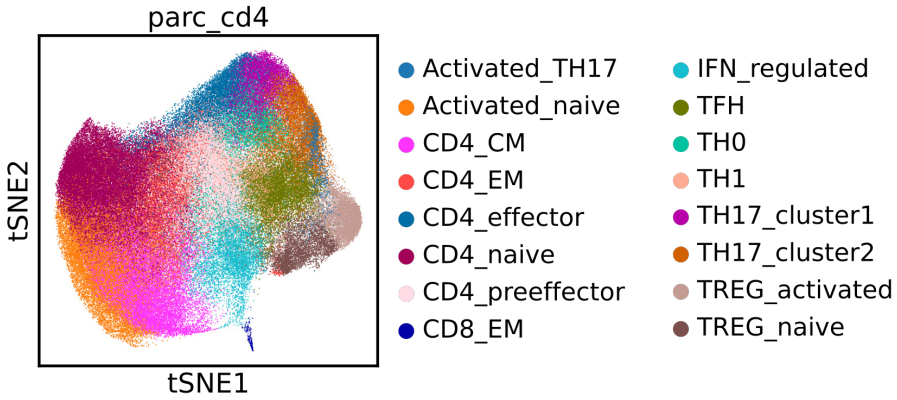


Figure S42: *t-SNE representation of $CD4^+$ T cells found in the covid-blood-severity-hq data.* The t-SNE representation and clustering was computed on DISCERN reconstructed expression, using covid-blood-severity-lq and covid-blood-lq data as input and bulk-hq as reference. Cell types are labeled by color. It is interesting to observe that the detected cell types largely overlap for the two studies.

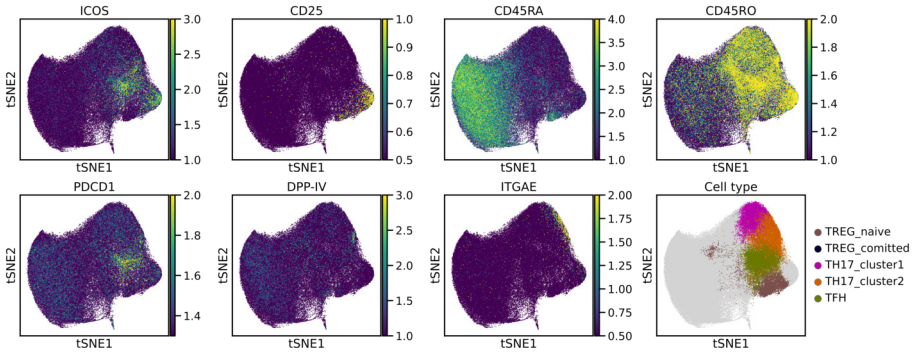


Figure S43: *t-SNE representation of the marker protein abundance in the covid-blood-severity dataset for TREG, TH17 and TFH cells provided by CITE-seq information.* The t-SNE representation and clustering was computed on DISCERN reconstructed expression, using covid-blood-severity-lq and covid-blood-lq data as input and bulk-hq as reference. The CITE-seq protein abundance of the covid-blood-severity data for seven marker proteins is displayed in color (blue - low to yellow - high abundance). The region we identified as regulatory T cells is positive for CD25 and CD45RO^{+} and activated TREGs are high in ICOS as described for highly suppressive TREG [29]. TFH cells are PDCD1 and ICOS surface protein positive cells [30] and DPP-IV is markedly increased in activated TH17 cells expressing IL17A, a TH17-specific signal as previously described [31]. ITGAE abundance was described for resident T Helper cells in the skin reentering circulation [32]. In general, the CITE-seq information confirms the cell type identification of DISCERN reconstructed covid-blood-severity-hq data.

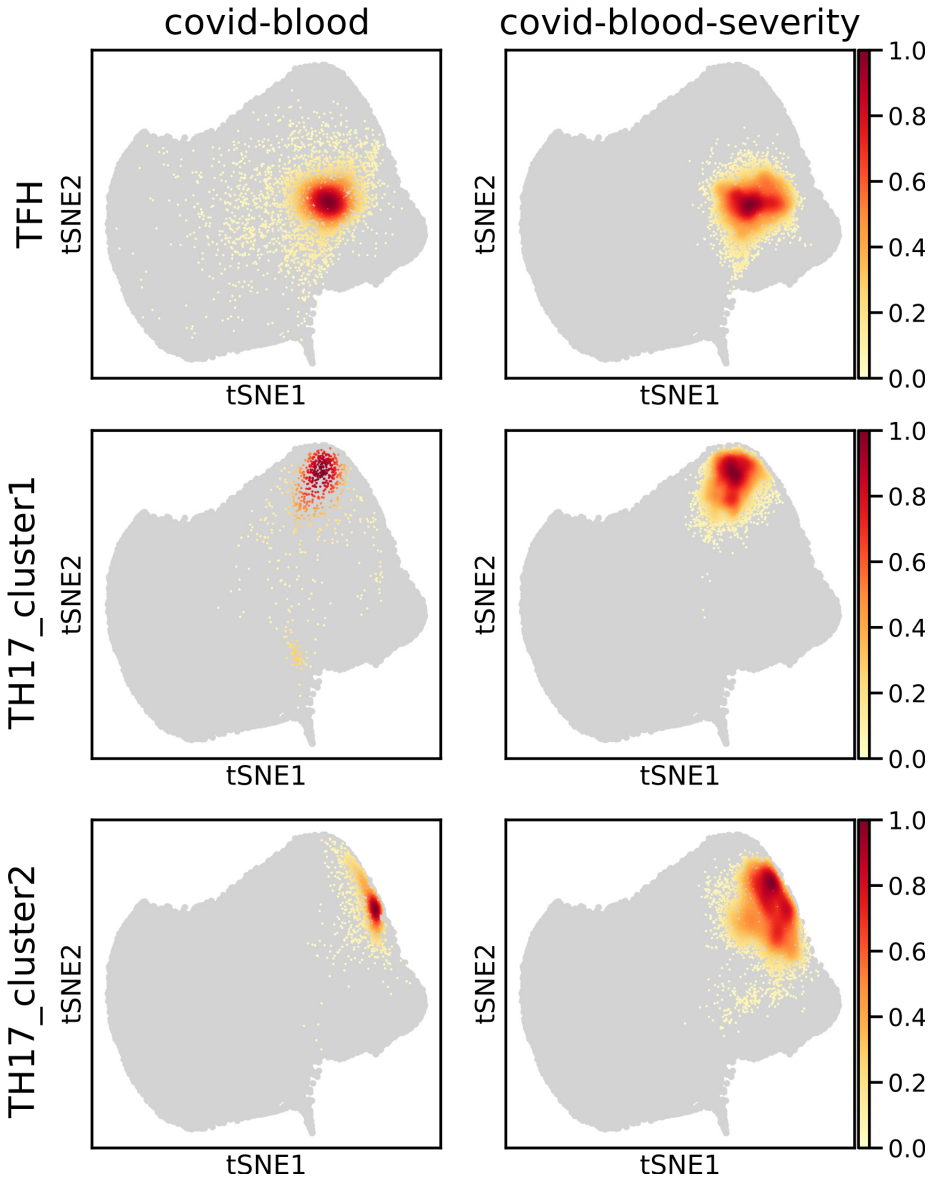


Figure S44: *t-SNE representation of three T helper cell clusters found in reconstructed covid-blood-hq and covid-blood-severity-hq data.* The t-SNE representation and clustering was computed on DISCERN reconstructed expression, using covid-blood-severity-lq and covid-blood-lq data as input and bulk-hq as reference. Cell type annotations for covid-blood-hq data are shown in the first column and covid-blood-severity-hq data in the second column. Cell densities are represented using white - low to red - high cell type density.

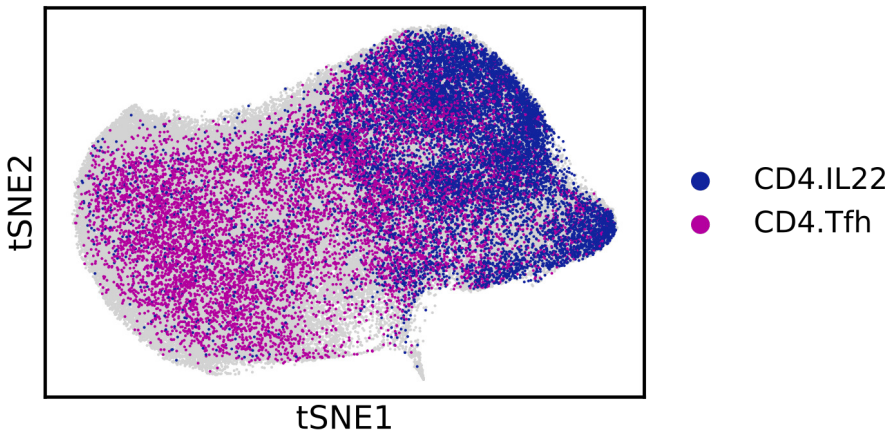


Figure S45: *tSNE representation of three T helper cell clusters of the reconstructed covid-blood and covid-blood-severity datasets.* The t-SNE representation and clustering was computed on DISCERN reconstructed expression, using covid-blood-severity-lq and covid-blood-lq data as input and bulk-hq as reference. Cell types are color-coded according to the covid-blood-severity-hq dataset. TFH cells from the original publication (CD4.Tfh) show significant overlap with naive $CD4^+$ T cells and $CD4^+$ $IL22^+$ cells (CD4.IL22) show marked overlap with TREG cells (compared with fig. S42).

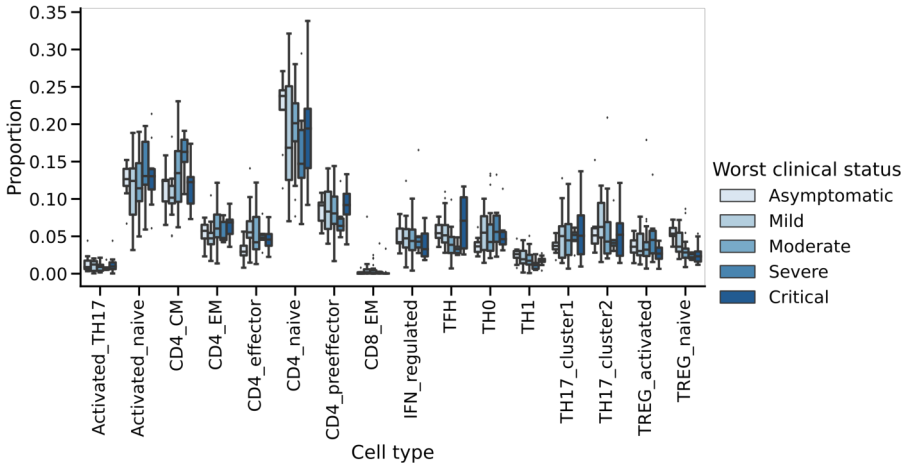


Figure S46: *Proportion of T cell subtypes in the covid-blood-severity-hq data grouped by disease severity.* The disease severity per patient is determined as the worst clinical status during hospitalization. Colors indicate disease severity, from light blue - asymptomatic to dark blue - critical. Boxplots represent median, quantiles, minimum, maximum, and potential outliers.

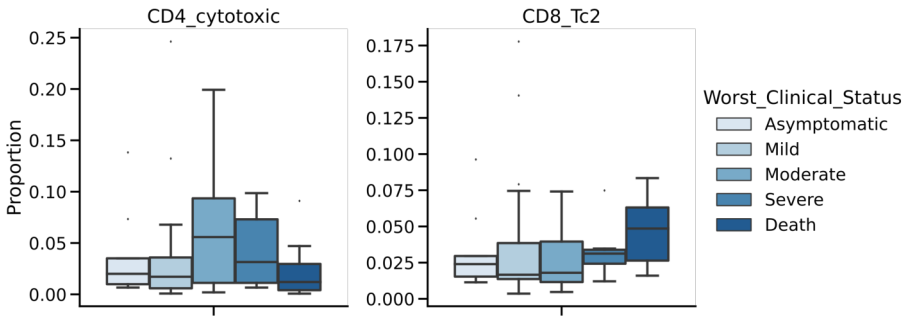


Figure S47: Proportion of 'unexpected' CD4_cytotoxic and CD8_Tc2 cell subtypes in the covid-blood-severity-hq data grouped by disease severity. The disease severity per patient is determined as the worst clinical status during hospitalization. Colors indicate disease severity, from light blue - asymptomatic to dark blue - critical. Boxplots represent median, quantiles, minimum, maximum, and potential outliers.

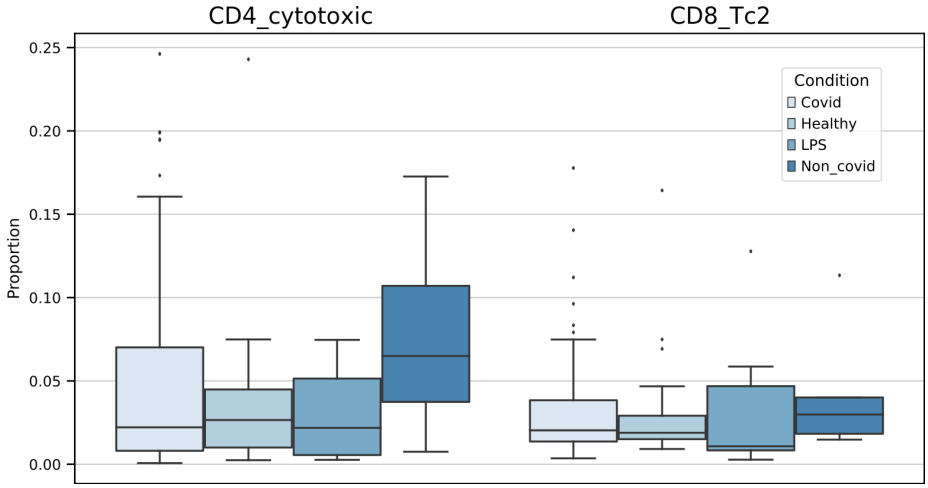


Figure S48: Proportion of 'unexpected' CD4_cytotoxic and CD8_Tc2 cell subtypes in the covid-blood-severity-hq data grouped by disease etiology. Colors indicate disease etiology, from light blue - COVID-19 to dark blue - non COVID-19. Boxplots represent median, quantiles, minimum, maximum, and potential outliers.

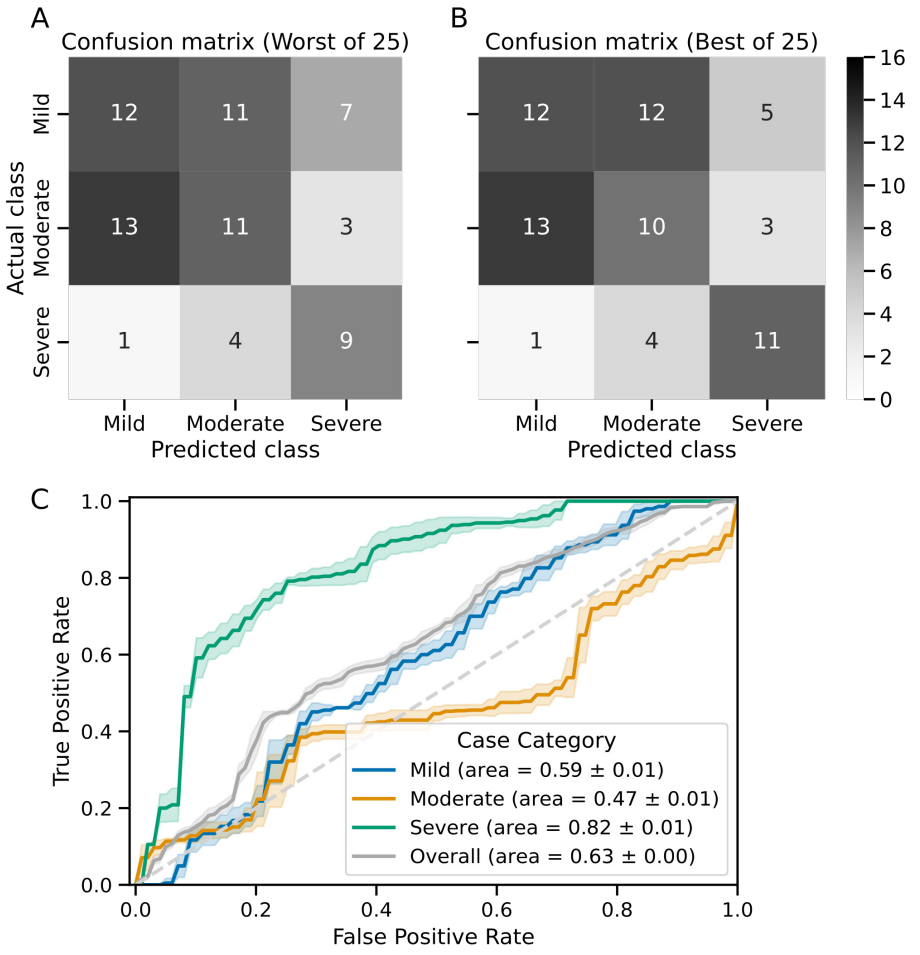


Figure S49: Disease-severity prediction using GBM classifiers trained on fractions of five T cell types of the covid-blood-severity-hq data. The five T cell types (CD8_EM, CD8_Tc2, TFH, TH17_cluster1, Treg_active) were selected using forward feature selection of the reconstructed covid-blood-severity-hq data. Confidence intervals were calculated using 25 runs of LOOCV. The disease category "critical" was combined with "severe" and "asymptomatic" with "mild". **A:** Confusion matrix for the worst run of LOOCV. **B:** Confusion matrix for the best run of LOOCV. **C:** ROC curve for the prediction of mild - blue, moderate - yellow, severe - green, and all categories (overall) - gray. Confidence intervals indicate one standard deviation.

Table S1: *Overview of all single cell and bulk sequencing datasets used in this study.* The table shows the dataset name, size of the dataset, the sequencing technology, cell types as annotated in the original study and a hyperlink to the publication.

Dataset	***Method***	***Cell Types***	***Publication or Download link***
pancreas (8569 cells)	SMARTSeq2, Fluidigm C1, CelSeq, CelSeq2, inDrops	alpha, beta, ductal, acinar, delta, gamma, activated_stellate, endothelial, quiescent_stellate, macrophage, mast, epsilon, schwann	[33]
difftec (31 021 cells)	10x Chromium v2, 10x Chromium v3, SMARTSeq2, Seq-Well, inDrops, Drop-seq, CelSeq2	Cytotoxic T cell, $CD4^+$ T cell, $CD14^+$ monocyte, B cell, Natural killer cell, Megakaryocyte, $CD16^+$ monocyte, Dendritic cell, Plasmacytoid dendritic cell, Unassigned	[34]
snRNA-seq & scRNA-seq (12 423 cells)	snRNA-seq and scRNA-seq using Chromium single-cell 3' v3	Epithelial cells, Macrophages, Hepatocytes, T cells, Endothelial cell, Fibroblasts, B cells, NK cells	`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4186980` `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4186974`

Table S1: *Overview of all single cell and bulk sequencing datasets used in this study continued.*

Dataset	***Method***	***Cell Types***	***Publication or Download link***
covid-lung (56 645 cells)	10X Genomics Chromium Single Cell 5'v1.1	CD8 T, TREG, CD4_CD8 proliferating, B cell, CD4_TCM, TRM1, TR1, CD8_TCM, T senescent, CD8_TEM, TEM17, T antiviral, alveolar MΦ, TRM17, M1, CD4_CD8 stressed TCM, CD4_TSCM, MAIT, Innate like, Neutrophils, doublets, CD4_CD8 lnc rich, aged Neutrophils, M1 HSP^+, Mast, DC, M1 Monoderived, M2 profibrotic, Epithelial, Neutrophil, Macrophage	[35]
covid-blood (83 709 cells)	10X Genomics Chromium Single Cell 5'v1.1	$CD3^+$ cells	[35]

Table S1: *Overview of all single cell and bulk sequencing datasets used in this study continued.*

Dataset	***Method***	***Cell Types***	***Publication or Download link***
citeseq (6592 cells)	10x Genomics Single Cell and CITE-seq	B cells, CD4 T cells, NK cells, $CD14^+$ Monocytes, $FCGR3A^+$ Monocytes, CD8 T cells	`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100866` `https://github.com/YosefLab/scVI-data/raw/master/pbmc_metadata.pickle`
bulk (9852 cells)	SMART-seq v4	Naive CD4, Memory CD4, TH1, TH2, TH17, Tfh, Fr. I nTreg, Fr. II eTreg, Fr. III T, Naive CD8, Memory CD8, CM CD8, EM CD8, TEMRA CD8, NK, Naive B, USM B, SM B, Plasmablast, DN B, CL Monocytes, Int Monocytes, NC Monocytes, mDC, pDC, Neutrophils, LDG	[36]

Table S1: *Overview of all single cell and bulk sequencing datasets used in this study continued.*

Dataset	***Method***	***Cell Types***	***Publication or Download link***
covid-blood-severity (636 836 cells)	10X Genomics Chromium Single Cell 5'v1.1	ASDC, B_exhausted, B_immature, B_malignant, B_naive, B_non-switched_memory, B_switched_memory, C1_CD16_mono, CD4.CM, CD4.EM, CD4.IL22, CD4.Naive, CD4.Prolif, CD4.Tfh, CD4.Th1, CD4.Th2', CD4.Th17, CD8.EM, CD8.Naive, CD8.Prolif, CD8.TE, CD14_mono, CD16_mono, CD83_CD14_mono, DC1, DC2, DC3, DC_prolif, HSC_CD38neg, HSC_CD38pos, HSC_MK, HSC_erythroid, HSC_myeloid, HSC_prolif, ILC1_3, ILC2, MAIT, Mono_prolif, NKT, NK_16hi, NK_56hi, NK_prolif, Plasma_cell_IgA, Plasma_cell_IgG, Plasma_cell_IgM, Plasmablast, Platelets, RBC, Treg, gdT, pDC	[37]

Table S1: *Overview of all single cell and bulk sequencing datasets used in this study continued.*

Dataset	***Method***	***Cell Types***	***Publication or Download link***
Kidney snRNA-seq & scRNA-seq (82 701 cells)	10x Genomics Chromium	None (not annotated)	`https://atlas.kpmp.org/repository/?facetTab=participants` Patients: 3010018, 3010034, 3210003, 3210034, 3310005, 3310006, 3410050, 3410184, 3410187

Table S2: *Detailed quality and batch information for all single cell and bulk sequencing datasets used in this study.* For each batch, the number of cells, the mean number of counts per cell, and the mean number of expressed genes per cell are listed. For the difftec dataset, the batch names were slightly adjusted. Their published batch names are written in brackets.

Dataset	**Batch**	**Number of cells**	**Mean number of counts per cell**	**Mean number of genes**
	smartseq2	2394	451021.4	6214.0
	fluidigmc1	638	1580155.4	8127.4
pancreas	celseq	2285	11161.1	3466.8

Table S2: *Detailed quality and batch information for all single cell and bulk sequencing datasets used in this study continued.*

Dataset	Batch	Number of cells	Mean number of counts per cell	Mean number of genes
difftec	celseq2	1004	23394.2	5274.9
	indrop	8569	5828.2	1887.2
	dropseq (pbmc1_Drop-seq)	3222	1282.0	676.0
	indrops (pbmc1_inDrops)	3222	566.3	362.4
	seqwell (pmbc1_Seq-Well)	3222	1035.3	567.2
	chromium-v3 (pbmc1_10x Chromium (v3))	3222	4891.3	1514.1
	chromium-v2 (pbmc1_10x Chromium (v2) A)	3222	2120.0	795.4
	chromium-v2B (pmbc1_10x Chromium (v2) B)	3222	2512.4	870.8
	smartseq2 (pbmc1_Smart-seq2)	253	385914.3	2434.6
	celseq2 (pbmc1_CEL-Seq2)	253	6057.3	2585.4
	dropseq-2 (pbmc2_Drop-seq)	3362	2141.0	977.7
	seqwell-2 (pbmc2_Seq-Well)	551	692.6	421.8

Table S2: *Detailed quality and batch information for all single cell and bulk sequencing datasets used in this study continued.*

Dataset	Batch	Number of cells	Mean number of counts per cell	Mean number of genes
	smartseq2-2 (pbmc2_Smart-seq2)	273	292924.3	2795.4
	celseq2-2 (pbmc2_CEL-Seq2)	273	5949.3	2556.6
	chromium-v2-2 (pbmc2_10x Chromium (v2))	3362	2860.7	1131.4
	indrops-2 (pbmc2_inDrops)	3362	1249.5	619.5
snRNA-seq	sn-lq	7260	2206.6	1308.7
& scRNA-seq	sc-hq	5163	4634.5	1214.6
covid-lung	Bacterial	14591	9627.2	1617.4
	SARS-CoV-2	42054	10284.4	1719.5
covid-blood	Bacterial	22199	5861.6	1703.0
	SARS-CoV-2	61510	5388.6	1700.7
citeseq	citeseq	6592	1391.8	797.8
bulk	bulk	9852	881440.6	13103.8
	cambridge	130637	4798.9	1485.9
covid-blood-severity	ncl	431733	3520.0	1276.1
	sanger	74466	3640.2	1445.1
Kidney snRNA-seq	kidney-lq (snRNA-seq)	52934	6532.8	2462.7
& scRNA-seq	kidney-hq (snRNA-seq)	29767	4449.6	1546.0

Table S3: Antibodies used in the CITE-seq experiments of the citeseq dataset (see table S1 & table S2).

Antibody	**Clone**	**Supplier**	**Target Protein**	**Target Gene**
CD3e	UCHT1	BioLegend, USA	CD3	*CD3E, CD3D*
CD19	HIB19	BioLegend, USA	CD19	*CD19*
CD4	RPA-T4	BioLegend, USA	CD4	*CD4*
CD8a	RPA-T8	BioLegend, USA	CD8	*CD8A*
CD56	MEM-188	BioLegend, USA	NCAM1	*NCAM1*
CD16	B73.1	BioLegend, USA	FCG3A	*FCGR3A*
CD11c	B-ly6	BD Pharmingen, USA	CD11c	*ITGAX*
CCR7	150603	RD Systems, USA	CCR7	*CCR7*
CCR5	J418F1	BioLegend, USA	CCR5	*CCR5*
CD34	581	BioLegend, USA	CD34	*CD34*
CD14	M5E2	BioLegend, USA	CD14	*CD14*
CD10	HI10a	BioLegend, USA	NEP	*MME, CD10*
CD45RA	HI100	BioLegend, USA	PTPRC, CD45RA	*PTPRC*
CD2	RPA-2.10	BioLegend, USA	CD2	*CD2*
CD57	H-NK1	BioLegend, USA	B3GA1, CD57	*B3GAT1*

Table S4: *Disease-severity prediction performance using GBM classifiers trained on T cell fractions.* Column one (3 - classes) and column two (2 - classes) displays the classification performance using cell type fractions obtained with reconstructed covid-blood-severity-hq data for three classes (mild, moderate, and severe) and two classes (mild and severe), respectively. The third column (2 - classes published) shows the classification performance for the fractions based on the originally published T cell annotations. All classifications were conducted with a GBM using 25 runs of LOOCV (confidence intervals) and forward feature selection. The T cell subtypes used by the GBM for column one are CD8_EM, CD8_Tc2, TFH, TH17_cluster1, Treg_active. For column two the features are CD4_CM, CD4_cytotoxic, CD4_naive, CD8_EM, CD8_effector. For column three CD4.CM, CD4.Tfh, CD8.EM, NKT, and Treg cells were used. It is striking to observe the strong increase in performance in the 2 class case between reconstructed cell type (column 2) and originally published (column 3) cell type information.

	3 - classes (mild, moderate, severe)	**2 - classes** (mild, severe)	**2 - classes published** (mild, severe)
F1-Score (Micro)	0.46 ± 0.01	$\mathbf{0.82 \pm 0.01}$	0.61 ± 0.01

F1-Score (Macro)	0.47 ± 0.01	$\mathbf{0.82 \pm 0.01}$	0.58 ± 0.01
AUROC	0.63 ± 0.00	$\mathbf{0.81 \pm 0.00}$	0.55 ± 0.01
Accuracy	0.46 ± 0.01	$\mathbf{0.82 \pm 0.01}$	0.61 ± 0.01

References

[1] S. Said, P. J. Kurtin, S. H. Nasr, R. P. Graham, S. Dasari, J. A. Vrana, S. Yasir, M. S. Torbenson, L. Zhang, T. Mounajjed, et al., Carboxypeptidase a1 and regenerating islet-derived 1α as new markers for pancreatic acinar cell carcinoma, Human Pathology 103 (2020) 120–126.

[2] M. Braun, The somatostatin receptor in human pancreatic β-cells, Vitamins & Hormones 95 (2014) 165–193.

[3] X. Wang, Y. He, Q. Zhang, X. Ren, Z. Zhang, Direct comparative analyses of 10x genomics chromium and smart-seq2, Genomics, proteomics & bioinformatics 19 (2) (2021) 253–266.

[4] A. Capone, E. Volpe, Transcriptional regulators of t helper 17 cell differentiation in health and autoimmune diseases, Frontiers in immunology 11 (2020) 348.

[5] R. Yagi, J. Zhu, W. E. Paul, An updated view on transcription factor gata3-mediated regulation of th1 and th2 cell differentiation, International immunology 23 (7) (2011) 415–420.

[6] E. Stolarczyk, G. M. Lord, J. K. Howard, The immune cell transcription factor t-bet: A novel metabolic regulator, Adipocyte 3 (1) (2014) 58–62.

[7] A. Chaudhry, R. M. Samstein, P. Treuting, Y. Liang, M. C. Pils, J.-M. Heinrich, R. S. Jack, F. T. Wunderlich, J. C. Brüning, W. Müller, et al., Interleukin-10 signaling in regulatory t cells is required for suppression of th17 cell-mediated inflammation, Immunity 34 (4) (2011) 566–578.

[8] Y. D. Mahnke, T. M. Brodie, F. Sallusto, M. Roederer, E. Lugli, The who's who of t-cell differentiation: human memory t-cell subsets, European journal of immunology 43 (11) (2013) 2797–2809.

[9] S. Yang, F. Liu, Q. J. Wang, S. A. Rosenberg, R. A. Morgan, The shedding of cd62l (l-selectin) regulates the acquisition of lytic activity in human tumor reactive t lymphocytes, PloS one 6 (7) (2011) e22560.

[10] M. Janakiram, J. M. Chinai, A. Zhao, J. A. Sparano, X. Zang, Hhla2 and tmigd2: new immunotherapeutic targets of the b7 and cd28 families, Oncoimmunology 4 (8) (2015) e1026534.

[11] G. Monaco, B. Lee, W. Xu, S. Mustafah, Y. Y. Hwang, C. Carré, N. Burdin, L. Visan, M. Ceccarelli, M. Poidinger, et al., Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types, Cell reports 26 (6) (2019) 1627–1640.

[12] A. S. Dejean, E. Joulia, T. Walzer, The role of eomes in human cd4 t cell differentiation: a question of context, European journal of immunology 49 (1) (2019) 38–41.

[13] Z. Yan, Y. Lijuan, W. Yinhang, J. Yin, X. Jiamin, W. Wei, P. Yuefen, H. Shuwen, Screening and analysis of rnas associated with activated memory cd4 and cd8 t cells in liver cancer, World journal of surgical oncology 20 (1) (2022) 1–15.

[14] R. J. Bonnal, G. Rossetti, E. Lugli, M. De Simone, P. Gruarin, J. Brummelman, L. Drufuca, M. Passaro, R. Bason, F. Gervasoni, et al., Clonally expanded eomes+ tr1-like cells in primary and metastatic tumors are associated with disease progression, Nature Immunology 22 (6) (2021) 735–745.

[15] T. Riaz, L. M. Sollid, I. Olsen, G. A. de Souza, Quantitative proteomics of gut-derived th1 and th1/th17 clones reveal the presence of cd28+ nkg2d-th1 cytotoxic cd4+ t cells, Molecular & Cellular Proteomics 15 (3) (2016) 1007–1016.

[16] S. L. Colpitts, N. M. Dalton, P. Scott, Interleukin-7 receptor (il7r) expression provides the potential for long-term survival of both cd62lhigh central memory t cells and th1 effector cells during leishmania major infection (96.6) (2009).

[17] K. Shadidi, T. Aarvak, J. Henriksen, J. Natvig, K. Thompson, The chemokines ccl5, ccl2 and cxcl12 play significant roles in the migration of th1 cells into rheumatoid synovial tissue, Scandinavian journal of immunology 57 (2) (2003) 192–198.

[18] K. Eshima, K. Misawa, C. Ohashi, K. Iwabuchi, Role of t-bet, the master regulator of th1 cells, in the cytotoxicity of murine cd4+ t cells, Microbiology and immunology 62 (5) (2018) 348–356.

[19] Y. Deng, Z. Huang, C. Zhou, J. Wang, Y. You, Z. Song, M. Xiang, B. Zhong, F. Hao, Gene profiling involved in immature cd4+ t lymphocyte responsible for systemic lupus erythematosus, Molecular immunology 43 (9) (2006) 1497–1507.

[20] D. Matza, A. Badou, K. S. Kobayashi, K. Goldsmith-Pestana, Y. Masuda, A. Komuro, D. McMahon-Pratt, V. T. Marchesi, R. A. Flavell, A scaffold protein, ahnak1, is required for calcium signaling during t cell activation, Immunity 28 (1) (2008) 64–74.

[21] Y. Bordon, Tox for tired t cells, Nature reviews Immunology 19 (8) (2019) 476–476.

[22] M. H. Sofi, J. Heinrichs, M. Dany, H. Nguyen, M. Dai, D. Bastian, S. Schutt, Y. Wu, A. Daenthanasanmak, S. Gencer, et al., Ceramide synthesis regulates t cell activity and gvhd development, JCI insight 2 (10) (2017).

[23] T. Ye, J. Feng, M. Cui, J. Yang, X. Wan, D. Xie, J. Liu, Lncrna miat services as a noninvasive biomarker for diagnosis and correlated with immune infiltrates in breast cancer, International journal of women's health 13 (2021) 991.

[24] A. Machicote, S. Belén, P. Baz, L. A. Billordo, L. Fainboim, Human cd8+ hla-dr+ regulatory t cells, similarly to classical cd4+ foxp3+ cells, suppress immune responses via pd-1/pd-l1 axis, Frontiers in immunology (2018) 2788.

[25] S. Nakamura, Y. Nagata, L. Tan, T. Takemura, K. Shibata, M. Fujie, S. Fujisawa, Y. Tanaka, M. Toda, R. Makita, et al., Transcriptional repression of cdc25b by ier5 inhibits the proliferation of leukemic progenitor cells through nf-yb and p300 in acute myeloid leukemia, PLoS One 6 (11) (2011) e28011.

[26] S. Paul, G. Lal, Regulatory and effector functions of gamma–delta ($\gamma\delta$) t cells and their therapeutic potential in adoptive cellular therapy for cancer, International journal of cancer 139 (5) (2016) 976–985.

[27] C.-M. Scarlata, C. Celse, P. Pignon, M. Ayyoub, D. Valmori, Differential expression of the immunosuppressive enzyme il4i1 in human induced aiolos+, but not natural helios+, foxp3+ treg cells, European Journal of Immunology 45 (2) (2015) 474–479.

[28] S. Cerboni, U. Gehrmann, S. Preite, S. Mitra, Cytokine-regulated th17 plasticity in human health and diseases, Immunology 163 (1) (2021) 3–18.

[29] M. Vocanson, A. Rozieres, A. Hennino, G. Poyet, V. Gaillard, S. Renaudineau, A. Achachi, J. Benetiere, D. Kaiserlian, B. Dubois, et al., Inducible costimulator (icos) is a marker for highly suppressive antigen-specific t cells sharing features of th17/th1 and regulatory t cells, Journal of Allergy and Clinical Immunology 126 (2) (2010) 280–289.

[30] X. Shi, Z. Qu, L. Zhang, N. Zhang, Y. Liu, M. Li, J. Qiu, Y. Jiang, Increased ratio of icos+/pd-1+ follicular helper t cells positively correlates with the development of human idiopathic membranous nephropathy, Clinical and Experimental Pharmacology and Physiology 43 (4) (2016) 410–416.

[31] B. Bengsch, B. Seigel, T. Flecken, J. Wolanski, H. E. Blum, R. Thimme,

Human th17 cells express high levels of enzymatically active dipeptidylpeptidase iv (cd26), The Journal of Immunology 188 (11) (2012) 5438–5447.

[32] M. M. Klicznik, P. A. Morawski, B. Höllbacher, S. R. Varkhande, S. J. Motley, L. Kuri-Cervantes, E. Goodwin, M. D. Rosenblum, S. A. Long, G. Brachtl, et al., Human cd4+ cd103+ cutaneous resident memory t cells are found in the circulation of healthy individuals, Science immunology 4 (37) (2019) eaav8995.

[33] S. Lab, panc8.SeuratData: Eight Pancreas Datasets Across Five Technologies, r package version 3.0.2 (2019).

[34] J. Ding, X. Adiconis, S. K. Simmons, M. S. Kowalczyk, C. C. Hession, N. D. Marjanovic, T. K. Hughes, M. H. Wadsworth, T. Burks, L. T. Nguyen, et al., Systematic comparison of single-cell and single-nucleus rna-sequencing methods, Nature biotechnology 38 (6) (2020) 737–746.

[35] Y. Zhao, C. Kilian, J.-E. Turner, L. Bosurgi, K. Roedl, P. Bartsch, A.-C. Gnirck, F. Cortesi, C. Schultheiß, M. Hellmig, et al., Clonal expansion and activation of tissue-resident memory-like th17 cells expressing gm-csf in the lungs of patients with severe covid-19, Science Immunology 6 (56) (2021) eabf6692.

[36] M. Ota, Y. Nagafuchi, H. Hatano, K. Ishigaki, C. Terao, Y. Takeshima, H. Yanaoka, S. Kobayashi, M. Okubo, H. Shirai, et al., Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases, Cell 184 (11) (2021) 3006–3021.

[37] E. Stephenson, G. Reynolds, R. A. Botting, F. J. Calero-Nieto, M. D. Morgan, Z. K. Tuong, K. Bach, W. Sungnak, K. B. Worlock, M. Yoshida, et al., Single-cell multi-omics analysis of the immune response in covid-19, Nature medicine 27 (5) (2021) 904–916.