Sort-Seq: immune repertoire-based scRNA-Seq systematization

Kriukova V.V.[1,2,*], Lukyanov D.K.[2,*], Shagina I.A.[2,3,*], Sharonov G.V.[2,3,4], Lupyr K.R.[3], Blagodatskikh K.A.[3], Staroverov D.B.[2,3], Ladell K.[5], Miners K.L. [5], Salnikova M.A.[2], Nikolaev R.V.[6], Shelikhova L.[6], Maschan M.A.[6], Britanova O.V.[2,3], Franke A.[1], Price D.A.[5,7], Chudakov D.M.[2,3,8,9]

[1] Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany
[2] Genomics of Adaptive Immunity Department, Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia
[3] Institute of Translational Medicine, Pirogov Russian National Research Medical University, Moscow, Russia
[4] Institute of Experimental Oncology and Biomedical Technologies, Privolzhsky Research Medical University, Nizhny Novgorod, Russia
[5] Division of Infection and Immunity, Cardiff University School of Medicine, University Hospital of Wales, Cardiff, United Kingdom
[6] Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia.
[7] Systems Immunity Research Institute, Cardiff University School of Medicine, University Hospital of Wales, Cardiff, United Kingdom
[8] Central European Institute of Technology, Brno, Czech Republic
[9] Abu Dhabi Stem Cell Center, Al Muntazah, United Arab Emirates

Abstract

The functional programs chosen by B and T cell clones fundamentally determine the architecture of immune response to distinct challenges. Advances in scRNA-Seq have improved our understanding of the diversity and stability of these programs, but it has proven difficult to link this information with known lymphocyte subsets. Here, we introduce Sort-Seq, an immune repertoire-based method that allows exact positioning of phenotypically defined lymphocyte subsets within scRNA-Seq data. Sort-Seq outperformed CITE-Seq for accurate mapping of the classical CD4$^+$ T helper (Th) cell subsets (Th1, Th1-17, Th17, Th22, Th2a, Th2, and Treg), offering a more powerful approach to the surface phenotype-based scRNA-Seq classification of adaptive lymphocyte subpopulations. Using integrated scRNA-Seq, Sort-Seq, and CITE-Seq data from 122 donors, we provide a comprehensive Th cell scRNA-Seq reference map. Exploration of this dataset revealed the low plasticity and extreme sustainability of the Th17, Th22, Th2, and Th2a cell programs over years. We also develop Cultivation-based Antigen-specific T cell identificatoR in Replicates (CultivAToRR), which identified >80 SARS-CoV-2-specific CD4$^+$ TCRβ clonotypes in a single donor across a wide frequency range. We complemented these results with frequency-based capturing of COVID-19-responsive clonotypes and screening against known SARS-CoV-2-specific TCRs. Positioning within the annotated scRNA-Seq map revealed functional subtypes of Th cell clones involved in primary and secondary responses against SARS-CoV-2. The ability to capture low-frequency antigen-specific T cell clones in combination with Sort-Seq-based scRNA-Seq annotation creates an integral pipeline that links challenge-responsive clones with their exact functional subtypes, providing a solid foundation for investigating T cell roles in healthy and pathological immune responses and vaccine development.

Introduction

Adaptive immunity controls pathogen clearance, tissue homeostasis, and vaccine efficacy. These processes depend largely on antigen immunogenicity, the functionality of cognate naive and/or memory B and T cells, and the type of acquired immune response.

Clonal populations of $CD4^+$ T helper (Th) cells essentially orchestrate the course of an immune response via specific interactions with peptide epitopes presented in complex with major histocompatibility class II (MHCII) molecules. Their functional and antigen-specific diversity allows them to guide both classical (B cells, dendritic cells, macrophages) and non-classical (endothelial cells, epithelial cells, granulocytes) antigen-presenting cells to optimize effector functionality[1, 2, 3]. As a consequence, inappropriate Th responses to certain antigens have been associated with impaired pathogen clearance[4, 5, 6], inefficient responses to vaccination[7], acute and chronic hypersensitivity and inflammation[8, 9, 10], inflammaging[11], autoimmunity[12, 13, 14, 15], and cancer[16]. Accordingly, when investigating T cell responses, it is critical not only to quantify the magnitude of the antigen-specific T cell clonal expansion but also to understand functional programs, and related phenotypes of the responding and memory helper T cells.

Single-cell RNA-Seq (scRNA-Seq) techniques are shedding light on the true diversity of Th cell programs[17, 18, 19], among which classical subsets described on the basis of cytokine release profiles and patterns of surface marker expression have to find their place. However, the expression of transcripts encoding characteristic surface markers is often low in scRNA-Seq datasets, with indirect correlations between mRNA abundance and protein density[20, 21]. This problem can be overcome to some extent by the incorporation of protein-level expression data into single-cell experiments, a task that is essentially implemented in CITE-Seq technology, which makes use of barcoded antibodies directed against markers of interest expressed on the cell surface[22, 23, 24].

Here, we report an alternative approach to the surface protein-based mapping of lymphocyte subsets to scRNA-Seq data, termed Sort-Seq. This approach makes use of immune repertoires from sorted lymphocyte subsets, which are then mapped to scRNA-Seq+scBCR/TCR-Seq datasets obtained from the same donors, employing natural barcodes in the form of sequence-defined BCRs/TCRs. We demonstrate that this method outcompetes CITE-Seq for the accurate mapping of phenotypically defined Th subsets. We next integrate scRNA-Seq, Sort-Seq, and CITE-Seq outputs of 122 donors to provide a Th scRNA-Seq reference dataset and employ it to analyze long-term Th subsets stability and plasticity at the scRNA-Seq level.

In addition, dataset and improved method for the identification of low-frequency antigen-specific T cell clones among PBMCs, termed Cultivation-based Antigen-specific T cell identificatoR in Replicates (CultivAToRR).

These approaches in combination allowed the identification of functional subsets of the SARS-CoV-2-responsive Th cell clonotypes via scRNA-Seq. Collectively, our results validate a seamless integration of wet laboratory data and bioinformatic tools, enabling unique insights into the true nature of the adaptive immune system.

Results

Sort-Seq

Functional subtypes of human helper T cells, such as Th1, Th2, Th17, Th1-17, Th22, Tfh, and Treg, are classically distinguished based on the surface markers, such as CD127, CD25, CCR10, CXCR5, CXCR3, CCR6, CCR4, and CRTh2 (e.g. **Fig. 1**, middle panel).

Here we conceptualized the precise annotation of Th cells in scRNA-Seq data powered by overlapping scTCR-Seq repertoires and TCR repertoires of FACS-sorted Th cell subsets (**Fig**. **1**). We hypothesized that, due to relatively high phenotypic stability and low plasticity between the helper T cell subsets[25], this approach would accurately map phenotypically defined subsets within scRNA-Seq data.

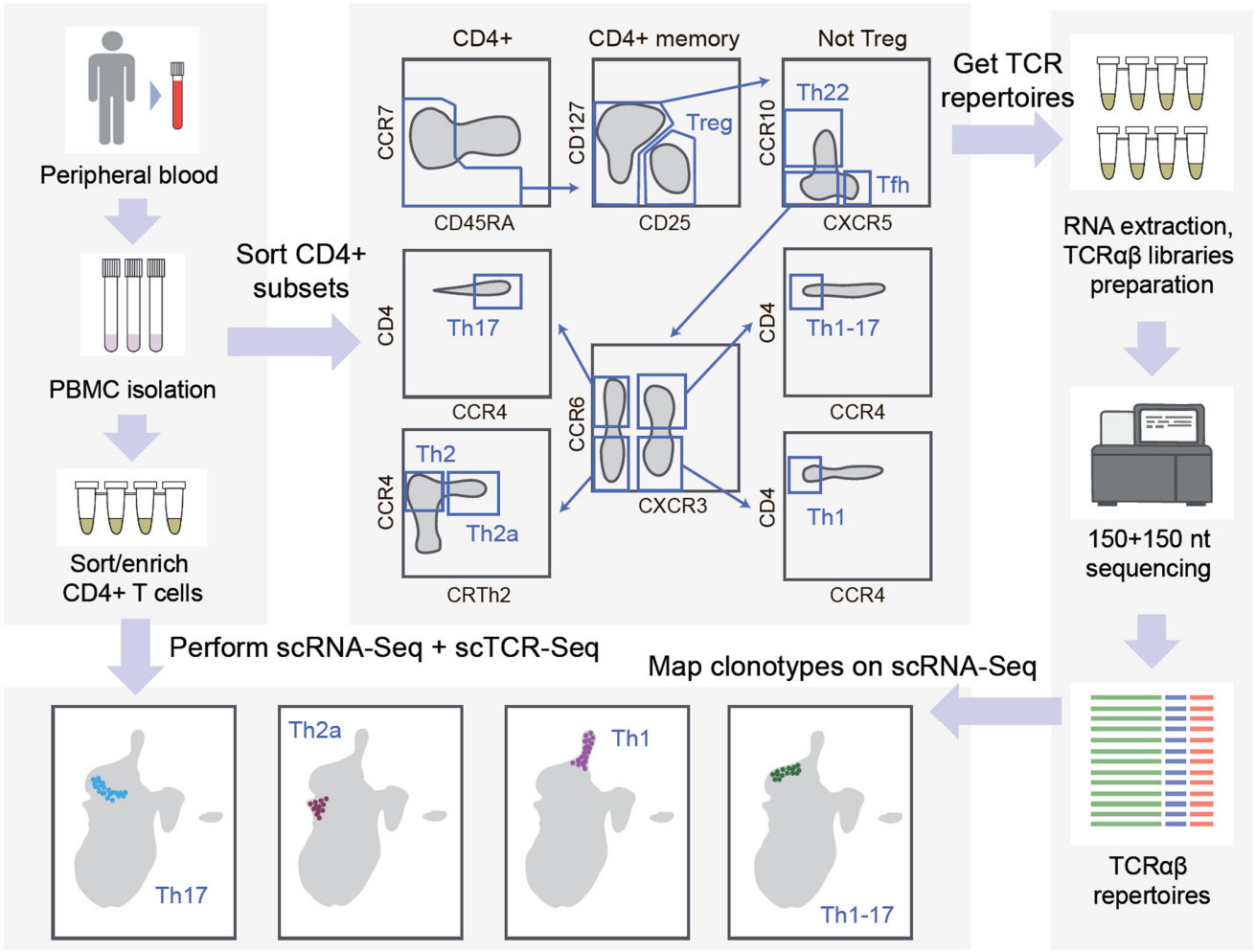


Figure 1. Sort-Seq concept. T cell or B cell subsets of interest are stained and FACS-sorted using conventional surface markers from a sample of interest, such as PBMCs, lymph nodes, or tumor tissue. Immune receptor repertoires are obtained for each sorted subset and mapped to the scRNA-Seq data obtained for the same patient or animal. Complementary scTCR-Seq or scBCR-Seq data is used to position clones of FACS-sorted subsets within scRNA-Seq. The figure is built to illustrate the content of the current work. The same concept is applicable for mapping various T cell and B cell subsets in human and model animal scRNA-Seq data.

To verify this concept, named Sort-Seq, we exploited previously obtained TCRα and TCRβ repertoires of Th1, Th17, Th1-17, Th22, Th2, Th2a, Treg, and Tfh CD4+ T cell subsets thoroughly sorted from peripheral blood of healthy donors[25]. Three participants of the latter study were available for repeated blood donation at the moment of the current study. We performed scRNA-Seq and scTCR-Seq profiling from their sorted bulk effector/memory CD4+ T cells (gated as CD4+, NOT CCR7+CD45RA+ cells to deplete naive T cells). The obtained scRNA-Seq data has then been integrated with the CD4+ scRNA-Seq reported in Ref. 26 in order to: 1) increase the power of the downstream analysis such as clustering and UMAP visualization on a larger number of donors and 2) enable comparison of our approach with the CITE-Seq method employed in the latter work. This resulted in a reference dataset composed of 147,677 cells (**Fig. 2a**), without notable donor-specific or study-specific batch effects (**Supplementary Fig**. **1a**).

In order to link scRNA-Seq clusters to the classic Th phenotypes, we mapped TCR clonotypes of the sorted Th subsets using scTCRs as natural barcodes. Remarkably, T cell clones from each of the sorted Th subpopulations formed clearly defined spots on the scRNA-Seq UMAP (**Fig. 2b,c,d, Supplementary Fig**. **1b,c**).

Of note, most of the scRNA-Seq clusters showed low clonality in the scTCR-Seq data, thus excluding clonal biases in the Sort-Seq-based annotation (**Supplementary Fig**. **1e**). High clonality was exclusively observed for the cluster annotated as Temra cytotoxic Th1 subset, which is in line with the previous reports[27].

We next attempted to annotate the same surface-marker defined Th subsets within scRNA-Seq using CITE-Seq. For this, we used CITE-Seq data from Ref. 26, and performed sequential *in silico* sorting-like gating as shown in **Supplementary Fig**. **2**.

Altogether, Sort-Seq resulted in much more accurate cluster annotations compared to CITE-Seq (**Fig**. **2b-d**, **Supplementary Figs**. **1**,**3**). Both approaches yielded similar accuracy of cluster annotations for certain subsets, such as Tregs, defined by $CD25^{high}CD127^{low}$ gating. CITE-Seq based annotation also readily distinguished Th17 and Th22 subsets. At the same time, CITE-Seq failed to annotate Th1 (gated as NOT $CD25^{high}CD127^{low}$, CCR10-CCR6-CXCR3+CCR4-) and poorly annotated Th2 and Th2a subsets (gated as NOT $CD25^{high}CD127^{low}$, CCR10-CCR6-CXCR3-CCR4+ and CRTh2- or CRTh2+, respectively), presumably due to poor surface expression of the corresponding markers. We concluded that, due to the high accuracy of FACS cell sorting, Sort-Seq approach is preferable for scRNA-Seq cluster annotation to find accurate correspondence with the studied T cell subpopulations of interest.

Based on the integrated data of scRNA-Seq, Sort-Seq, and CITE-Seq outputs of 122 donors and taking into account findings and considerations of Sakaguchi and colleagues[18], we propose a comprehensive Th scRNA-Seq reference map with 16 major Th clusters (**Fig. 2a, Supplementary Table 1**). Owing to Sort-Seq-based classification, it for the first time accurately positions Th2 and Th2a subsets. Classic Th1, Th17, Th22, as well as Temra cytotoxic[27, 28, 29], but not Th1-17 (see below) subsets also find their exact locations within the reference scRNA-Seq dataset. “Eff-mem IFN response” cluster is nearly absent in healthy donors’ blood but is well-detectable in moderate and severe COVID patients (**Supplementary Fig. 4**), probably representing the typical behavior of Th cells in acute viral infection[30]. Tfh subset is found in two dissimilar clusters differing in the expression of CXCR5 and CXCR3. This points at functional heterogeneity of Tfh subset[31], which warrants a deeper focused investigation.

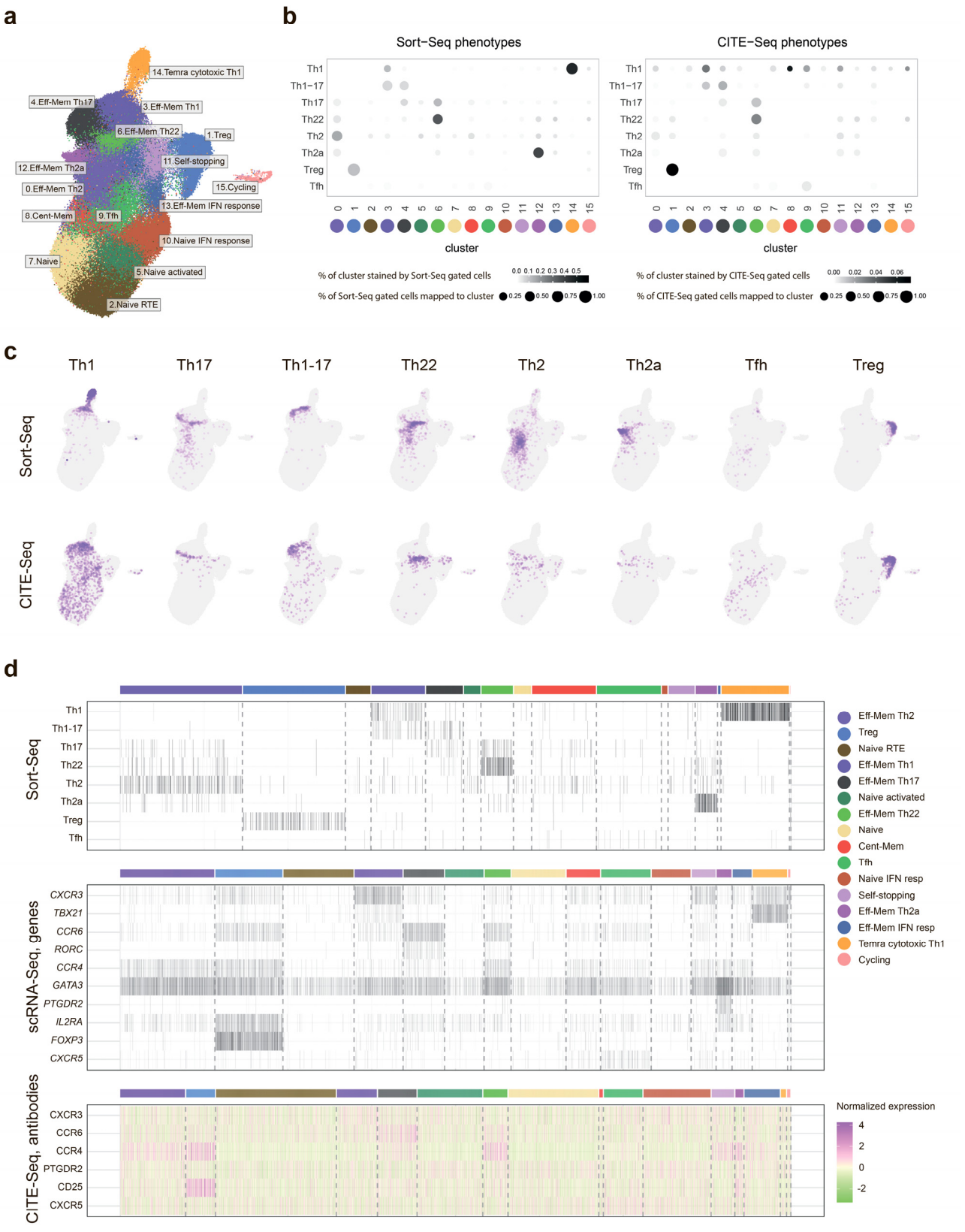

a
14.Temra cytotoxic Th1
4.Eff-Mem Th17
3.Eff-Mem Th1
6.Eff-Mem Th22
1.Treg
12.Eff-Mem Th2a
11.Self-stopping
0.Eff-Mem Th2
15.Cycling
13.Eff-Mem IFN response
8.Cent-Mem
9.Tfh
10.Naive IFN response
7.Naive
5.Naive activated
2.Naive RTE
b
Sort−Seq phenotypes
CITE−Seq phenotypes
Th1
Th1−17
Th17
Th22
Th2
Th2a
Treg
Tfh
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
cluster
% of cluster stained by Sort-Seq gated cells
0.0 0.1 0.2 0.3 0.4 0.5
% of Sort-Seq gated cells mapped to cluster
0.25 0.50 0.75 1.00
% of cluster stained by CITE-Seq gated cells
0.00 0.02 0.04 0.06
% of CITE-Seq gated cells mapped to cluster
0.25 0.50 0.75 1.00
c
Th1
Th17
Th1-17
Th22
Th2
Th2a
Tfh
Treg
Sort-Seq
CITE-Seq
d
Sort-Seq
Th1
Th1-17
Th17
Th22
Th2
Th2a
Treg
Tfh
scRNA-Seq, genes
CXCR3
TBX21
CCR6
RORC
CCR4
GATA3
PTGDR2
IL2RA
FOXP3
CXCR5
CITE-Seq, antibodies
CXCR3
CCR6
CCR4
PTGDR2
CD25
CXCR5
Eff-Mem Th2
Treg
Naive RTE
Eff-Mem Th1
Eff-Mem Th17
Naive activated
Eff-Mem Th22
Naive
Cent-Mem
Tfh
Naive IFN resp
Self-stopping
Eff-Mem Th2a
Eff-Mem IFN resp
Temra cytotoxic Th1
Cycling
Normalized expression
4
2
0
-2

Figure 2. Mapping the classic Th subsets with scRNA-Seq. a. UMAP visualization of the reference scRNA-Seq dataset of peripheral blood Th cells. Dataset built via Seurat integration of publicly available and our scRNA-Seq data. The proposed classification is based on previous knowledge and findings of the current work. **b**. Dot plots summarizing the positioning of sorted (Sort-Seq) and *in silico* gated (CITE-Seq) Th subsets within the scRNA-Seq clusters shown on the panel (c). For normalization, we used 20,000 randomly selected scRNA-Seq cells with matched CITE-Seq or Sort-Seq data for each plot. Dot intensity shows the stained proportion of the scRNA-Seq cluster. Dot size shows the proportion of Sort-Seq-identified or proportion of *in silico* CITE-Seq-based gated scRNA-Seq cells mapped to the scRNA-Seq cluster. **c**. UMAP plots showing the localization of Sort-Seq and CITE-Seq defined subsets. TCRβ clonotypes were used to define Th subsets in Sort-Seq method. Expression of the surface markers was used to gate Th subsets in the CITE-Seq-based annotation. The color intensity in Sort-Seq is proportional to the clonal frequencies in the original sorted Th bulk TCRβ repertoires. **d**. Heatmap summarizing the positioning of Sort-Seq subsets, scRNA-Seq expression of characteristic genes, and CITE-Seq signal for the corresponding surface proteins. For normalization, equal numbers of 10,000 scRNA-Seq cells were randomly selected from Sort-Seq and CITE-Seq experimental datasets. Each tile of the heatmap represents one cell. The color intensity in the scRNA-Seq plot visualizes the gene expression in a cell (white color depicts zero expression). Protein expression measured by CITE-Seq was Z-scored, and all values exceeding 99.5 percentile were trimmed.

Phenotypic versus intrinsic program plasticity of Th cells

To evaluate the relationships and cross-subset plasticity of T cell clones, we analyzed cluster stability at different clustering resolution levels (**Supplementary Fig. 5a**) and clonal intersections between the clusters (**Supplementary Fig**. **5b**). We previously suggested notable plasticity between Th22/Th17, Th17/Th2, and Th2/Th2a subsets based on the corresponding intersections of sorted Th cell subset repertoires[25]. However, the new data on clonal overlaps and cluster stability at various resolutions in the current study prompts us to reconsider these interpretations.

Indeed, T cell clones sorted as classic Th17 subset (gated as NOT $CD25^{high}CD127^{low}$, CCR10-CCR6+CXCR3-CCR4+) are found in both Th17 and Th22 scRNA-Seq clusters. However, clonal overlap between these clusters is low while the stability of both clusters is high (**Supplementary Figs**. **5a**,**b**). Furthermore, clones sorted as classic Th22 (gated as NOT $CD25^{high}CD127^{low}$, CCR10+) are almost exclusively found in Th22 scRNA-Seq cluster. This indicates that while certain Th22 clones may exhibit similar surface markers to Th17 clones due to the phenotypic plasticity, their intrinsic programs remain stable. In other words, our data suggests that Th17 clones mainly remain Th17, and Th22 mainly remain Th22. The independent origin of human Th22 clones was initially suggested[32] and reported in mouse models[33, 34]. Distinct TCR repertoire features also suggested the existence of clonally discrete Th22 subset[25]. According to this logic, the population classically sorted by phenotypic markers and described as Th17, represents the mixture of *bona fide* Th17 and *bona fide* Th22 cells. In contrast, classically sorted CCR10+ Th22 subsets almost fully coincide with the corresponding scRNA-Seq cluster, representing an almost pure population of uniformly programmed T cells (**Supplementary Table 2**).

Similarly, based on this data we suggest the self-standing nature of Th17 and Th1 subsets. T cell clones sorted as Th1-17 subset were localized in both of these scRNA-Seq clusters, but clonal overlap between Th17 and Th1 clusters is low, and stability of both clusters is high (**Supplementary Fig**. **5a**,**b**). This data suggests the existence of stable clonal T cell populations that carry either Th17 or Th1 programs, and that only their phenotypic features but not the true underlying program could change. The Th1-17 subset, in this paradigm, does not exist as a stable combo of the two major programs of Th cells (**Supplementary Table 2**).

Th2a cluster appears to be self-standing: it is stable at different clustering resolutions, and is clonally well separated from Th2 cluster (**Fig. 2**, **Supplementary Figs**. **5a**,**b**). In contrast, Th1 and Temra cytotoxic Th1 clusters demonstrate notable clonal exchange (**Supplementary Fig**. **5b**), indicating that the former could convert into the latter, as some of the recent works suggest[35, 36].

In the proposed scRNA-Seq classification shown in **Fig**. **2a**, we give priority to the cells partitioning into stable scRNA-Seq clusters (**Supplementary Fig**. **5a**), while Sort-Seq data are used to match classical surface phenotype-based sorted T cell subsets with those clusters.

Of note, T cell repertoires of the sorted subsets were obtained in 2018, while scRNA-Seq experiment was performed in 2022. Despite the 4-year distance, Sort-Seq mapped sorted clones to clearly defined positions, generally limited to one or two neighboring scRNA-Seq clusters (**Fig**. **2**, **Supplementary Fig**. **1**), indicating long-term stability of a program once chosen by an activated T cell clone.

COVID-specific TCRs discovery

To test Sort-Seq capability to determine the phenotypes of antigen-responsive T cells, we focused on searching SARS-CoV2-specific CD4+ T cell clones in donor D11, who was vaccinated with adenoviral SARS-CoV2 vaccine in Dec2020/Jan2021 and got two registered, PCR-confirmed SARS-CoV2 infections in April 2021 and January 2022 (**Supplementary Note 1, Supplementary Fig**. **6, Supplementary Table 3**).

To identify TCR variants responding to SARS-CoV2 antigens for this donor, we implemented several approaches:

A. Cultivation-based Antigen-specific T cell identificatoR in Replicates method (CultivAToRR).

In this assay, we relied on the concept of antigen-specific T cell expansion, followed by TCR repertoire profiling[37, 38]. This approach, either used in bulk[38] or focused on dividing T cells that progressively dilute cytoplasmic dyes such as carboxyfluorescein succinimidyl ester (CFSE)[37], allows the identification of T cell clones that selectively proliferate in presence of particular antigens and consequently become enriched within the TCR repertoire.

Notably, the random nature of cell sampling and further interactions between T cells and potent antigen-presenting cells in a culture, as well as the varying proliferation potential of individual memory T cells, substantially limits applicability of such approaches to relatively large clonal T cell expansions[38]. To account for the sampling issue, we exploited our previous experience with time-lapse T cell clonal tracking[39, 40] and identification of T cell clones that are locally expanded in tumor sections[41]. In both scenarios, we obtained independent biological (at the level of cells) replicates for the TCR profiling, followed by identification of reproducibly expanded clones, where we also accounted for the sampling noise defined as variability between the replicates.

Here, we cultured CFSE-labeled D11 PBMCs in the presence of three distinct mixtures of antigenic peptides, each in three independent replicates. After 7 days, divided ($CFSE^{low}$) CD4+ and CD8+ subsets were sorted from each replica, yielding 224-7,322 T cells, followed by the sensitive TCR library preparation technique. Obtained TCRβ CDR3 repertoires have been then analyzed with the pipeline based on EdgeR, a widely used method to identify differentially expressed genes in RNA-Seq[42]. Here, the same EdgeR algorithm allowed us to identify TCR CDR3 variants that were reproducibly enriched within repertoires of

replicates cultured in the presence of the corresponding mixture of antigenic peptides, compared to the other mixtures of antigens (following the same logic as in Ref. 39). Those antigenic peptide mixtures included Miltenyi peptivators comprising the overlapping 15mers of 1) SARS-CoV-2 spike protein (S) and 2) pooled membrane glycoprotein (M) + nucleoprotein (N), and 3) a mixture of irrelevant control peptides. See **Fig**. **3a** for the overview of the CultivAToRR approach.

EdgeR analysis of obtained repertoires revealed 93 TCRβ clonotypes that specifically and reproducibly proliferated *in vitro* in the presence of S (28 CD4 and 4 CD8 clonotypes) or M+N (59 CD4 and 2 CD8 clonotypes) peptivators. This bias towards CD4+ T cells could be explained by the nature of 15mers peptivators, since proper processing towards shorter 9-10mers that are presented by MHCI to CD8 T cells may not always proceed efficiently, the question that requires further investigation.

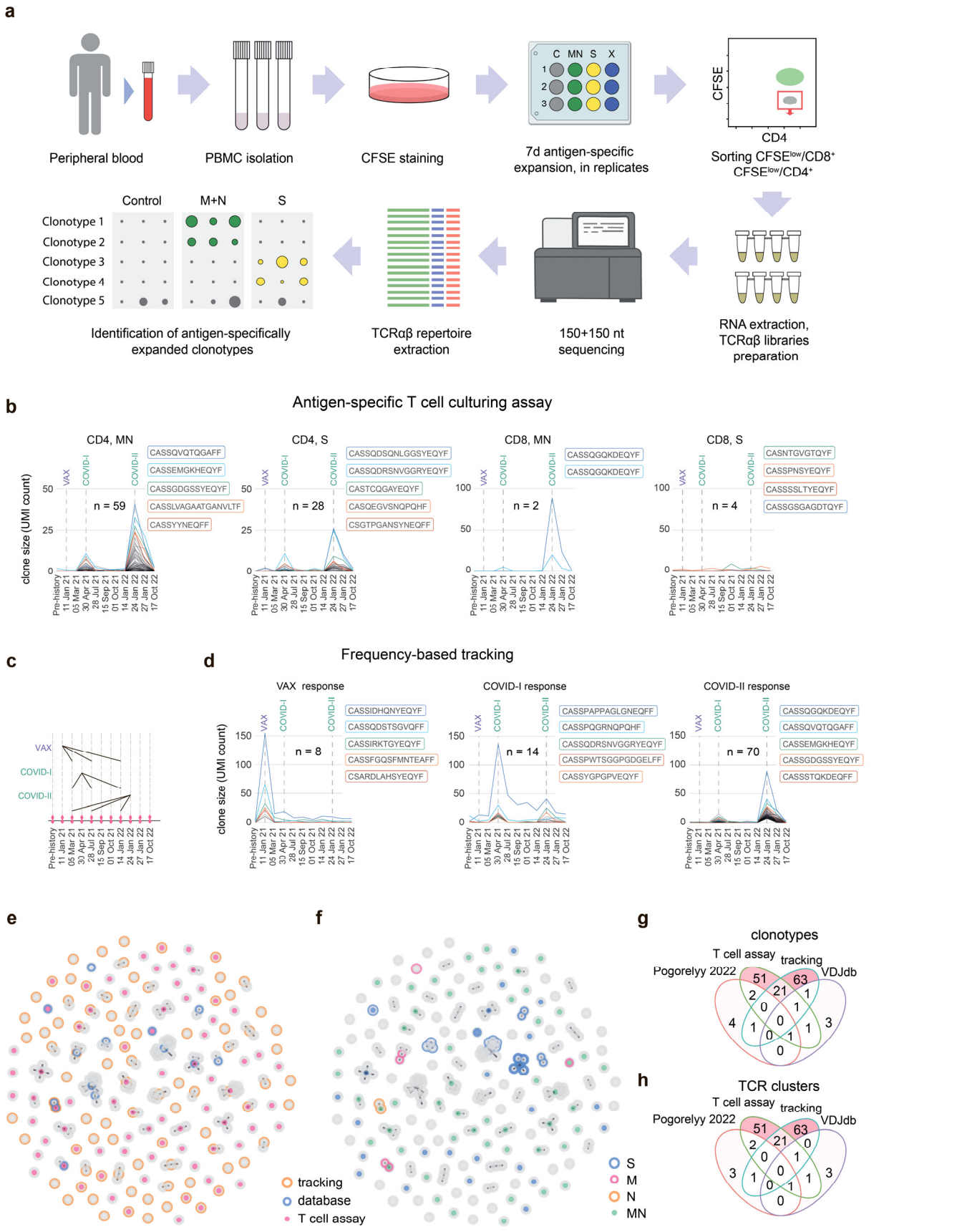
The frequencies of these clonotypes were tracked within the deep peripheral TCRβ repertoires of the donor, which were obtained at multiple time points. The analysis revealed that these clonotypes expanded in the donor's blood exclusively on two time points, which coincided with one or both SARS-CoV-2 infections (**Fig**. **3b**), confirming the accuracy of the CultivAToRR approach. A group of CD4+ Spike-specific TCRβ CDR3s (CASQEGVSNQPQHF, CASSEGASNQPQHF, CASSEGSSSQPQHF, TRBV6-1/TRBJ1-5) was identical or highly homologous to recently described TCRs specific to Spike epitope NLLLQYGSFCTQLNRAL in DRB1*15:01 context[43], also carried by D11. One of the CD4+ Spike-specific clonotypes that first appeared after vaccination also responded to the 1st SARS-CoV-2 infection and then dominated in the 2nd SARS-CoV-2 infection, suggesting vaccine-mediated protection (CASSQDSQNLGGSYEQYF, TRBV4-2/TRBJ2-7, **Fig**. **3b**).

Notably, the frequencies of each of these clonotypes were relatively low, mostly below 0.1% of the whole peripheral repertoire at the peak of response to the 2nd infection. This frequency is generally below the sensitivity of MHC tetramer staining or antigen-specific activation assays[44]. This result demonstrates the potency of our method, which integrates antigen-specific T cell culturing, replicates, and comparative data post-analysis, in identifying low-frequency T cell clones specific to the antigens of interest.

B. Clonal expansion in corresponding time points.

Although CultivAToRR approach allowed us to identify more than 90 SARS-CoV-2-specific TCRβ variants for D11, its results were limited to the three tested antigens, and biased towards CD4+ TCRs, thus essentially missing the CD8+ TCRs as well as response to non-M/N/S epitopes. To widen our fishing net, we applied a frequency-based approach to identify the TCRβ variants that specifically expanded at the time points of vaccination or SARS-CoV-2 infection in the bulk peripheral TCRβ repertoire of D11, a technique applied previously to reveal vaccine-responsive[39] or pathogen-specific[40, 45] T cell clones.

Again, edgeR analysis was applied to the bulk peripheral TCRβ repertoires, obtained in biological replicates and normalized to the 45,000 randomly sampled UMI (unique TCRβ cDNA molecules) per replicate per time point. This analysis revealed 8, 14 and 70 TCRβ clonotypes expanding at the time points of vaccination, 1st or 2nd SARS-CoV-2 infections, respectively, compared to the three reference time points (**Fig**. **3c**,**d**). A number of clonotypes that were identified as expanded at the time point of the 1st SARS-CoV-2 infection were also expanding at the 2nd infection, and *vice versa*. At the same time, similar analysis revealed only from 1 to 2 clonal expansions in the control time points compared to the challenge time points (**Supplementary Fig**. **7a**). This suggests that most of the edgeR-identified clonotypes that were expanded in peripheral blood in the challenge time points are truly associated with the corresponding challenges.


a
Peripheral blood
PBMC isolation
CFSE staining
C MN S X
7d antigen-specific expansion, in replicates
CFSE
CD4
Sorting CFSE^low/CD8^+ CFSE^low/CD4^+
RNA extraction, TCRαβ libraries preparation
150+150 nt sequencing
TCRαβ repertoire extraction
Control M+N S
Clonotype 1
Clonotype 2
Clonotype 3
Clonotype 4
Clonotype 5
Identification of antigen-specifically expanded clonotypes
b
Antigen-specific T cell culturing assay
CD4, MN
CD4, S
CD8, MN
CD8, S
clone size (UMI count)
VAX
COVID-I
COVID-II
n = 59
n = 28
n = 2
n = 4
CASSQVQTQGAFF
CASSEMGKHEQYF
CASSGDGSSYEQYF
CASSLVAGAATGANVLTF
CASSYYNEQFF
CASSQDSQNLGGSYEQYF
CASSQDRSNVGGRYEQYF
CASTCQGAYEQYF
CASQEGVSNQPQHF
CSGTPGANSYNEQFF
CASSQGQKDEQYF
CASSQGQKDEQYF
CASNTGVGTQYF
CASSPNSYEQYF
CASSSSLTYEQYF
CASSGSGAGDTQYF
Pre-history
11 Jan 21
05 Mar 21
30 Apr 21
28 Jul 21
15 Sep 21
01 Oct 21
14 Jan 22
24 Jan 22
27 Jan 22
17 Oct 22
c
d
Frequency-based tracking
VAX response
COVID-I response
COVID-II response
n = 8
n = 14
n = 70
CASSIDHQNYEQYF
CASSQDSTSGVQFF
CASSIRKTGYEQYF
CASSFGQSFMNTEAFF
CSARDLAHSYEQYF
CASSPAPPAGLGNEQFF
CASSPQGRNQPQHF
CASSQDRSNVGGRYEQYF
CASSPWTSGGPGDGELFF
CASSYGPGPVEQYF
CASSQGQKDEQYF
CASSQVQTQGAFF
CASSEMGKHEQYF
CASSGDGSSYEQYF
CASSSTQKDEQFF
e
f
g
h
tracking
database
T cell assay
S
M
N
MN
clonotypes
TCR clusters
T cell assay
tracking
Pogorelyy 2022
VDJdb
51
63
21
2
1
0
4
3
1
0

Figure

3. Three methods of capturing COVID-specifc TCR clonotypes. a. Scheme of the antigen-specific T cell culturing approach, CultivAToRR. **b**,**d**. Longitudinal tracking of TCRβ clonotypes (CDR3nt + V + J) identified with CultivAToRR (b) and based on expansion of frequency in target time points (d), using PBMC TCRβ profiling in replicates in sequential time points. Each PBMC TCR-Seq library is downsampled to 45,000 TCRβ encoding cDNA molecules (UMI). TCRβ CDR3 amino acid sequences are shown for the five largest clonotypes on each plot. **c**. Schematic representation of target and reference time points for frequency-based discovery of expanding TCRβ clonotypes. **e**,**f**. TCRβ CDR3 homology clusters identified by the three methods. Each node represents a single CD3aa + V + J clonotype, nodes having one amino acid mismatch in CDR3β are connected with edges. Method (e) and target antigen, if known (f) are shown. **g**,**h**. Venn diagrams representing clonal (g) and CDR3 homology cluster (h) overlaps between the methods.

C. Screening against known SARS-CoV-2-specific TCRs.

To exploit the accumulated knowledge of SARS-CoV-2-specific TCRs with known target peptides and/or restricting human leukocyte antigens (HLAs), we also overlapped peripheral time-tracked TCRβ repertoires of our donor with the VDJdb database[46] and the dataset of CD4+ TCRs enriched in SARS-CoV-2 patients with computationally implied HLA allele specificities[47]. We narrowed our search to the HLA class I and II alleles matching the donor's HLA (**Supplementary Note 1**). This analysis revealed 7 and 8 clonotypes matching those listed in VDJdb and Pogorelyy *et al.*, respectively. The largest expansions of identified clonotypes were observed at time points corresponding to infection with SARS-CoV-2 (**Supplementary Fig**. **7b**,**c**).

Finally, we integrated the data on SARS-CoV-2-specific TCRβ clonotypes obtained using the three approaches: 1) antigen-specific culturing; 2) frequency-based capturing; 3) database search (**Supplementary Table 4**). To account for the presence of convergent CDR3 groups[48] we built single aa-mismatch TCRβ CDR3 homology clusters on the pooled 1)+2)+3) data, visualized on **Fig 3e**,**f**. Notably, a number of CDR3 homology clusters and clonotypes were independently supported by 2 or all 3 of the approaches, cross-confirming their specificity (**Fig 3e-h**).

Mapping functionality of SARS-CoV-2-responsive T cell clones

To functionally characterize SARS-CoV-2-responsive Th cell clones, we performed scRNA-Seq of effector/memory Th cells at the time points right after COVID-I and COVID-II for the same donor D11. We next mapped D11 scRNA-Seq data on the bigger landscape of scRNA-Seq data integrated from 122 donors and assigned each D11 T cell to the corresponding scRNA-Seq cluster (**Supplementary Fig**. **8a**). Next, we highlighted SARS-CoV-2-responsive TCRβ clonotypes identified in the previous section within the scTCR-Seq of D11 (**Fig**. **4**, **Supplementary Fig**. **8b**).

17 of identified SARS-CoV-2-responsive clonotypes of D11 have been successfully mapped on Th cells localized in several scRNA-Seq clusters. Most of the cells represented "Self-stopping" (*PD-1+, TIGIT+*), Th2, Th2a, and Th17 clusters (**Fig**. **4b, Supplementary Fig**. **8c**). Most cells recognized as responding to COVID-I based on frequency expansion were also assigned to "Self-stopping", Th2, Th2a, and Th17 clusters. Most cells responding to the COVID-II as well as identified by antigen-specific culturing assay were assigned to the "Self-stopping" cluster (**Fig**. **4a**,**b, Supplementary Fig**. **8c**).

Notably, none of the SARS-CoV-2-responsive clonotypes have been identified among Th1, Temra-Th1, or "Eff-Mem IFN response" clusters. We can suggest that the Th response to COVID-I (moderate illness)

was suboptimal in the context of viral infection. The presence of “Self-stopping” SARS-CoV-2-specific T cells and their higher clonal heterogeneity could reflect the generally more prominent T cell response to COVID-II (mild illness).

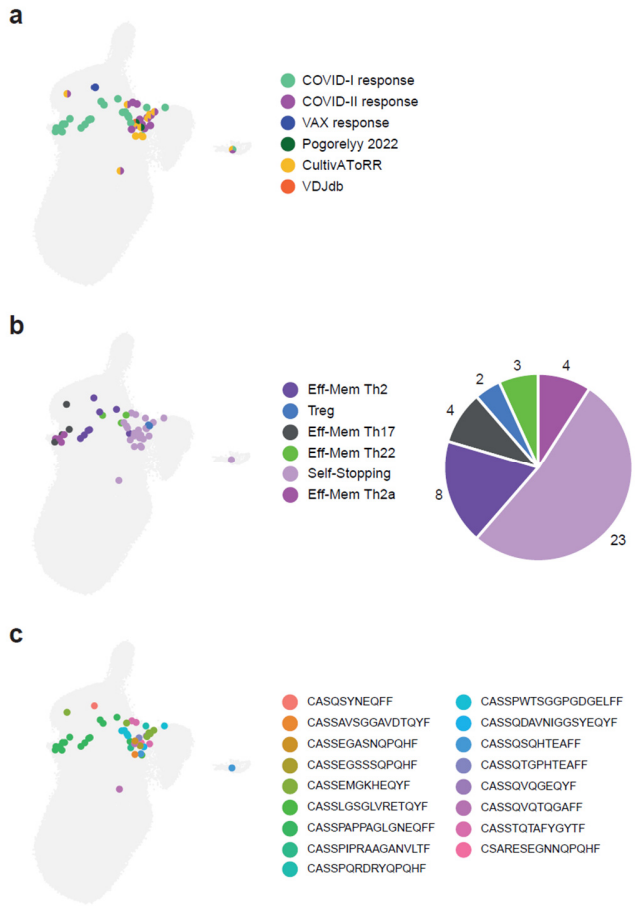


Figure 4. Mapping to the reference dataset shows functional clusters of SARS-CoV-2-specific clones. Positioning of D11 SARS-CoV-2-specific TCRβ CDR3 clonotypes. **a**. SARS-CoV-2-specific cells are colored based on identification method, clonotypes verified by more than one method are shown as small pie charts. **b**. Corresponding scRNA-Seq clusters. Large pie chart shows the proportions of scRNA-Seq clusters in the anti-SARS-CoV-2 response. **c**. TCRβ CDR3 clonotypes.

Discussion

The architecture of T cell memory essentially determines the entire pattern of our interaction with the antigens of the surrounding world, our microbiota, and our self-antigens[1, 2]. This architecture starts formation in the prenatal period[49, 50, 51, 52, 53], is actively formed in the first years of life in the contact with pathogens, airborne and food antigens, and microbiota maturation[54, 55] and then continues to be actively shaped by vaccinations and further contacts with infectious and non-infections challenges and antigens.

Clonal populations of T cells that are instructively primed by professional antigen-presenting cells[1, 2, 3] make decisions about which reaction programs to choose to respond to each specific antigen. They remember these programs as memory clones, forming stable response patterns to familiar challenges, patterns of regulation and cross-regulation of immune responses to friends and foes.

Mistakes made in such decisions cost us dearly: they lead to autoimmune diseases, inefficient elimination of pathogens, chronic inflammation, cancer, and may essentially underlie the entire phenomenon of inflammaging. Presumably for this reason, some mammalian species are likely to avoid forming such a long-term clonal memory[11].

In humans, however, the clonal memory of both CD8+ and CD4+ T lymphocytes can persist for years and decades[56, 57]. In the present work, this allowed us to map Th subset repertoires on the very same T cell clones of the scRNA-Seq data obtained for the donor 4 years later, which emphasizes the stability of the huge number of clones accumulated over previous years[58]. We also show that these clones are mapped predominantly or even exclusively within their corresponding and independent scRNA-Seq clusters, highlighting the persistence of program decisions once made by each T cell clone.

Repertoire-based Sort-Seq annotation of scRNA-Seq data turned out to be a powerful aid in matching classical subsets of lymphocytes sorted by surface markers with stable scRNAseq clusters that more comprehensively describe the diversity of functional lymphocyte programs. This effort structures our understanding of functional diversity of Th cells, critical for further progress in cancer and autoimmunity immunotherapy and vaccine development.

CITE-seq can also be used to phenotype cell populations on a single-cell level based on surface markers, having an advantage of simultaneously measuring hundreds of those. However, our work shows that Sort-Seq significantly outperforms CITE-Seq in terms of resolution capacity. Additionally, some of the CITE-Seq staining antibodies may affect cell signaling and transcriptomic profiles, introducing bias in the scRNA-Seq landscape. In contrast, in the Sort-Seq pipeline, cells stained with surface antibodies for sorting and bulk TCR-Seq are obtained and analyzed independently of the scRNA-Seq/scTCR-Seq experiment.

At the same time, Sort-Seq approach has certain limitations that must be taken into account. Firstly, similar to CITE-Seq, the method relies solely on surface markers to define target cell subpopulations. Secondly, the method is applicable only to T- and B-cells with specific clonal receptors acting as living barcodes. Thirdly, Sort-Seq is based on lymphocyte clonality, where we need to capture the same clone in sorting and scRNA-seq experiments. Therefore, Sort-Seq based annotation of e.g. central memory T cell subsets or peripheral blood Tfh cells, which are almost as diverse as naive T cells[25, 31], may require substantially deeper scRNA-Seq and bulk TCR profiling. For the naive T cell subsets, Sort-Seq application would be probably limited to innate-like, relatively clonal naive T cell subpopulations of fetal origin[51]. However, once the laborious work is done for the first time for each tissue and subsets of interest, scientific community may use the newly generated datasets to build well-annotated by Sort-Seq scRNA-Seq references of lymphocytes - as we did in the current study for peripheral blood CD4+ helper T cell subsets.

In the future works, Sort-Seq can be used to match and classify: 1) CD4+ T cell populations in the peripheral blood, lymph nodes, tertiary lymphoid structures at the sites of chronic inflammation and tumor environment, other tissues in health and disease; 2) diverse follicular helper T cells[31]; 3) known and unknown types of invariant and semi-invariant T cells, such as iNKT, NKT, MAIT, and CAIT[59, 60, 61]; 4) gamma delta T cells[62]; 5) central memory and stem cell memory T cells that would probably require deeper scRNA-Seq coverage due to their relatively lower clonality; 6) CD8+ T cell subsets[63]; 7) B cell functional subsets[64].

In the second part of the work, we report CultivAToRR, a cost-efficient method for identifying antigen-specific T cell clones in the patient's blood. This method, representing the modification of previously

described approaches[37, 38], was here empowered by biological replicates and appropriate statistical analysis of TCR repertoires. It should be widely applicable in studies of cancer, autoimmunity, infections, and vaccinations as a convenient easy-to-use tool for the rapid identification of effector/memory T cell clones specific to particular antigenic peptides or peptide mixes. Here, we employed CultivAToRR to identify more than 80 CD4+ SARS-CoV2-specific T cell clones in a single donor. Their expansion in longitudinal TCR repertoire-based tracking ideally coincided with two SARS-CoV2 infections of the donor, confirming accuracy of the method.

Finally, linking the two branches of the current work, we mapped SARS-CoV2-specific T cell clones with the scRNA-Seq data of the donor, integrated and annotated within the global CD4+ T cell reference map. This allowed us to determine their functional scRNA-Seq clusters, and to accordingly make certain assumptions about the nature of patients' immune response to the first and second infections.

Altogether, we hope that our work:

1) Clarifies correspondence between the well-studied Th subsets and scRNA-Seq landscape.
2) Proposes more accurate classification of CD4+ T cell scRNA-Seq, delineating positioning of Th2 and Th2a clusters, and refining a number of other details.
3) Shows long-term program stability and low intrinsic plasticity of Th memory clones.
4) Offers Sort-Seq, a new powerful approach for classification of T- and B-lymphocyte subsets and their exact positioning within scRNA-Seq landscape.
5) Provides an integrated easy-to-use scRNA-Seq reference dataset of peripheral Th lymphocytes.
6) Offers CultivAToRR, a powerful, cost-efficient, and easy-to-use method to identify antigen-specific TCRs in a 10-days assay that starts directly from PBMCs and does not require many laborious operations.

More generally, we hope that this work is helping to move the study of the role of programmed populations of memory T cells to a new level, making it possible to clearly distinguish the functional nature of each immune response, as well as to investigate the true plasticity between the known T cell subsets. This level of understanding serves as a necessary stepping stone to rational development of better immunotherapeutic approaches in oncology and autoimmunity, and vaccine development, where the chosen T cell programs fundamentally determine the type of immune response and decide the matter.

Data availability

Raw sequencing data is available in NCBI Sequence Read Archive (BioProject: PRJNA995237). Processed Seurat objects with Th reference and scRNA-Seq datasets from D11 are available for download at https://figshare.com/projects/T_helper_subsets_Kriukova_et_al_/173466. PBMC ds45k dataset (D11, TCRβ repertoires from PBMCs in replicates, downsampled to 45,000 UMI in each sample) is deposited in our GitHub repository (Th_kriukova/outs/ds45k) at https://github.com/kriukovav/Th_kriukova.

Code availability

Code for analysis (TCRβ repertoires from PBMCs) and all figures (scRNA-Seq and TCRβ repertoires) is available at https://github.com/kriukovav/Th_kriukova. R package wrapping Seurat[65] functions to perform reference mapping to our Th scRNA-Seq dataset is available at https://github.com/kriukovav/CD4map.

Acknowledgements

We thank Vadim Karnaukhov, Denis Syrko, Viktor Kotlyar, and Dmitry Bolotin for assistance with data analysis and Maria Vakhitova, Tatiana Grigorova, Ilgar Mamedov, and the NGS Laboratory (IKMB, Kiel University) for assistance with sequencing, which was also supported by the DFG Research Infrastructure NGS_CC (project 407495230) as part of the Next Generation Sequencing Competence Network (project 423957469).

ONLINE METHODS

IFN-γ ELISpot

PBMCs were isolated via standard density gradient centrifugation, quantified using a LUNA-II Automated Cell Counter (Logos Biosystems), and resuspended at 3 × 10^6 cells/mL in warm (37°C) CTL-Test™ Medium (Cellular Technology Ltd.) containing 1% L-glutamine (Thermo Fisher Scientific). Cells were then plated in 96-well anti-IFN-γ-coated ELISpot plates (Human IFN-γ Single-Color ELISPOT Kit, Cellular Technology Ltd.) at 3 × 10^5 cells/well in the presence of recombinant SARS-CoV-2 nucleocapsid phosphoprotein (N) or recombinant SARS-CoV-2 spike glycoprotein (S) in duplicate (each 1 µg/mL, Miltenyi Biotec). Negative control wells lacked viral proteins, and positive control wells contained phytohemagglutinin (2.5 ng/mL, Sigma-Aldrich). Plates were incubated for 18 h at 37°C in the presence of 5% CO_2. Assays were developed according to the manufacturer's instructions, and spots were counted using an ImmunoSpot S6 Universal Analyzer (Cellular Technology Ltd.).

CultivAToRR assay

Freshly isolated PBMCs were stained with CFSE (450 nM) and seeded in 48-well cell culture plates at 0.5 × 10^6 cells/well in complete RPMI (RPMI 1640 + 10% heat-inactivated fetal bovine serum) supplemented with IL-21 (10 ng/mL). The experiment included three conditions, each with three replicates: stimulation with PepTivator SARS-CoV-2 Prot_S (Miltenyi Biotec), stimulation with a mixture of PepTivator SARS-CoV-2 Prot_M and PepTivator SARS-CoV-2 Prot_N (Miltenyi Biotec), and control stimulation with a pool of irrelevant peptides (all at a final concentration of 0.6 nmol/peptide/mL). Cells were stained with anti-CD4–Alexa Fluor 647 (clone RPA-T4, BioLegend) and anti-CD8–BV421 (clone SK1, BioLegend) on day 6. $CFSE^{low}CD4^+$ and $CFSE^{low}CD8^+$ T cells were sorted directly into RLT buffer (200 µL, QIAGEN) using a FACSAria III (BD Biosciences). RNA was isolated using RNeasy Micro Kit (>500 cells/sample, QIAGEN) or with TRIzol Reagent (<500 cells/sample, Thermo Fisher Scientific). TCRβ libraries were constructed using all extracted RNA. Sequencing and raw data analysis were performed as described below. Significantly expanded clonotypes were identified using EdgeR[42, 66, 67]. UMI counts were filtered using the `filterByExpr` function with the following parameters: `min.count = 1, min.total.count = 5, large.n = 1, and min.prop = 0.5`. Count normalization and estimated dispersion by the default method were performed using the trimmed mean of M values (TMM). Each group was tested against all others using a quasi-likelihood (QL) negative binomial generalized log-linear model followed by the F-test.

Final clonotypes had an FDR <0.05 and a logFC ≥4. The code is available as part of the TCRgrapher library at https://github.com/KseniaMIPT/tcrgrapher (edgeR_pipeline function).

Bulk TCR-Seq library preparation and sequencing

PBMC replicates were lysed either fresh or after a freeze/thaw cycle in Buffer RLT (QIAGEN) (**Supplementary Table 3**). Total RNA was isolated using RNeasy Mini Kit (QIAGEN). UMI-based TCRβ repertoire libraries were prepared using a Human TCR RNA Multiplex Kit (MiLaboratories). Pooled samples were sequenced across 150+150 bp with a coverage of 50–100 reads per input cell on a NextSeq550 (Illumina).

Bulk TCR-Seq data analysis and clonotype table preprocessing

Raw fastq data were analyzed using MiXCR[68] (v4.1.0) with the appropriate built-in preset (either `milab-human-tcr-rna-multiplex-cdr3` or `milab-human-tcr-rna-race-cdr3`, the latter used only for pre-vaccination bulk PBMC TCR repertoire extraction, time point "pre-history"). Only cDNA molecules sequenced at least twice (based on UMI coverage) were considered in the generation of TCR clonotype tables (parameters `Massemble.consensusAssemblerParameters.assembler.minRecordsPerConsensus=2` and `MrefineTagsAndSort.parameters.postFilter=null` in MiXCR). TCR clonotype tables were additionally filtered to include only productive CDR3β amino acid sequences starting with the conserved cysteine and ending with the conserved phenylalanine. The resulting TCRβ repertoires from bulk PBMCs were all normalized to the same depth by randomly sampling 45,000 unique cDNA molecules from each sample (ds45k dataset). To complement the missing time point before the first vaccination, an artificial "pre-history" time point was generated with a pre-vaccination TCRβ repertoire composed of exactly 45,000 cDNA molecules, based on Rep-Seq data from sorted $CD4^+$ and $CD8^+$ T cells from the same donor (several time points in replicates, all collected in 2017). These samples were divided into two sets, each incorporating $CD4^+$ and $CD8^+$ TCR repertoires, and 30,000 $CD4^+$ TCR cDNA molecules and 15,000 $CD8^+$ TCR cDNA molecules were randomly sampled from each set. This procedure resulted in two replicates of TCRβ repertoires from $CD4^+$ and $CD8^+$ T cells pooled at a ratio of 2:1 (pseudo-bulk PBMCs). All table data manipulations were performed using R (v4.1.2) and tidyverse (v1.3.2).

Clonal expansions at corresponding time points

TCR clonotypes were defined as unique nucleotide TCRβ sequences in the ds45k dataset. Differentially abundant TCRs between a considered time point and three relevant reference time points, where no immune challenge was documented, were identified using edgeR (v3.36.0). Biological replicates were used to fuel edgeR statistics. The exact combinations of experimental time points together with the relevant reference time points are shown on **Fig. 3c**. Clonotypes were defined as expanded at an FDR <0.05 and a $\log2(count_{exp}/count_{ref})$ >1. Artificially generated pre-vaccination TCR repertoires (pre-history) were not used in the differential abundance analysis and were only adopted for the purpose of visualization.

Screening against known SARS-CoV-2-specific TCRs

Overlaps were sought between clonotypes in ds45k dataset and the VDJdb-2020-03-30 database[46] or the Pogorelyy *et al.* dataset[47], with exact TRBV and TRBJ segment matches and a maximum of one amino acid mismatch within the CDR3β (cdr3aa-v-j-1mm intersect type). Analysis and data visualization were confined to clonotypes specific for SARS-CoV-2 restricted by a donor-matched HLA. Clonotypes represented by at least three unique cDNA molecules at the VAX, COVID1, or COVID2 time points in the ds45k dataset were also selected for reliability.

TCRβ CDR3 homology clusters identified by the three methods

Clonotypes (defined by amino acid CDR3β sequence and TRBV-TRBJ segments) identified via CultivAToRR, clonal expansion at corresponding time points, and mapping against known SARS-CoV-2-specific TCRs were pulled together before clustering, linked by TRBV-TRBJ segment and CDR3β sequence identity. This set was complemented by including clonotypes from the ds45k dataset with one amino acid mismatch. Only clonotypes with the same TRBV-TRBJ combination were allowed in any one cluster. For cluster visualization, each node represents one TCRβ clonotype, and each edge links nodes with one amino acid mismatch.

scRNA-Seq and scTCR-Seq library preparation and sequencing

Frozen PBMCs from D11 (COVID1 and COVID2 samples) were thawed according to 10x Genomics recommendations and rested for 1 h in complete RPMI (RPMI 1640 + 10% autologous serum) at 37°C in the presence of 5% CO_2. Cells were then stained with anti-CD4–Alexa Fluor 647 (clone RPA-T4, BioLegend), anti-CD8–PerCP (clone SK1, BioLegend), anti-CD19–FITC (clone J3-119, Beckman Coulter), anti-CD27–eFluor 780 (clone LG.7F9, Thermo Fisher Scientific), anti-CD45RA–eFluor 450 (clone HI100, Thermo Fisher Scientific), and 7-AAD (Thermo Fisher Scientific). Effector/memory $CD4^+$ T cells gated as $CD4^+$ $CD8^-$ $CD19^-$ after excluding $CD27^+$ $CD45RA^+$ events were sorted into complete RPMI using a FACSAria III (BD Biosciences). Dead cells were excluded based on morphology and staining with 7-AAD. Sorted effector/memory $CD4^+$ T cells were washed twice in phosphate-buffered saline containing 0.04% bovine serum albumin and loaded onto a Next GEM Chip (10x Genomics) in one replicate with a target count of 10,000 cells for the COVID1 sample and in two replicates with a target count of 10,000 cells for the COVID2 sample. The emulsion from the COVID1 sample was divided into two parts, resulting in the generation of two technical replicates, each originating from approximately 5,000 cells. Samples were prepared using a Chromium Next GEM Single Cell V(D)J Reagent Kit v1.1 (10x Genomics). Pooled samples were sequenced with a coverage of 16,000 reads per input cell for scRNA-Seq and 5,000 reads per input cell for scTCR-Seq on a NextSeq 550 System (Illumina).

Fresh PBMCs from D01, D04, D05 were stained with anti-CCR7–PE-Cy7 (clone 3D12, BD Biosciences), anti-CD3–APC-Fire750 (clone SK7, BioLegend), anti-CD4–PE-Cy5.5 (clone S3.5, Thermo Fisher Scientific), anti-CD14–V500 (clone M5E2, BD Biosciences), anti-CD19–V500 (clone HIB19, BD Biosciences), anti-CD45RA–PE-Cy5 (clone HI100, BioLegend), and LIVE/DEAD Fixable Aqua (Thermo Fisher Scientific). Viable effector/memory $CD4^+$ T cells gated as $CD3^+$ $CD4^+$ $CD14^-$ $CD19^-$ after excluding $CCR7^+$ $CD45RA^+$ events were sorted in two replicates for D01 and without replicates for D04 and D05 using a custom-modified FACSAria II (BD Biosciences) and loaded onto a Chromium Controller (10x

Genomics). Samples were prepared using a Chromium Next GEM Single Cell 5' Reagent Kit v2 (10x Genomics). Pooled samples were sequenced with a coverage of 100,000 reads per input cell for scRNA-Seq and 25,000 reads per input cell for scTCR-Seq on a NovaSeq 6000 System with an S4 Flow Cell (Illumina).

scRNA-Seq and scTCR-Seq data analysis

Raw scRNA-Seq and scTCR-seq fastq files were processed using the `count` and `vdj` pipelines in cellranger (v6.1.2). Filtered gene expression matrices were uploaded into the Seurat R package (v4.2.0)[65]. TCR data were processed using the CellaRepertorium R package (v1.4.0). The most abundant TCR chain was used for cells with two relevant transcripts (TRA or TRB). Publicly available data were downloaded via E-MTAB-10026 and converted to Seurat object. $CD4^+$ T cell clusters were chosen based on the average cluster expression of *CD4* and *CD3E*. The public dataset was then split by sample origin. CITE-Seq ADT counts were normalized using the CLR method. Additional filtering steps were performed on individual datasets, including the elimination of outlier clusters with low UMI counts and/or high percentages of mitochondrial genes and cells expressing *CD8A* and *CD8B*. The stress score was calculated for each cell using `AddModuleScore()` function in Seurat and included genes upregulated as a consequence of the dissociation procedure[69]. Integration was carried out using the Seurat reference-based reciprocal PCA protocol, with the largest 3' and 5' datasets chosen as references to account for differences in methodology[65]. The percentage of mitochondrial genes and the stress gene signature were regressed out of individual and integrated datasets. All TCR and IG genes were removed from variable features used in the PCA and from the anchor features used for integration. The number of dimensions used for running the UMAP algorithm was 25.

Reference mapping

D11_COVID1, D11_COVID2_rep1, and D11_COVID2_rep2 samples were mapped to the integrated reference dataset Seurat Reference Mapping procedure with default parameters[65]. Plots were designed to highlight amino acid-defined TCRβ clonotypes with specificity for SARS-CoV-2.

Ethics statement

Ethical approval was granted by the Cardiff University School of Medicine Research Ethics Committee (16/55) and the Institutional Review Board of the Pirogov Russian National Research Medical University.

References

1. Kunzli, M. & Masopust, D. CD4(+) T cell memory. *Nat Immunol* **24**, 903-914 (2023).
2. Sallusto, F. Heterogeneity of Human CD4(+) T Cells Against Microbes. *Annual review of immunology* **34**, 317-334 (2016).
3. Borst, J., Ahrends, T., Babala, N., Melief, C.J.M. & Kastenmuller, W. CD4(+) T cell help in cancer immunology and immunotherapy. *Nat Rev Immunol* **18**, 635-647 (2018).

4. McDonald, D.R. TH17 deficiency in human disease. *The Journal of allergy and clinical immunology* **129**, 1429-1435; quiz 1436-1427 (2012).

5. Cook, M.C. & Tangye, S.G. Primary immune deficiencies affecting lymphocyte differentiation: lessons from the spectrum of resulting infections. *Int Immunol* **21**, 1003-1011 (2009).

6. Hernandez-Santos, N. *et al.* Th17 cells confer long-term adaptive immunity to oral mucosal Candida albicans infections. *Mucosal immunology* **6**, 900-910 (2013).

7. Misiak, A. *et al.* Addition of a TLR7 agonist to an acellular pertussis vaccine enhances Th1 and Th17 responses and protective immunity in a mouse model. *Vaccine* **35**, 5256-5263 (2017).

8. Bacher, P. & Scheffold, A. Antigen-specific regulatory T-cell responses against aeroantigens and their role in allergy. *Mucosal immunology* **11**, 1537-1550 (2018).

9. McGee, H.S. & Agrawal, D.K. TH2 cells in the pathogenesis of airway remodeling: regulatory T cells a plausible panacea for asthma. *Immunologic research* **35**, 219-232 (2006).

10. Finotto, S. T-cell regulation in asthmatic diseases. *Chem Immunol Allergy* **94**, 83-92 (2008).

11. Izraelson, M. *et al.* Distinct organization of adaptive immunity in the long-lived rodent Spalax galili. *Nat Aging* **1**, 179-189 (2021).

12. Osnes, L.T., Nakken, B., Bodolay, E. & Szodoray, P. Assessment of intracellular cytokines and regulatory cells in patients with autoimmune diseases and primary immunodeficiencies - novel tool for diagnostics and patient follow-up. *Autoimmunity reviews* **12**, 967-971 (2013).

13. Costa, N. *et al.* Two separate effects contribute to regulatory T cell defect in systemic lupus erythematosus patients and their unaffected relatives. *Clin Exp Immunol* **189**, 318-330 (2017).

14. Bonelli, M. *et al.* Quantitative and qualitative deficiencies of regulatory T cells in patients with systemic lupus erythematosus (SLE). *Int Immunol* **20**, 861-868 (2008).

15. Miyara, M. *et al.* Global natural regulatory T cell depletion in active systemic lupus erythematosus. *J Immunol* **175**, 8392-8400 (2005).

16. Protti, M.P., De Monte, L. & Di Lullo, G. Tumor antigen-specific CD4+ T cells in cancer immunity: from antigen identification to tumor prognosis and development of therapeutic strategies. *Tissue Antigens* **83**, 237-246 (2014).

17. Andreatta, M. *et al.* A CD4(+) T cell reference map delineates subtype-specific adaptation during acute and chronic viral infections. *eLife* **11** (2022).

18. Yasumizu, Y. *et al.* Single-cell transcriptome landscape of circulating CD4⁺ T cell populations in human autoimmune diseases. *bioRxiv*, 2023.2005.2009.540089 (2023).

19. Radtke, D. *et al.* Th2 single-cell heterogeneity and clonal distribution at distant sites in helminth-infected mice. *eLife* **11** (2022).

20. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387 (2014).

21. Reimegard, J. *et al.* A combined approach for single-cell mRNA and intracellular protein expression analysis. *Commun Biol* **4**, 624 (2021).

22. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865-868 (2017).

23. Peterson, V.M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol* **35**, 936-939 (2017).

24. Shahi, P., Kim, S.C., Haliburton, J.R., Gartner, Z.J. & Abate, A.R. Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Scientific reports* **7**, 44447 (2017).

25. Kasatskaya, S.A. *et al.* Functionally specialized human CD4(+) T-cell subsets express physicochemically distinct TCRs. *eLife* **9** (2020).

26. Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nat Med* **27**, 904-916 (2021).

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

28. Cheroutre, H. & Husain, M.M. CD4 CTL: living up to the challenge. *Semin Immunol* **25**, 273-281 (2013).

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

30. Scharf, L. *et al.* Longitudinal single-cell analysis of SARS-CoV-2-reactive B cells uncovers persistence of early-formed, antigen-specific clones. *JCI insight* **8** (2023).

31. Brenna, E. *et al.* CD4(+) T Follicular Helper Cells in Human Tonsils and Blood Are Clonally Convergent but Divergent from Non-Tfh CD4(+) Cells. *Cell reports* **30**, 137-152 e135 (2020).

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

33. Barnes, J.L. *et al.* T-helper 22 cells develop as a distinct lineage from Th17 cells during bacterial infection and phenotypic stability is regulated by T-bet. *Mucosal immunology* **14**, 1077-1087 (2021).

34. Perez, L.G. *et al.* TGF-beta signaling in Th17 cells promotes IL-22 production and colitis-associated colon cancer. *Nature communications* **11**, 2608 (2020).

35. Knudson, C.J. *et al.* Mechanisms of Antiviral Cytotoxic CD4 T Cell Differentiation. *J Virol* **95**, e0056621 (2021).

36. Hoeks, C., Duran, G., Hellings, N. & Broux, B. When Helpers Go Above and Beyond: Development and Characterization of Cytotoxic CD4(+) T Cells. *Frontiers in immunology* **13**, 951900 (2022).

37. Klinger, M. *et al.* Combining next-generation sequencing and immune assays: a novel method for identification of antigen-specific T cells. *PLoS One* **8**, e74231 (2013).

38. Danilova, L. *et al.* The Mutation-Associated Neoantigen Functional Expansion of Specific T Cells (MANAFEST) Assay: A Sensitive Platform for Monitoring Antitumor Immunity. *Cancer immunology research* **6**, 888-899 (2018).

39. Pogorelyy, M.V. *et al.* Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proc Natl Acad Sci U S A* **115**, 12704-12709 (2018).

40. Minervina, A.A. *et al.* Primary and secondary anti-viral response captured by the dynamics and phenotype of individual T cell clones. *eLife* **9** (2020).

41. Yuzhakova, D.V. *et al.* Measuring Intratumoral Heterogeneity of Immune Repertoires. *Frontiers in oncology* **10**, 512 (2020).

42. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

43. Wragg, K.M. *et al.* Establishment and recall of SARS-CoV-2 spike epitope-specific CD4(+) T cell memory. *Nat Immunol* **23**, 768-780 (2022).

44. Dolton, G. *et al.* Optimized Peptide-MHC Multimer Protocols for Detection and Isolation of Autoimmune T-Cells. *Frontiers in immunology* **9**, 1378 (2018).

45. Minervina, A.A. *et al.* Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. *eLife* **10** (2021).

46. Goncharov, M. *et al.* VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. *Nat Methods* **19**, 1017-1019 (2022).

47. Pogorelyy, M.V. *et al.* Resolving SARS-CoV-2 CD4(+) T cell specificity via reverse epitope discovery. *Cell Rep Med* **3**, 100697 (2022).

48. Venturi, V. *et al.* Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc Natl Acad Sci U S A* **103**, 18691-18696 (2006).

49. King, C.L. *et al.* Acquired immune responses to Plasmodium falciparum merozoite surface protein-1 in the human fetus. *J Immunol* **168**, 356-364 (2002).

50. Li, N. *et al.* Memory CD4(+) T cells are generated in the human fetal intestine. *Nat Immunol* **20**, 301-312 (2019).

51. Pogorelyy, M.V. *et al.* Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS computational biology* **13**, e1005572 (2017).

52. Prescott, S.L. *et al.* Development of allergen-specific T-cell memory in atopic and normal children. *Lancet* **353**, 196-200 (1999).

53. Huygens, A., Dauby, N., Vermijlen, D. & Marchant, A. Immunity to cytomegalovirus in early life. *Frontiers in immunology* **5**, 552 (2014).

54. Renz, H. & Skevaki, C. Early life microbial exposures and allergy risks: opportunities for prevention. *Nat Rev Immunol* **21**, 177-191 (2021).

55. Pieren, D.K.J., Boer, M.C. & de Wit, J. The adaptive immune system in early life: The shift makes it count. *Frontiers in immunology* **13**, 1031924 (2022).

56. Naniche, D. *et al.* Decrease in measles virus-specific CD4 T cell memory in vaccinated subjects. *J Infect Dis* **190**, 1387-1395 (2004).

57. Jokinen, S., Osterlund, P., Julkunen, I. & Davidkin, I. Cellular immunity to mumps virus in young adults 21 years after measles-mumps-rubella vaccination. *J Infect Dis* **196**, 861-867 (2007).

58. Yoshida, K. *et al.* Aging-related changes in human T-cell repertoire over 20years delineated by deep sequencing of peripheral T-cell receptors. *Experimental gerontology* **96**, 29-37 (2017).

59. LeBlanc, G., Kreissl, F.K., Melamed, J., Sobel, A.L. & Constantinides, M.G. The role of unconventional T cells in maintaining tissue homeostasis. *Semin Immunol* **61-64**, 101656 (2022).

60. Rosati, E. *et al.* A novel unconventional T cell population enriched in Crohn's disease. *Gut* (2022).

61. Crowther, M.D. *et al.* Genome-wide CRISPR-Cas9 screening reveals ubiquitous T cell cancer targeting via the monomorphic MHC class I-related protein MR1. *Nat Immunol* **21**, 178-185 (2020).

62. Davey, M.S. *et al.* The human Vdelta2(+) T-cell compartment comprises distinct innate-like Vgamma9(+) and adaptive Vgamma9(-) subsets. *Nature communications* **9**, 1760 (2018).

63. Galletti, G. *et al.* Two subsets of stem-like CD8(+) memory T cell progenitors with distinct fate commitments in humans. *Nat Immunol* **21**, 1552-1562 (2020).

64. Stewart, A. *et al.* Single-Cell Transcriptomic Analyses Define Distinct Peripheral B Cell Subsets and Discrete Development Pathways. *Frontiers in immunology* **12**, 602539 (2021).

65. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529 (2021).

66. McCarthy, D.J., Chen, Y. & Smyth, G.K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40**, 4288-4297 (2012).

67. Chen, Y., Lun, A.T. & Smyth, G.K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Res* **5**, 1438 (2016).

68. Bolotin, D.A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* **12**, 380-381 (2015).

69. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol* **21**, 130 (2020).