

Longitudinal analysis of T-cell receptor repertoires reveals persistence of antigen-driven CD4+ and CD8+ T-cell clusters in Systemic Sclerosis

Authors and affiliations

N.H. Servaas^{a,b}, F. Zaaraoui-Boutahar^c, C.G.K. Wichers^{a,b}, A. Ottria^{a,b}, E. Chouri^{a,b}, A.J. Affandi^{a,b}, S. Silva-Cardoso^{a,b}, M. van der Kroef^{a,b}, T. Carneiro^{a,b}, F. van Wijk^b, T.R.D.J. Radstake^{a,b}, A.C. Andeweg^c, A. Pandit^{a,b*}

^aDepartment of Rheumatology & Clinical Immunology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; ^bCenter for Translational Immunology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; ^cDepartment of Viroscience, Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands

***Corresponding Author**

Aridaman Pandit, PhD

Email: a.pandit@umcutrecht.nl

Telephone: +31 88 75 535 69

Full postal address:

University Medical Center Utrecht

KC.02.046.0, Lundlaan 6

3583EA Utrecht, The Netherlands

Abstract

The T-cell receptor (TCR) is a highly polymorphic surface receptor that allows T-cells to recognize antigenic peptides presented on the major histocompatibility complex (MHC). Changes in the TCR repertoire have been observed in several autoimmune conditions, and these changes are suggested to predispose autoimmunity. Multiple lines of evidence have implied an important role for T-cells in the pathogenesis of Systemic Sclerosis (SSc), a complex autoimmune disease. One of the major questions regarding the roles of T-cells is whether expansion and activation of T-cells observed in the diseases pathogenesis is (auto)antigen driven.

To investigate the temporal TCR repertoire dynamics in SSc, we performed high-throughput sequencing of CD4+ and CD8+ TCR β chains on longitudinal samples obtained from four SSc patients collected over a minimum of two years. Repertoire overlap analysis revealed that samples taken from the same individual over time shared a high number of TCR β sequences, indicating a clear temporal persistence of the TCR β repertoire in CD4+ as well as CD8+ T-cells. Moreover, the TCR β s that were found with a high frequency at one time point were also found with a high frequency at the other time points (even after almost four years), showing that frequencies of dominant TCR β s are largely consistent over time. We also show that TCR β generation probability and observed TCR frequency are not related in SSc samples, showing that clonal expansion and persistence of TCR β s is caused by antigenic selection rather than convergent recombination. Moreover, we demonstrate that TCR β diversity is lower in CD4+ and CD8+ T-cells from SSc patients compared to healthy memory T-cells, as SSc TCR β repertoires are largely dominated by clonally expanded persistent TCR β sequences. Lastly, using “Grouping of Lymphocyte Interactions by Paratope Hotspots” (GLIPH2), we identify clusters of TCR β sequences with homologous sequences that potentially recognize the same antigens and contain TCR β s that are persist in SSc patients.

In conclusion, our results show that that CD4+ and CD8+ T-cells are highly persistent in SSc patients over time, and this persistence is likely a result from antigenic selection. Moreover, persistent TCRs form high similarity clusters with other (non-)persistent sequences, that potentially recognize the same epitopes. These data provide evidence for an (auto-)antigen driven expansion of CD4/CD8+ T-cells in SSc.

Keywords

T-cell receptor; Repertoire sequencing; Systemic Sclerosis; Autoimmunity; Longitudinal Immunosequencing

1. Introduction

Systemic Sclerosis (SSc) is a complex chronic autoimmune disease, characterized by vascular abnormalities and widespread fibrosis affecting the skin and internal organs[1]. Although the pathogenic mechanisms underlying SSc remain largely unknown, multiple lines of evidence imply an important role for CD4+ and CD8+ T-cells in the progression of the disease. Activated T-cells infiltrate the skin of SSc patients already in the early phase of the disease[2,3]. These infiltrating T-cells can cross-talk with fibroblasts, inducing fibroblast activation and apoptosis through secretion of pro-inflammatory cytokines and fas/fas ligand engagement[4,5]. Moreover, T-cells isolated from SSc patients undergo clonal expansion when cultured together with autologous fibroblasts, suggesting that auto-antigens presented by SSc fibroblasts can induce auto-reactive T-cell responses[6]. Apart in the from skin, peripheral blood T-cells from SSc patients also exhibit signs of activation and express activation markers, including IL-2R, HLA-DR, and CD29[7–9], and secrete pro-inflammatory and pro-fibrotic factors[10–12].

The T-cell receptor (TCR) is a highly polymorphic surface receptor that allows T-cells to recognize antigenic peptides presented on the major histocompatibility complex (MHC)[13]. CD4+ T-cells recognize peptides presented on the MHC class II complex, while CD8+ T-cells

recognize peptides presented on the MHC class I complex. Classical TCRs are heterodimers consisting of a paired α - and β -chain. These chains making up the TCR are generated through somatic recombination of V (variable), D (diversity) and J (joining) gene segments accompanied by pseudorandom insertions and deletions of nucleic acids at their joining regions[14], thereby giving rise to an enormously diverse TCR repertoire in every individual. By this process of VDJ recombination, the small set of genes that encode the TCR can be used to create over 10^{15} potential TCR clonotypes[13,15]. Previous estimates of number of unique T-cells in a human range from 10^6 to 10^{11} [16–18], meaning that every individual only carries a small fraction of the potential repertoire.

High throughput sequencing of TCR repertoires is emerging as a valuable tool to unravel the exact role of T-cells in autoimmune diseases. The TCR repertoire has been proposed to serve as diagnostic biomarker for various autoimmune diseases, and recent studies have identified disease-associated TCR sequences in autoimmune diseases including autoimmune encephalomyelitis (AE), systemic lupus erythematosus (SLE), and rheumatoid arthritis (RA)[19–21]. Moreover, changes in T-cell repertoire diversity have been suggested to predispose the pathological manifestations in RA patients[22]. Prior studies examining the TCR repertoire in SSc have shown that there is an oligoclonal expansion of T-cells in the skin, lungs and blood of SSc patients[23–25], suggesting that expanded T-cells are involved in the disease pathogenesis. However, the results from current studies examining at the TCR repertoire in SSc patients are limited as: a) they lack the use of high-throughput techniques; b) consider either CD4+ or CD8+ or unsorted T-cell populations; and c) they study T-cells obtained only from a single time point thereby providing a static snapshot of the TCR repertoire in SSc.

Two major hypotheses have been postulated to explain mechanism of the expansion of T-cells in the context of autoimmunity[26][27]. The first hypothesis states that T-cells might expand non-specifically or by chance (bystander activation) due to chronic inflammation

observed in autoimmune patients[27,28][29]. In this case, proliferation of T-cells is induced through non-specific activation in the presence of TLR ligands and cytokines during an immune response. Due to inherent biases in the V(D)J recombination process, some TCR β sequences are more prevalent as they have a high generation probability[30]. As a result of this bias, during bystander activation, naïve T-cell clones with TCR β s that have high generation probabilities have a larger chance of being at the site of action due to their increased prevalence, and therefore have a higher chance to expand. In this case, expansion is a result of chance. The second hypothesis states that clonal expansion in autoimmunity is driven by a chronic, specific response to (auto-)antigens that selectively skew the TCR repertoire[26]. Here expansion is driven by antigen specific selection rather than chance. It remains to be unraveled which of these two mechanisms contributes to the activation and expansion of autoreactive T-cells in SSc.

To better understand T-cell responses in SSc pathogenesis, here we investigate the temporal TCR repertoire dynamics in SSc patients. We performed high-throughput sequencing of TCR β chains of sorted CD4+ and CD8+ non-naïve T-cells isolated from longitudinal samples from four SSc patients collected over a minimum of two years.

2. Materials and Methods

2.1 Sample collection

Whole heparinized blood samples from SSc patients were obtained from the University Medical Center Utrecht. SSc patients were classified according to the ACR/EULAR criteria[31]. This study was conducted in accordance with the Declaration of Helsinki and was performed with approval of the Institutional Review Board of the University Medical Centre Utrecht, The Netherlands. The medical ethics committee of the UMC Utrecht approved the study (METC nr 13-697). All participants enrolled in the study signed informed, and patient samples were anonymized upon collection. Peripheral blood mononuclear cells (PBMCs) were isolated by density gradient centrifugation on Ficoll-PlaqueTM Plus (GE Healthcare,

Uppsala, Sweden). pDCs, mDCs, B-cells and monocytes were first depleted by magnetic bead sorting using the autoMACs Pro Separator (Miltenyi Biotec, Bergisch Gladbach, Germany) according to the manufacturer's protocol. The remaining peripheral blood lymphocytes (PBLs) were resuspended in freezing medium (80% FCS, (Sigma-Aldrich, Saint Louis, Missouri, USA) / 20% DMSO, (Sigma-Aldrich)) and stored in sterile cryovials in liquid nitrogen (-196°C) until further use. From all patients PBLs were collected at baseline (T0), at least one year after inclusion (T1, ranging from 12-19 months) and at least two years after inclusion (T2, ranging from 24-46 months) (Figure 1a).

2.2 T-cell sorting

PBLs were thawed in RPMI 1640 (Gibco, Thermo Fisher Scientific, Waltham, Massachusetts, USA) supplemented with 20% FCS (Sigma-Aldrich), and washed with PBS. Subsequently, the cells were resuspended in FACS buffer (PBS supplemented with 1% BSA (Sigma-Aldrich) and 0.1% sodium azide (Sigma-Aldrich)) and stained using the following antibodies: CD3-AF700 (clone UCHT1, Biolegend, San Diego, California, USA), CD4-BV711 (clone OKT4, Sony Biotechnology, San Jose, California, USA), CD8-V500 (clone RPA-T8, BD Bioscience, San Jose, California, USA), CD56-PE-CF594 (clone B159, BD Bioscience), CD16-BV785 (clone 3G8, Sony Biotechnology), CD14-PerCP-Cy5.5 (clone HCD14, Biolegend), HLADR-BV421 (clone L243, Biolegend), CD45RO-PE-Cy7 (clone L243, BD Bioscience), CD27- APC-eFluor780 (clone O323, eBioscience, San Diego, California, USA). Multiparametric flow cytometry sorting of non-naive CD4⁺ T-cells (CD3⁺CD4⁺CD45RO⁺CD27^{+/-}) and non-naive CD8⁺ T-cells (CD3⁺CD8⁺CD45RO⁺CD27^{+/-}) was performed on the BD FACSAria II (BD Bioscience), according to the gating strategy described in Figure 1b. After sorting, cells were washed in PBS, lysed in TRIzol™ Reagent (Invitrogen, Carlsbad, California, USA), and stored at -20°C until RNA isolation.

2.3 RNA isolation and TCR sequencing

RNA was isolated using the RNeasy Mini Kit (Qiagen, Venlo, Netherlands) by adding ethanol to the upper aqueous phase of processed TRIzol samples and transferring directly to the RNeasy spin columns. TCR amplification was performed according to the protocol published by Mamedov et al[32]. Primer and barcode sequences are provided in Supplementary Table 1. Briefly, cDNA was generated by RACE using a primer directed to the TCR β constant region. Thirteen nucleotide long unique molecular identifiers (UMIs) were incorporated during cDNA synthesis. Subsequently, two-stage semi-nested PCR amplification was performed including a size selection/agarose gel purification step after the first PCR. To minimize cross-sample contamination, 5-nucleotide sample specific barcodes were introduced at two steps during the library preparation process[32]. Resulting TCR amplicons were subjected to high-throughput sequencing using the Ovation Low Complexity Sequencing System kit (NuGEN, San Carlos, California, USA) according to the manufacturer's instructions, and the Illumina MiSeq system (Illumina, San Diego, California, USA), using indexed paired-end 300 cycle runs.

2.4 TCR repertoire analysis

Raw paired-end reads were assembled using Paired-End reAd mergeR (PEAR)[33]. Sample specific barcode correction was performed using the 'umi_group_ec' command from the Recover T Cell Receptor (RTCR) pipeline[34], allowing zero mismatches in the barcode seed sequence for UMI detection (sample specific barcodes are provided in Supplementary Table 1). This strict barcode selection resulted in about 50% loss of reads, but ensured that there was minimal cross-sample contamination. Subsequently, barcode sequence reads having the same UMI were collapsed into consensus sequences using the RTCR pipeline to accurately recover TCR β sequences. Downstream data analysis of TCR β repertoires was performed using the tcR R package[35].

Healthy longitudinal sequencing data[dataset][36] was obtained from the immuneACCESS portal of Adaptive Biotechnologies repository at: <https://clients.adaptivebiotech.com/pub/healthy-adult-time-course-TCRB>.

2.5 Statistical analyses

Statistical analyses were performed using R version 3.4.1[37], and figures were produced using the R package ggplot2[38]. Generation probabilities of TCR β amino acid sequences were computed using the generative model of V(D)J recombination implemented by OLGA (Optimized Likelihood estimate of immunoGlobulin Amino-acid sequences)[39], using the default parameters. Diversity estimates were calculated by sample-size-based rarefaction and extrapolation using the R package iNEXT (iNterpolation/EXTrapolation)[40]. Clustering analysis was performed using the GLIPH2[41] webtool (<http://50.255.35.37:8080/>). Significant clusters were considered based on the following parameters: number of samples=3, number of CDR3>=3, vb_score<0.05, length_score<0.05. After filtering for significance, clusters were ordered based on final_score obtained from GLIPH2. Network graphs of clusters were produced using the R package igraph[42]. Unless indicated otherwise, analysis of differences was performed using Student t test. For multiple group comparisons, one-way anova was used. P-values <0.05 were considered statistically significant.

2.6 Availability of data

The TCR β sequencing data presented in this study have been deposited in NCBI's Gene Expression Omnibus (GEO) database under GEO: GSE156980. Both raw data and processed data are available.

3. Results

3.1 High-throughput TCR sequencing of SSc patients

To investigate the TCR repertoire dynamics in SSc, we performed high-throughput sequencing (HTS) of TCR β chains on longitudinal samples obtained from four SSc patients collected over a minimum of two years. The clinical characteristics of the SSc patients are included in Table 1. Among the SSc patients included, two were limited SSc (lcSSc), and two were diffuse SSc patients (dcSSc). From all patients PBLs were collected at baseline (T0), at least one year after inclusion (T1, ranging from 12-19 months) and at least two years after inclusion (T2, ranging from 24-46 months) (Figure 1a). Frozen PBL samples were subjected to FACS sorting and sorted non-naive CD4⁺ T-cells (CD3⁺CD4⁺CD45RO^{+/-}CD27^{+/-}) and non-naive CD8⁺ T-cells (CD3⁺CD8⁺CD45RO^{+/-}CD27^{+/-}) were used for TCR β sequencing (Figure 1b). Sample specific barcodes and UMIs to barcode individual mRNA molecules were used to accurately recover TCR β sequences using the RTCR pipeline[34]. We produced a total of 906 448 and 125 962 TCR β UMI corrected amino acid (AA) sequence reads for CD4⁺ and CD8⁺ T-cells respectively. The average number of UMI corrected AA sequence reads per sample was +/-75 000 for CD4⁺ T-cells and +/- 10 500 for CD8⁺ T-cells (details see Supplementary Table 2).

3.2 Frequency of V β and J β gene segment usage indicates temporal persistence of TCR β sequences in SSc patients

We first assessed the frequency of V β and J β gene segment usage in SSc patients over time (Figure 1c). The most frequently used V β segments across all samples were V20-1, V5-1 and V7-9, for both CD4⁺ and CD8⁺ T-cells. When looking at J β segment usage, J2-7, J2-1 and J2-3 were most frequently observed. In previous studies, these V β (V20-1, V5-1 and V7-9) and J β genes (J2-7, J2-1 and J2-3) were also identified as the most frequently used genes in both healthy and diseased individuals[43,44], reflecting known intrinsic biases in the V-D-J rearrangement process[45]. Additionally, V β 2, which we identified with a relatively high frequency in CD4⁺ and CD8⁺ T-cells, was previously found to be one of the most frequent V β chains in peripheral blood T-cells in SSc patients in another study[46], showing that disease specific patterns are also present. Furthermore, we also observed individual specific

patterns of V β and J β segments. As an example, SSc patient 2 displayed a lower frequency of V-28, V7-2 and J2-5 in both CD4+ and CD8+ T-cells across all time points compared to the other patients (Figure 1c).

In order to quantify the relative similarity of V β and J β gene segment usage across all samples, we calculated the Jensen-Shannon divergence (JSD) between them (Figure 1d). The JSD ranges from 0-1. A JSD of 0 indicates identical V β and J β segment usage, while a JSD of 1 indicates that the V β and J β segment usage is distinct between two samples. When comparing the JSD between samples taken from the same patient, we observed that the use of V β and J β segments for both CD4+ and CD8+ T-cells was rather consistent over time, while samples taken from different patients displayed a higher divergence amongst each other (Figure 1d). Moreover, when comparing the differences in JSD between V β and J β segment usage, we observed that V β usage (Figure 1d, right panels) was more variable than J β segment usage (Figure 1d, left panels) across different individuals. This difference in variability between V β and J β usage is to be expected since the TCR β locus has more V β than J β gene segments (according to the ImMunoGeneTics (IMGT) database[47]), resulting in a greater potential variability for V β segment usage. Overall this analysis shows that, V β and J β gene segment usage is largely persistent within SSc patients over time.

3.3 SSc TCR β repertoires are highly stable over time

Next to examining persistence in the use of V β and J β segments, we wanted to further examine the persistence of full CDR3 amino acid sequences within SSc patients over time. In order to quantify the overlap in TCR β repertoire between different samples, Morisita's overlap index was calculated to intersect amino acid CDR3 sequences. This index ranges from 0 (no overlapping sequences) to 1 (identical repertoires). Overlap analysis revealed that samples taken from the same patient shared a high number of sequences, indicating a clear temporal persistence of the repertoire within patients, while overlap was extremely limited between samples taken from different patients. This pattern was consistent over all time

points, for both CD4+ and CD8+ T-cells (Figure 2a and Figure 2b respectively). These results demonstrate that the TCR β repertoire in SSc patients is highly unique and stable over long periods of time, also at the level of exact CDR3 amino acid sequences.

Next we examined whether the frequencies of TCR β sequences were also consistent over time. We observed that TCR β sequences that were found with a high frequency at one time point were also found with a high frequency at the other time points collected from the same patient, for both CD4+ and CD8+ T-cells (Figure 2c and Figure 2d respectively). This shows that the frequencies of highly abundant TCR β s within SSc patients are largely consistent over long periods of time, even after almost four years of follow-up for SSc patient 4. Persistence of dominant TCR β sequences was observed for all the SSc patients included in our study, for both CD4+ and CD8+ T-cells (Supplementary Figure S1). The exact number of TCR β amino acid sequences overlapping between samples taken from the same patient are shown in Figure 2e. Although the absolute number of sequences that are overlapping within patients over time (i.e. persistent sequences) are low, they make up a substantial part of the samples in terms of abundance, based on UMI corrected reads, as shown in Figure 2f. These results clearly demonstrate that SSc repertoires are largely dominated by persistent sequences.

In order to investigate whether persistent TCR β sequences have any known antigen specificity, we queried the sequences that were persistently present in all three time points for every patient in VDJdb (a curated database of TCR sequences with known antigen specificities)[48]. The results of this analysis are shown in Table 2. In this table we show the hits for peptides presented on MHC II for CD4+ T-cells and peptides presented on MHC I for CD8+ T-cells. Hits were also found for peptides presented on MHC I for CD4+ T-cells and peptides presented on MHC II for CD8+ T-cells, even though technically these peptides cannot be recognized due to MHC restriction (Supplementary Table 3). For persistent TCR β s from CD4+ T-cells, we found most sequences to be associated with peptides presented on

MHC I (Supplementary Table 3). Since CD4+ T-cells only recognize peptides presented on MHC II, these findings probably do not represent the true antigenic specificity of the TCR β sequences queried. In general, the majority of the records present in VDJdb are based on TCR sequences obtained from MHC class I multimer assays.

For persistent TCR β s from CD8+ T-cells, we mainly found associations with peptides related to viral antigens including EBV, CMV and HIV (Table 2). For patient 2 and 3, TCR β sequences reactive against HIV-1 epitope B2M were identified, although these patients are not HIV positive. In SSc patient 2 we also identified one TCR β sequence (CASSLGQAYEQYF) that is associated with two human antigens, namely MLANA (melanoma antigen) and ABCD3 (ATP binding cassette subfamily D member 3). In patient 4 we identified one TCR β sequence (CASSLDLYEQYF) that is associated with the human antigen TKT (transketolase). Interestingly, MLANA is a known antigen that is widely expressed in skin, and the presence of MLANA-specific CD8+ T-cells has been associated with autoimmune reactions in melanoma patients treated with immune checkpoint inhibitors[49].

Overall, these results reveal a clear temporal persistence of clonally expanded CD4+ and CD8+ T-cells in SSc patients. These longitudinal dynamics show that the TCR β repertoire in SSc patients is highly stable over time, and this stability is potentially driven by a chronic response against (auto)antigens.

3.4 Persistent clones display common features across SSc patients

Previously, TCRs in the context of autoimmune disease have been associated with certain characteristics such as shorter length and a bias in V/J- gene segment usage[21,50,51]. To further investigate the potential involvement of persistent TCR β s identified in SSc patients in autoimmune responses, we computed the distribution of lengths of all the TCR β amino acid sequences and compared the lengths of persistent and non-persistent TCR β s. The lengths

were calculated based on both incidence (without weighing sequences by their abundance) and abundance (also taking into account the frequency of the sequences). A Gaussian distribution of lengths was observed for both persistent and non-persistent sequences, in CD4+ as well as CD8+ T-cells, when looking at incidence (Figure 3a), and distribution of lengths of sequences was similar between persistent and non-persistent sequences in all samples (two sample Kolmogorov-Smirnov tests >0.05 for all comparisons). When comparing the distribution of CDR3 lengths of persistent and non-persistent sequences based on abundance, we again observed no significant differences in the distributions neither for CD4+ nor CD8+ T-cells (two sample Kolmogorov-Smirnov tests >0.05 for all comparisons, Figure 3b). Although the CDR3 length distributions were not significantly different between persistent and non-persistent sequences, in SSc patient 1 and 2, more shorter length sequences were observed in the persistent TCR β s, while in SSc patient 4, longer sequences were found (Figure 3b). However, this skewness in lengths is mainly caused by the expansion of few dominant TCR β sequences.

Next to comparing lengths, we also compared the frequencies of V β and J β segment usage between persistent and non-persistent TCR sequences to see if there was any preferential usage (Figure 3c). Although most differences observed were small, we identified various V β and J β gene segments that had either higher or lower frequencies in persistent sequences across all SSc patients. As an example, TRBJ1-2 had a significantly lower frequency in persistent sequences in CD4+ as well as CD8+ T-cells across all SSc patients, while the frequency of TRBV7-2 and TRBV7-3 was higher in persistent sequences in CD4+ and CD8+ T-cells respectively (Figure 3c). This analysis demonstrates that, although the number of exact sequences that are shared between SSc patients is low, TCR β sequences that are persistently present in SSc patients over time show similarities in terms of V β and J β usage. Moreover, these are significantly different from non-persistent sequences, showing that persistent sequences display preferential usage of V β and J β segments across SSc patients. Given that similar TCR sequences are thought to be involved in T-cell responses to similar

antigens[52–55], the preferential segment V β and J β segment usage across SSc patients over time might reflect chronic immune responses against antigens that are commonly present across patients.

3.5 SSc TCR β repertoires are less diverse than healthy memory repertoires

The persistence of highly abundant TCRs is not necessarily unique to autoimmune repertoires and has in fact previously been observed in healthy individuals[36]. Therefore, we also compared our SSc data to a public dataset of longitudinal TCR β sequences from healthy donors[36]. To investigate whether TCR β s in SSc patient repertoires are aberrantly expanded compared to healthy repertoires, we compared the Shannon diversities of healthy memory repertoires to those of SSc. As the samples differed in their sizes, to compare the diversity between samples we performed rarefaction analysis. The healthy control dataset from Chu *et al.* included here did not contain singletons (sequences represented by one read). We therefore performed rarefaction analysis and estimated Shannon diversity in our SSc patient data by including and excluding singletons. In Figure 4a, we show the estimated Shannon diversity based on the rarefied and extrapolated data excluding singletons. The estimated species richness and Shannon diversity was significantly lower in SSc CD4+ and CD8+ T-cell repertoires as compared to the healthy memory repertoire (p-value <0.05, Figure 4b). To estimate the effect of excluding singletons on the diversity, the same analysis was also performed including the SSc singletons. In this analysis, the estimated Shannon diversity of TCR β sequences obtained from SSc CD8+ T-cells were still significantly lower than the diversity of healthy memory cells, while the significant difference in diversity with the CD4+ T-cells was no longer observed (Supplementary Figure S2). Overall, these results demonstrate that the TCR β repertoire diversity is lower in SSc patients compared to healthy individuals. This provides evidence for a skewed, clonally expanded repertoire in SSc, potentially due to chronic (auto-)antigen driven T-cell responses.

3.6 Persistent frequency of dominant TCR sequences is driven by antigenic selection rather than a-specific expansion

The hypothesis of 1) bystander activation due to chronic inflammation and 2) clonal expansion driven by the chronic presence of (auto-)antigens skewing the TCR repertoire have been proposed. To test the hypothesis of bystander activation, we investigated whether easy to generate TCR β sequences (having high generation probabilities) are present at high frequencies in SSc patients. To this extend, we calculated the generation probabilities (pgens) of the TCR β s identified in SSc patients using OLGA[39]. We then used linear regression to model the relationship between TCR β frequency and pgen. Using this analysis, we show that in SSc CD4+ and CD8+ T-cells, TCR β frequency and pgen are not related (p-value >0.05, Figure 5a and Figure 5b for CD4+ and CD8+ T-cells respectively). This indicates that T-cells persistent in SSc patients have not expanded because of random bystander effects.

For naive T-cells isolated from healthy individuals, we found a significant positive correlation between the pgen and abundance of TCR β s (p-value <0.05, Figure 5c). For memory T-cells isolated from the same healthy individuals, we also observed a significant positive correlation between pgen and abundance of TCR β s (p-value <0.05, Figure 5d). Notably, the slope for memory T-cells was lower than that observed for the naïve T-cells (0.00018 versus 0.0165, respectively). From this analysis we show that for naïve and memory TCR β s obtained from healthy individuals, pgen and abundance are positively correlated, whereas in non-naïve SSc T-cells no relationship between pgen and abundance is observed. However, as the samples obtained from healthy individuals were sequenced more deeply than our SSc samples, this observed difference in correlation might be confounded by sequencing depth. Therefore, we repeated the linear regression analysis for 100 random subsamples obtained from the healthy dataset and compared these results to the results obtained from SSc samples. Upon subsampling of the naïve healthy T-cells, 89% of the slopes observed the in linear regression were significantly different from zero (linear regression p-value <0.05 and slope >0, as

indicated by the red bars in Figure 5e), confirming that indeed there is a positive relationship between pgen and abundance for TCR β s obtained from healthy naïve T-cells. However, when looking at healthy memory T-cells, we did not confirm the correlation between frequency and pgen that was observed in the full sample (p-value <0.05 in 7% of subsamples, Figure 5f), showing that in healthy memory T-cells there is no clear correlation between TCR β pgen and abundance. Thus, subsampling analysis shows that in both healthy memory T-cells and SSc non-naïve CD4+ and CD8+ T-cells, TCR β pgen and abundance are not correlated. These results suggest that the more abundant TCR β s in SSc repertoires are likely there because of selection, similar to what is observed in healthy memory repertoires, rather than a-specific expansion of naïve T-cells.

3.7 Clusters of similar TCR β sequences can be identified in SSc patients over time

In order to identify TCR β s in SSc patients that are potentially involved in chronic autoimmune responses, we used “Grouping of Lymphocyte Interactions by Paratope Hotspots” (GLIPH2), that employs sequence similarity and motif analysis to group TCR sequences that potentially recognize the same epitope[41]. To screen for antigen specific TCR clusters, we used all sequences obtained from every time point for each individual SSc patient as an input for GLIPH2. In order to exclude false positives, for each patient we considered the clusters that contained sequences from all time points, had at least three unique CDR3s, had similar CDR3 lengths (length score <0.05), and shared similar V β -gene frequency distributions (V β score <0.05). The number of clusters that were obtained for every patient for CD4+ and CD8+ T-cells are given in Table 3.

Significant clusters were identified in all patients, either based on global similarity (CDR3 sequence differing by maximum one amino acid) or local similarity (shared motif within CDR3 amino acid region). All clusters identified by GLIPH2 are given in Supplementary Table 4. In Figure 6a, an example network of TCR β sequences based on clustering analysis by GLIPH2 in CD4+ T-cells is given. Red and purple nodes within this network represent TCR β s that are

persistently present in three or two time points respectively, while blue shaded nodes represent sequences are found only in one time point. Nodes are connected when they are part of the same cluster as identified by GLIPH2. Global and local motifs are indicated in green and orange respectively. A “%” sign within the motif indicates a variable amino acid across the sequences in which that particular motif is present.

Within the network, persistent TCRβs share motifs with other persistent and non-persistent TCRβs sequences. This shows that persistent TCRβ sequences cluster together with other, similar TCRβ sequences, potentially representing groups of T-cells responding to the same antigen. The top 10 clusters, based on final cluster score outputted by GLIPH2, and their corresponding motifs for this network are given in Figure 6b. Some clusters also shared TCRβ sequences with other clusters within a patient, showing that clusters also display convergence between each other (Figure 6c, top 3 clusters are shown). Similar results were obtained for CD8+ T-cells, where we also identified many clusters of persistent and non-persistent TCRβs sharing motifs within their CDR3 region (Figure 6d-f). These results demonstrate that, apart from the presence of individual persistent clonally expanded T-cells, clusters of T-cells with potentially similar specificities are present within SSc patients over time.

Lastly, we performed an overlap analysis of all motifs from significant clusters identified by GLIPH2 between the different SSc patients (Figure 6g and Figure 6h). We did not observe any motifs for CD4+ T-cells or for CD8+ T-cells that overlap between all four patients. For the CD4+ T-cells, there were 11 motifs that were identified in clusters from three out of four SSc patients (figure 6g). These represent groups of T-cells that are likely to recognize the same or highly similar antigens across SSc patients, which could potentially be involved in SSc pathogenesis.

4. Discussion

Identification of TCR sequences that are associated with the chronic autoimmune response in SSc will help to get more insights into the autoimmune pathogenesis of the disease, and will help to identify the antigenic triggers that underlie these responses. Our analysis reveals that the peripheral blood TCR β repertoire of SSc patients is highly stable over time. Moreover, the TCR β sequences that were found within a patient with a high frequency at one time point were also found with a high frequency at the other time points (even after four years), showing that frequencies of dominant TCR β s are largely consistent over time. These persistent, clonally expanded CD4+ and CD8+ T-cells are potentially involved in the autoimmune responses underlying SSc pathogenesis. Furthermore, we have shown that the persistent expansion of these T-cells is likely a result of antigenic selection rather than recombination bias, as TCR β frequencies were not related to their respective generation probabilities.

When we queried the persistent TCR β s found in our SSc patients in VDJdb, we obtained several hits for TCR β sequences that are known to be associated with viral antigens from HIV, CMV and EBV, especially in the CD8+ T-cell compartment. However, as the SSc patients included in our study are all HIV negative, these hits could be potentially false positive. SSc patient 4 has a CMV and EBV positive status, and TCR β sequences associated with CMV and EBV, as well as one TCR β sequence (CASSLDLYEQYF) associated with the human autoantigen TKT were identified in this patient. For the other SSc patients included in this study, the CMV and EBV status are unknown. Interestingly, in SSc patient 2 we identified CASSLGQAYEQYF as a persistent sequence that is associated with EBV epitopes BMLF1 and EBNA3, as well as the human autoantigens MLANA and ABCD3. This TCR β sequence might thus be cross-reactive to EBV and human autoantigen epitopes, potentially representing molecular mimicry. In fact, molecular mimicry between chronic viral antigens and human autoantigens has been proposed as a potential driver for autoimmune disease[56], and EBV and CMV infections have been shown to be environmental risk factors for SSc[57–60].

Chu *et al.* have previously shown that subsets of persistent TCR β s are also present within healthy individuals[36]. Thus, persistence of TCR sequences in itself is likely not just a characteristic of autoimmune related repertoires. However, the temporal dynamics of the TCR repertoire in healthy individuals in the aforementioned study have only been investigated over a period of one year, so it remains to be seen whether this stability is also maintained in healthy individuals over longer periods of time, as is observed in SSc in our study. Moreover, whereas the previous study looked into the total pool of memory T-cells, we show that persistent TCR β s can be identified in both the CD4+ and CD8+ memory T-cell compartments separately.

When further comparing the TCR repertoires of CD4+ and CD8+ T-cells from SSc patients to repertoires obtained from healthy memory T-cells, we found that SSc repertoires have lower diversity. Indeed, decreased diversities of TCR β repertoires as compared to healthy have been observed in other autoimmune diseases[20,44,61], and have been proposed as a characteristic of autoimmune repertoires. Interestingly, in SSc, differences in the diversity of the TCR repertoire have also been observed between responders and non-responders after autologous hematopoietic stem-cell transplantation (AH SCT, the only therapy with long-term clinical benefit in SSc), with non-responders having a less diverse repertoire[62]. This provides further evidence that decreased TCR repertoire diversity contributes to the autoimmune pathogenesis of SSc.

Predicting T-cell reactivity towards antigens is one of the major areas currently investigated in the field of TCR research. Since prediction of TCR binding to a specific antigen is extremely challenging, current efforts are more focused on identifying groups of TCRs that contain certain motifs within their CDR3 region[52–55]. These groups of TCRs comprise of clones that potentially respond to the same antigen. Apart from exact sharing of TCR β sequences between samples, we also identified clusters of TCR β s that share sequence

motifs and were persistent within patients. This indicates that antigen selection reshapes the TCR β repertoire in SSc. Potential antigens could include self-antigens (such as MLANA as associated TCR β sequences were found to be persistent within SSc patients studied here), or chronic infections with pathogens (e.g., CMV or EBV, for which we also identified associated persistent TCR β s). Interestingly, we also found clusters of TCR β s from CD4+ T-cells within patients that shared motifs with other TCR β clusters between patients. These could represent clusters of similar TCR β s that might contribute to more public autoimmune responses underlying SSc pathogenesis. For CD8+ T-cells, we did not find any clusters overlapping between more than two patients. Notably, in general we obtained less clusters in CD8+ T-cells than we did in CD4+ T-cells. This could be due to the fact that we sequenced less CD8+ than CD4+ T-cells, and thus this difference might be explained by a difference in sequencing depth between the two cell types.

A limitation of our study is that HLA-genotype information from the SSc patients included was not obtained, so the contribution of HLA to the TCR β clusters and shaping of the repertoire in these patients remains unknown. To validate our findings and further study the presence of potential pathogenic role of antigen specific TCR β clusters in SSc, larger patient cohorts should be studied. In this cohort we included a limited number of patients with similar clinical characteristics which makes it difficult to account for factors such as age, sex and ethnicity influencing the immune system. Thus, studying larger longitudinal cohorts are needed to further define disease specific clusters of autoimmune associated TCR β s. Lastly, it would also be interesting to perform immune sequencing of SSc skin to see if these TCR β clusters/motifs can be traced back in the skin (the major organ affected by the disease) of SSc patients.

5. Conclusion

Our data provide evidence for an (auto-)antigen driven expansion of CD4/CD8+ T-cells in SSc. We have identified clusters of T-cell clones that are highly persistent over time, and we have shown that this persistence likely is a result of antigenic selection.

Acknowledgements

We thank all the Systemic Sclerosis patients included in this study who graciously donated their time and samples. Furthermore we thank Sanne Hiddingh and the Core Flow cytometry Facility (CFF) at the UMC Utrecht for their help in sorting the T-cells samples included in this study, and Wilfred van IJcken and the Erasmus Center Biomics team (Erasmus MC Rotterdam) for Illumina sequencing. We are grateful to Anneline Hinrichs for her help with gathering of clinical data, and Rob de Boer and Peter de Greef from the Theoretical Biology & Bioinformatics department of Utrecht University for their valuable input and discussions of the data.

Funding

The work was supported by the Center of Translational Immunology at the UMC Utrecht, which is a non-profit research organization. A.P. was supported by Netherlands Organization for Scientific Research (NWO) grant number: 016.Veni.178.027. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

N.H. Servaas: conceptualization, formal analysis, investigation, writing - original draft, writing - review & editing, visualization **F. Zaaraoui-Boutahar:** investigation, writing - review & editing **C.G.K. Wichers:** investigation, writing - review & editing **A. Ottria:** investigation, writing - review & editing **E. Chouri:** investigation, writing - review & editing **A.J. Affandi:**

investigation, writing - review & editing **S. Silva-Cardoso**: investigation, writing - review & editing **M. van der Kroef**: investigation, writing - review & editing **T. Carneiro**: investigation, writing - review & editing **F. van Wijk**: writing - review & editing **T.R.D.J. Radstake**: writing - review & editing, funding acquisition **A.C. Andeweg**: methodology, investigation, resources, writing - review & editing **A. Pandit**: conceptualization, methodology, writing - original draft, writing - review & editing, supervision, project administration, funding acquisition. All authors critically reviewed and approved the manuscript.

Disclosure of conflicts of interest

The authors declare no conflict of interest.

References

- [1] A. Gabrielli, E. V Avvedimento, T. Krieg, Scleroderma., *N. Engl. J. Med.* 360 (2009) 1989–2003. <https://doi.org/10.1056/NEJMra0806188>.
- [2] A.D. Roumm, T.L. Whiteside, T.A. Medsger, G.P. Rodnan, Lymphocytes in the skin of patients with progressive systemic sclerosis, *Arthritis Rheum.* 27 (1984) 645–653. <https://doi.org/10.1002/art.1780270607>.
- [3] A. Kalogerou, E. Gelou, S. Mountantonakis, L. Settas, E. Zafiriou, L. Sakkas, Early T cell activation in the skin from patients with systemic sclerosis, *Ann. Rheum. Dis.* 64 (2005) 1233–1235. <https://doi.org/10.1136/ard.2004.027094>.
- [4] R. de Palma, E. D’aiuto, S. Vettori, P.P. Cuoppolo, G. Abbate, G. Valentini, Peripheral T cells from patients with early systemic sclerosis kill autologous fibroblasts in co-culture: Is T-cell response aimed to play a protective role?, *Rheumatology.* (2010). <https://doi.org/10.1093/rheumatology/keq094>.
- [5] T. Hügler, S. O’Reilly, R. Simpson, M.D. Kraaij, V. Bigley, M. Collin, A. Krippner-Heidenreich, J.M. Van Laar, Tumor necrosis factor-costimulated T lymphocytes from patients with systemic sclerosis trigger collagen production in fibroblasts, *Arthritis*

- Rheum. 65 (2013) 481–491. <https://doi.org/10.1002/art.37738>.
- [6] R. De Palma, F. Del Galdo, S. Lupoli, P. Altucci, G.F. Abbate, G. Valentini, Peripheral T lymphocytes from patients with early systemic sclerosis co-cultured with autologous fibroblasts undergo an oligoclonal expansion similar to that occurring in the skin, *Clin. Exp. Immunol.* 144 (2006) 169–176. <https://doi.org/10.1111/j.1365-2249.2006.03041.x>.
- [7] A. Kahan, A. Kahan, F. Picard, C.J. Menkes, B. Amor, Abnormalities of T lymphocyte subsets in systemic sclerosis demonstrated with anti-CD45RA and anti-CD29 monoclonal antibodies, *Ann. Rheum. Dis.* 50 (1991) 354–358. <https://doi.org/10.1136/ard.50.6.354>.
- [8] L. Parolin Ercole, M. Malvezzi, A.C. Boaretti, S. Ramos da Rosa Utiyama, A. Rachid, Analysis of lymphocyte subpopulations in systemic sclerosis, *J. Investig. Allergol. Clin. Immunol.* (2003).
- [9] G. Papp, I.F. Horvath, S. Barath, E. Gyimesi, S. Sipka, P. Szodoray, M. Zeher, Altered T-cell and regulatory cell repertoire in patients with diffuse cutaneous systemic sclerosis, *Scand. J. Rheumatol.* 40 (2011) 205–210. <https://doi.org/10.3109/03009742.2010.528021>.
- [10] P. Fuschiotti, A.T. Larregina, J. Ho, C. Feghali-Bostwick, T.A. Medsger, Interleukin-13-producing CD8+ T cells mediate dermal fibrosis in patients with systemic sclerosis, *Arthritis Rheum.* 65 (2013) 236–246. <https://doi.org/10.1002/art.37706>.
- [11] M.E. Truchetet, N.C. Brembilla, E. Montanari, Y. Allanore, C. Chizzolini, Increased frequency of circulating Th22 in addition to Th17 and Th2 lymphocytes in systemic sclerosis: Association with interstitial lung disease, *Arthritis Res. Ther.* 13 (2011) R166. <https://doi.org/10.1186/ar3486>.
- [12] S. O'Reilly, T. Hügle, J.M. Van Laar, T cells in systemic sclerosis: A reappraisal, *Rheumatol. (United Kingdom)*. (2012). <https://doi.org/10.1093/rheumatology/kes090>.
- [13] M.M. Davis, P.J. Bjorkman, T-cell antigen receptor genes and T-cell recognition, *Nature.* 334 (1988) 395–402. <https://doi.org/10.1038/334395a0>.

- [14] C.H. Bassing, W. Swat, F.W. Alt, The mechanism and regulation of chromosomal V(D)J recombination, *Cell*. (2002). [https://doi.org/10.1016/S0092-8674\(02\)00675-X](https://doi.org/10.1016/S0092-8674(02)00675-X).
- [15] H.S. Robins, S.K. Srivastava, P. V. Campregher, C.J. Turtle, J. Andriesen, S.R. Riddell, C.S. Carlson, E.H. Warren, Overlap and effective size of the human CD8+ T cell receptor repertoire, *Sci. Transl. Med.* (2010). <https://doi.org/10.1126/scitranslmed.3001442>.
- [16] T.P. Arstila, A. Casrouge, V. Baron, J. Even, J. Kanellopoulos, P. Kourilsky, A direct estimate of the human $\alpha\beta$ T cell receptor diversity, *Science* (80-.). (1999). <https://doi.org/10.1126/science.286.5441.958>.
- [17] Q. Qi, Y. Liu, Y. Cheng, J. Glanville, D. Zhang, J.Y. Lee, R.A. Olshen, C.M. Weyand, S.D. Boyd, J.J. Goronzy, Diversity and clonal selection in the human T-cell repertoire, *Proc. Natl. Acad. Sci. U. S. A.* (2014). <https://doi.org/10.1073/pnas.1409155111>.
- [18] T. Mora, A.M. Walczak, How many different clonotypes do immune repertoires contain?, *Curr. Opin. Syst. Biol.* 18 (2019) 104–110. <https://doi.org/10.1016/j.coisb.2019.10.001>.
- [19] Y. Zhao, P. Nguyen, J. Ma, T. Wu, L.L. Jones, D. Pei, C. Cheng, T.L. Geiger, Preferential Use of Public TCR during Autoimmune Encephalomyelitis, *J. Immunol.* (2016). <https://doi.org/10.4049/jimmunol.1501029>.
- [20] C.M. Chang, Y.W. Hsu, H.S.C. Wong, J.C.C. Wei, X. Liu, H.T. Liao, W.C. Chang, Characterization of T-Cell Receptor Repertoire in Patients with Rheumatoid Arthritis Receiving Biologic Therapies, *Dis. Markers.* (2019). <https://doi.org/10.1155/2019/2364943>.
- [21] X. Liu, W. Zhang, M. Zhao, L. Fu, L. Liu, J. Wu, S. Luo, L. Wang, Z. Wang, L. Lin, Y. Liu, S. Wang, Y. Yang, L. Luo, J. Jiang, X. Wang, Y. Tan, T. Li, B. Zhu, Y. Zhao, X. Gao, Z. Wan, C. Huang, M. Fang, Q. Li, H. Peng, X. Liao, J. Chen, F. Li, G. Ling, H. Zhao, H. Luo, Z. Xiang, J. Liao, Y. Liu, H. Yin, H. Long, H. Wu, H. Yang, J. Wang, Q. Lu, T cell receptor β repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis, *Ann. Rheum. Dis.* 78 (2019) 1070–1078.

<https://doi.org/10.1136/annrheumdis-2019-215442>.

- [22] J.J. Goronzy, C.M. Weyand, T-cell senescence and contraction of T-cell repertoire diversity - Catalysts of autoimmunity and chronic inflammation, *Arthritis Res. Ther.* (2003). <https://doi.org/10.1186/ar974>.
- [23] L.I. Sakkas, B. Xu, C.M. Artlett, S. Lu, S.A. Jimenez, C.D. Platsoucas, Oligoclonal T Cell Expansion in the Skin of Patients with Systemic Sclerosis, *J. Immunol.* 168 (2002) 3649–3659. <https://doi.org/10.4049/jimmunol.168.7.3649>.
- [24] V. V. Yurovsky, F.M. Wigley, R.A. Wise, B. White, Skewing of the CD8+ T-cell repertoire in the lungs of patients with systemic sclerosis, *Hum. Immunol.* (1996). [https://doi.org/10.1016/0198-8859\(96\)00091-2](https://doi.org/10.1016/0198-8859(96)00091-2).
- [25] B. White, V. V Yurovsky, Oligoclonal expansion of V delta 1+ gamma/delta T-cells in systemic sclerosis patients, *Ann N Y Acad Sci.* (1995).
- [26] M. Attaf, E. Huseby, A.K. Sewell, $\alpha\beta$ T cell receptors as predictors of health and disease, *Cell. Mol. Immunol.* (2015). <https://doi.org/10.1038/cmi.2014.134>.
- [27] Y. Pacheco, Y. Acosta-Ampudia, D.M. Monsalve, C. Chang, M.E. Gershwin, J.-M. Anaya, Bystander activation and autoimmunity, (2019). <https://doi.org/10.1016/j.jaut.2019.06.012>.
- [28] H.G. Lee, J.U. Lee, D.H. Kim, S. Lim, I. Kang, J.M. Choi, Pathogenic function of bystander-activated memory-like CD4 + T cells in autoimmune encephalomyelitis, *Nat. Commun.* 10 (2019) 1–14. <https://doi.org/10.1038/s41467-019-08482-w>.
- [29] N.J. Burroughs, M. Ferreira, B.M.P.M. Oliveira, A.A. Pinto, Autoimmunity arising from bystander proliferation of T cells in an immune response model, *Math. Comput. Model.* (2011). <https://doi.org/10.1016/j.mcm.2010.01.020>.
- [30] V. Venturi, K. Kedzierska, D.A. Price, P.C. Doherty, D.C. Douek, S.J. Turner, M.P. Davenport, Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination, *Proc. Natl. Acad. Sci. U. S. A.* (2006). <https://doi.org/10.1073/pnas.0608907103>.
- [31] F. Van Den Hoogen, D. Khanna, J. Franssen, S.R. Johnson, M. Baron, A. Tyndall, M.

- Matucci-Cerinic, R.P. Naden, T.A. Medsger, P.E. Carreira, G. Riemekasten, P.J. Clements, C.P. Denton, O. Distler, Y. Allanore, D.E. Furst, A. Gabrielli, M.D. Mayes, J.M. Van Laar, J.R. Seibold, L. Czirjak, V.D. Steen, M. Inanc, O. Kowal-Bielecka, U. Müller-Ladner, G. Valentini, D.J. Veale, M.C. Vonk, U.A. Walker, L. Chung, D.H. Collier, M.E. Csuka, B.J. Fessler, S. Guiducci, A. Herrick, V.M. Hsu, S. Jimenez, B. Kahaleh, P.A. Merkel, S. Sierakowski, R.M. Silver, R.W. Simms, J. Varga, J.E. Pope, 2013 classification criteria for systemic sclerosis: An American college of rheumatology/European league against rheumatism collaborative initiative, *Ann. Rheum. Dis.* 72 (2013) 1747–1755. <https://doi.org/10.1136/annrheumdis-2013-204424>.
- [32] I.Z. Mamedov, O. V. Britanova, I. V. Zvyagin, M.A. Turchaninova, D.A. Bolotin, E. V. Putintseva, Y.B. Lebedev, D.M. Chudakov, Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling, *Front. Immunol.* 4 (2013) 1–10. <https://doi.org/10.3389/fimmu.2013.00456>.
- [33] J. Zhang, K. Kobert, T. Flouri, A. Stamatakis, PEAR: A fast and accurate Illumina Paired-End reAd mergeR, *Bioinformatics.* (2014). <https://doi.org/10.1093/bioinformatics/btt593>.
- [34] B. Gerritsen, A. Pandit, A.C. Andeweg, R.J. De Boer, RTCR: A pipeline for complete and accurate recovery of T cell repertoires from high throughput sequencing data, *Bioinformatics.* 32 (2016) 3098–3106. <https://doi.org/10.1093/bioinformatics/btw339>.
- [35] V.I. Nazarov, M. V. Pogorelyy, E.A. Komech, I. V. Zvyagin, D.A. Bolotin, M. Shugay, D.M. Chudakov, Y.B. Lebedev, I.Z. Mamedov, tcR: An R package for T cell receptor repertoire advanced data analysis, *BMC Bioinformatics.* 16 (2015) 175. <https://doi.org/10.1186/s12859-015-0613-1>.
- [dataset] [36] N.D. Chu, H.S. Bi, R.O. Emerson, A.M. Sherwood, M.E. Birnbaum, H.S. Robins, E.J. Alm, Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in highly public receptors, *BMC Immunol.* 20 (2019) 19. <https://doi.org/10.1186/s12865-019-0300-5>.

- [37] R Development Core Team, R: A language and environment for statistical computing., Vienna, Austria. (2017). <https://doi.org/10.1007/978-3-900051-07-0> R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- [38] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, 2016.
- [39] Z. Sethna, Y. Elhanati, C.G. Callan, A.M. Walczak, T. Mora, OLGA: Fast computation of generation probabilities of B-and T-cell receptor amino acid sequences and motifs, *Bioinformatics*. 35 (2019) 2974–2981. <https://doi.org/10.1093/bioinformatics/btz035>.
- [40] T.C. Hsieh, K.H. Ma, A. Chao, iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers), *Methods Ecol. Evol.* (2016). <https://doi.org/10.1111/2041-210X.12613>.
- [41] H. Huang, C. Wang, F. Rubelt, T.J. Scriba, M.M. Davis, Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening, *Nat. Biotechnol.* (2020) 1–9. <https://doi.org/10.1038/s41587-020-0505-4>.
- [42] G. Csardi, T. Nepusz, The igraph software package for complex network research, *InterJournal Complex Syst.* (2006).
- [43] L. Ma, L. Yang, B. Shi, X. He, A. Peng, Y. Li, T. Zhang, S. Sun, R. Ma, X. Yao, Analyzing the CDR3 Repertoire with respect to TCR - Beta Chain V-D-J and V-J Rearrangements in Peripheral T Cells using HTS, *Sci. Rep.* 6 (2016) 1–10. <https://doi.org/10.1038/srep29544>.
- [44] J.H. Cui, Y. bin Jin, K.R. Lin, P. Xiao, X. ping Chen, Y. ming Pan, W. Lin, Z. chang Wu, D. mei Guo, X. fan Mao, C. ling Zhang, W. lue Lian, W. Luo, Characterization of peripheral blood TCR repertoire in patients with ankylosing spondylitis by high-throughput sequencing, *Hum. Immunol.* 79 (2018) 485–490. <https://doi.org/10.1016/j.humimm.2018.03.007>.
- [45] J.E. Park, R.A. Botting, C.D. Conde, D.M. Popescu, M. Lavaert, D.J. Kunz, I. Goh, E. Stephenson, R. Ragazzini, E. Tuck, A. Wilbrey-Clark, K. Roberts, V.R. Kedlian, J.R.

- Ferdinand, X. He, S. Webb, D. Maunder, N. Vandamme, K.T. Mahbubani, K. Polanski, L. Mamanova, L. Bolt, D. Crossland, F. De Rita, A. Fuller, A. Filby, G. Reynolds, D. Dixon, K. Saeb-Parsy, S. Lisgo, D. Henderson, R. Vento-Tormo, O.A. Bayraktar, R.A. Barker, K.B. Meyer, Y. Saeys, P. Bonfanti, S. Behjati, M.R. Clatworthy, T. Taghon, M. Haniffa, S.A. Teichmann, A cell atlas of human thymic development defines T cell repertoire formation, *Science* (80-.). (2020). <https://doi.org/10.1126/science.aay3224>.
- [46] D. Farge, L.C.M. Arruda, F. Brigant, E. Clave, C. Douay, Z. Marjanovic, C. Deligny, G. Maki, E. Gluckman, A. Toubert, H. Moins-Teisserenc, Long-term immune reconstitution and T cell repertoire analysis after autologous hematopoietic stem cell transplantation in systemic sclerosis patients., *J. Hematol. Oncol.* 10 (2017) 21. <https://doi.org/10.1186/s13045-016-0388-5>.
- [47] M.P. Lefranc, IMGT, the international ImMunoGeneTics database®, *Nucleic Acids Res.* (2003). <https://doi.org/10.1093/nar/gkg085>.
- [48] D. V. Bagaev, R.M.A. Vroomans, J. Samir, U. Stervbo, C. Rius, G. Dolton, A. Greenshields-Watson, M. Attaf, E.S. Egorov, I. V. Zvyagin, N. Babel, D.K. Cole, A.J. Godkin, A.K. Sewell, C. Kesmir, D.M. Chudakov, F. Luciani, M. Shugay, VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium, *Nucleic Acids Res.* 48 (2020) D1057–D1062. <https://doi.org/10.1093/nar/gkz874>.
- [49] O. Klein, L.M. Ebert, T. Nicholaou, J. Browning, S.E. Russell, M. Zuber, H.M. Jackson, N. Dimopoulos, B.S. Tan, A. Hoos, I.F. Luescher, I.D. Davis, W. Chen, J. Cebon, Melan-a-specific cytotoxic T cells are associated with tumor regression and autoimmunity following treatment with anti-CTLA-4, *Clin. Cancer Res.* 15 (2009) 2507–2513. <https://doi.org/10.1158/1078-0432.CCR-08-2424>.
- [50] X. Hou, P. Zeng, X. Zhang, J. Chen, Y. Liang, J. Yang, Y. Yang, X. Liu, H. Diao, Shorter TCR β -Chains Are Highly Enriched During Thymic Selection and Antigen-Driven Selection, *Front. Immunol.* 10 (2019) 299. <https://doi.org/10.3389/fimmu.2019.00299>.

- [51] I. Gomez-Tourino, Y. Kamra, R. Baptista, A. Lorenc, M. Peakman, T cell receptor β -chains display abnormal shortening and repertoire sharing in type 1 diabetes, *Nat. Commun.* 8 (2017) 1–15. <https://doi.org/10.1038/s41467-017-01925-2>.
- [52] M. V. Pogorelyy, A.A. Minervina, M. Shugay, D.M. Chudakov, Y.B. Lebedev, T. Mora, A.M. Walczak, Detecting T cell receptors involved in immune responses from single repertoire snapshots, *PLoS Biol.* 17 (2019). <https://doi.org/10.1371/journal.pbio.3000314>.
- [53] J. Glanville, H. Huang, A. Nau, O. Hatton, L.E. Wagar, F. Rubelt, X. Ji, A. Han, S.M. Krams, C. Pettus, N. Haas, C.S.L. Arlehamn, A. Sette, S.D. Boyd, T.J. Scriba, O.M. Martinez, M.M. Davis, Identifying specificity groups in the T cell receptor repertoire, *Nature.* 547 (2017) 94–98. <https://doi.org/10.1038/nature22976>.
- [54] P. Dash, A.J. Fiore-Gartland, T. Hertz, G.C. Wang, S. Sharma, A. Souquette, J.C. Crawford, E.B. Clemens, T.H.O. Nguyen, K. Kedzierska, N.L. La Gruta, P. Bradley, P.G. Thomas, Quantifiable predictive features define epitope-specific T cell receptor repertoires, *Nature.* (2017). <https://doi.org/10.1038/nature22383>.
- [55] A. Madi, A. Poran, E. Shifrut, S. Reich-Zeliger, E. Greenstein, I. Zaretsky, T. Arnon, F. Van Laethem, A. Singer, J. Lu, P.D. Sun, I.R. Cohen, N. Friedman, T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences, *Elife.* (2017). <https://doi.org/10.7554/eLife.22057>.
- [56] M.F. Cusick, J.E. Libbey, R.S. Fujinami, Molecular mimicry as a mechanism of autoimmune disease, *Clin. Rev. Allergy Immunol.* (2012). <https://doi.org/10.1007/s12016-011-8294-7>.
- [57] C. Lunardi, C. Bason, R. Navone, E. Millo, G. Damonte, R. Corrocher, A. Puccetti, Systemic sclerosis immunoglobulin G autoantibodies bind the human cytomegalovirus late protein UL94 and induce apoptosis in human endothelial cells, *Nat. Med.* (2000). <https://doi.org/10.1038/80533>.
- [58] Y. Arnsion, H. Amital, S. Guiducci, M. Matucci-Cerinic, G. Valentini, O. Barzilai, R. Maya, Y. Shoenfeld, The role of infections in the immunopathogenesis of systemic

- sclerosis-evidence from serological studies, in: *Ann. N. Y. Acad. Sci.*, 2009.
<https://doi.org/10.1111/j.1749-6632.2009.04808.x>.
- [59] L. Sternbæk, A.H. Draborg, M.T. Østerlund, L. V. Iversen, L. Troelsen, E. Theander, C.T. Nielsen, S. Jacobsen, G. Houen, Increased antibody levels to stage-specific Epstein–Barr virus antigens in systemic autoimmune diseases reveal a common pathology, *Scand. J. Clin. Lab. Invest.* (2019).
<https://doi.org/10.1080/00365513.2018.1550807>.
- [60] L.C.M. Arruda, E. Clave, C. Douay, J.R. Lima-Júnior, S.N. Slavov, K.C.R. Malmegrim, A.J. Alberdi, M.C. Oliveira, A. Toubert, CMV-specific clones may lead to reduced TCR diversity and relapse in systemic sclerosis patients treated with AHSCT, *Rheumatology*. (2020). <https://doi.org/10.1093/rheumatology/keaa001>.
- [61] D.R. Thapa, R. Tonikian, C. Sun, M. Liu, A. Dearth, M. Petri, F. Pepin, R.O. Emerson, A. Ranger, Longitudinal analysis of peripheral blood T cell receptor diversity in patients with systemic lupus erythematosus by next-generation sequencing, *Arthritis Res. Ther.* (2015). <https://doi.org/10.1186/s13075-015-0655-9>.
- [62] L.C.M. Arruda, K.C.R. Malmegrim, J.R. Lima-Júnior, E. Clave, J.B.E. Dias, D.A. Moraes, C. Douay, I. Fournier, H. Moins-Teisserenc, A.J. Alberdi, D.T. Covas, B.P. Simões, P. Lansiaux, A. Toubert, M.C. Oliveira, Immune rebound associates with a favorable clinical response to autologous HSCT in systemic sclerosis patients, *Blood Adv.* 2 (2018) 126–141. <https://doi.org/10.1182/bloodadvances.2017011072>.

Patient	Timepoint	Time from T0	Subset	Age	Sex	ILD	ANA	mRSS
SSc 1	T0	0	dcSSc	60's	F	Yes	Yes	11
	T1	19	dcSSc		F	Yes	Yes	11
	T2	34	dcSSc		F	Yes	Yes	11
SSc 2	T0	0	dcSSc	30's	M	No	Yes	16
	T1	12	dcSSc		M	No	Yes	13
	T2	25	dcSSc		M	No	Yes	13
SSc 3	T0	0	lcSSc	40's	F	Yes	Yes	4
	T1	18	lcSSc		F	Yes	Yes	4
	T2	42	lcSSc		F	Yes	Yes	8
SSc 4	T0	0	lcSSc	50's	F	Yes	Yes	2
	T1	15	lcSSc		F	Yes	Yes	2
	T2	46	lcSSc		F	Yes	Yes	2

Table 1. Clinical characteristics of patients included. Time from T0 is indicated in months after the first sample was taken. Abbreviations: dcSSc=diffuse SSc, lcSSc=limited SSc, M=male, F=female, ILD=interstitial lung disease, ANA= anti-nuclear antibodies, mRSS= modified rodman skin score. Age is age range between inclusion (T0) and T2.

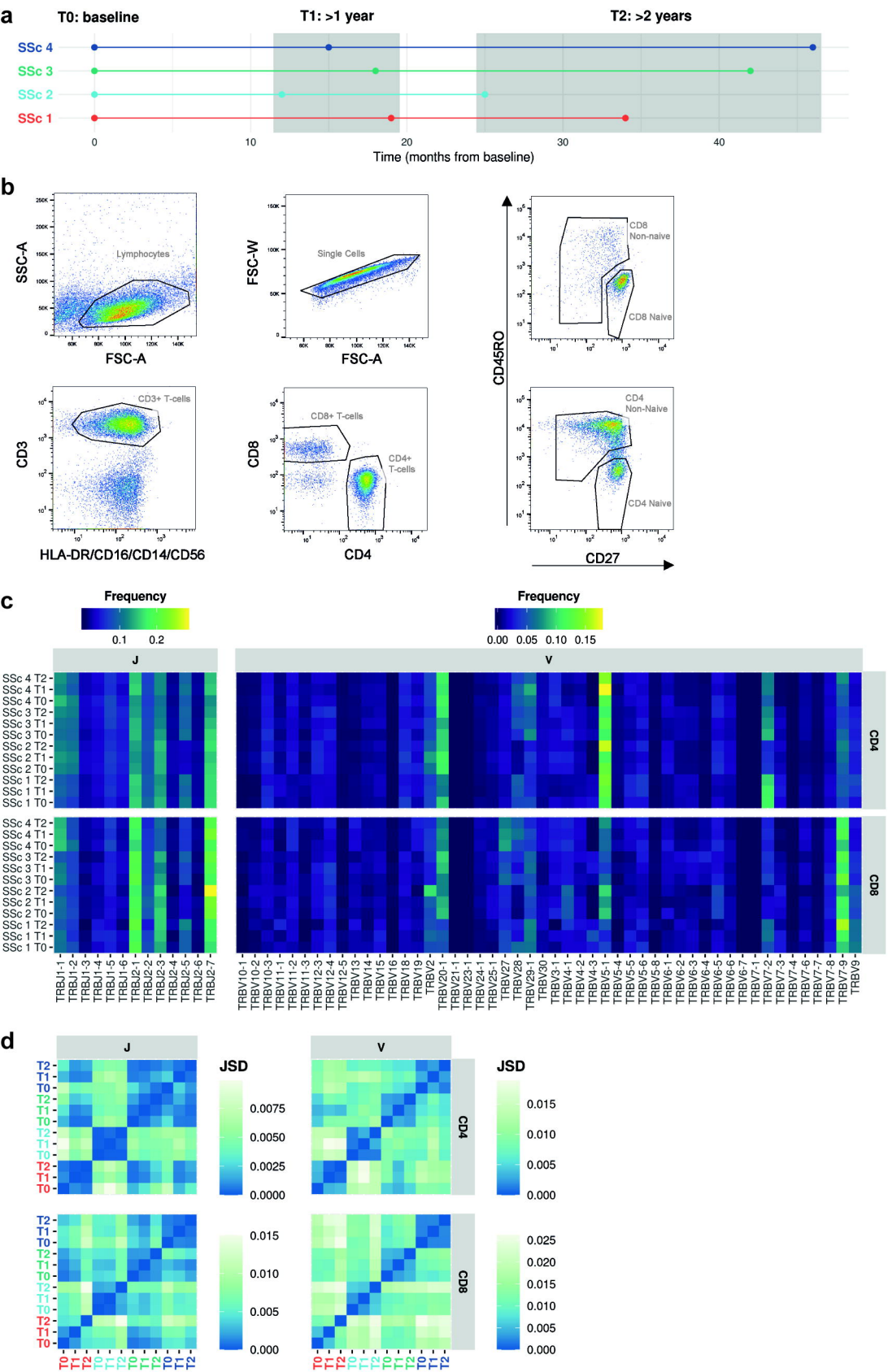
CD4 - MHC II					
	CDR3	MHC A	MHC B	Epitope	Species
SSc 2	CASSSQGYEQYF	HLA-DRA*01:01	HLA-DRB1*04:01	HA	InfluenzaA
SSc 2	CASSPGTGLDTQYF	HLA-DRA*01:01	HLA-DRB1*04:01	HA	InfluenzaA
SSc 2	CASSPRTSGPYEQYF	HLA-DRA*01:01	HLA-DRB5*01:01	p24	HIV-1
SSc 2	CASSPRTSGGSDTQYF	HLA-DRA*01:01	HLA-DRB5*01:01	p24	HIV-1
CD8 - MHC I					
	CDR3	MHC A	MHC B	Epitope	Species
SSc 1	CASSWGQGSNYGYTF	HLA-A*11:01	B2M	EBNA3B	EBV
SSc 2	CASSLGQAYEQYF	HLA-B*08:01:29	B2M	EBNA3	EBV
SSc 2	CASSLGQAYEQYF	HLA-B*44:05:01	B2M	MLANA	Homo Sapiens
SSc 2	CASSLGQAYEQYF	HLA-B*44:05:01	B2M	ABCD3	Homo Sapiens
SSc 2	CASSLGQAYEQYF	HLA-A*02:01	B2M	BMLF1	EBV
SSc 2	CASSPGTGEGYEQYF	HLA-B*08:01	B2M	BZLF1	EBV
SSc 2	CASSRLAGGTDQYF	HLA-B*08	B2M	IE1	CMV
SSc 2	CASSRLAGGTDQYF	HLA-B*08	B2M	p24	HIV-1
SSc 3	CASSFQGNYEQYF	HLA-B*08	B2M	p24	HIV-1
SSc 3	CASSPTDTQYF	HLA-A*02	B2M	p65	CMV
SSc 3	CSARDRTGNGYTF	HLA-A*02:01	B2M	BMLF1	EBV
SSc 4	CASRPLGGSYNEQFF	HLA-A*02	B2M	p65	CMV
SSc 4	CASSLAGVSYEQYF	HLA-A*02	B2M	p65	CMV
SSc 4	CASSLDLGSTEAFF	HLA-B*35:01	B2M	p65	CMV
SSc 4	CASSLDLYEQYF	HLA-A*02:01	B2M	TKT	Homo Sapiens
SSc 4	CASSLGTEAFF	HLA-A*02	B2M	p65	CMV
SSc 4	CASSLSGELFF	HLA-A*02	B2M	p65	CMV
SSc 4	CASSPGQGYEQYF	HLA-A*02	B2M	p65	CMV
SSc 4	CASSPSGVSYNEQFF	HLA-A*02	B2M	p65	CMV

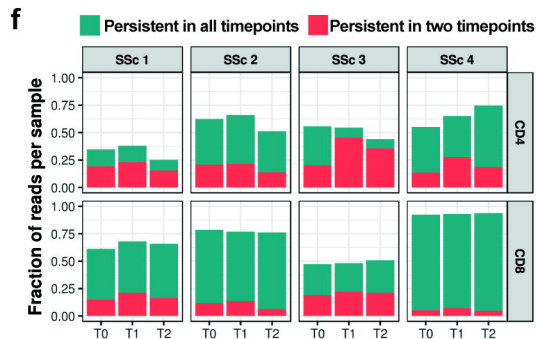
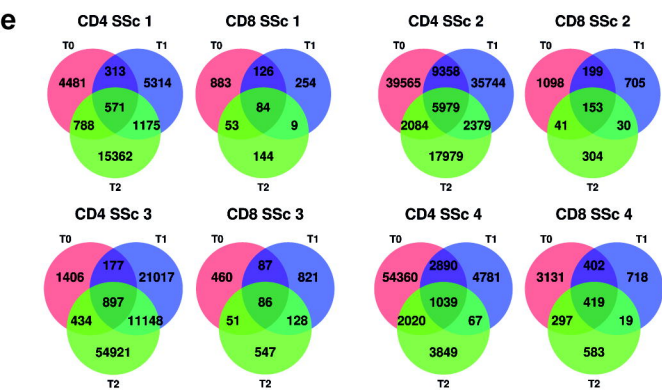
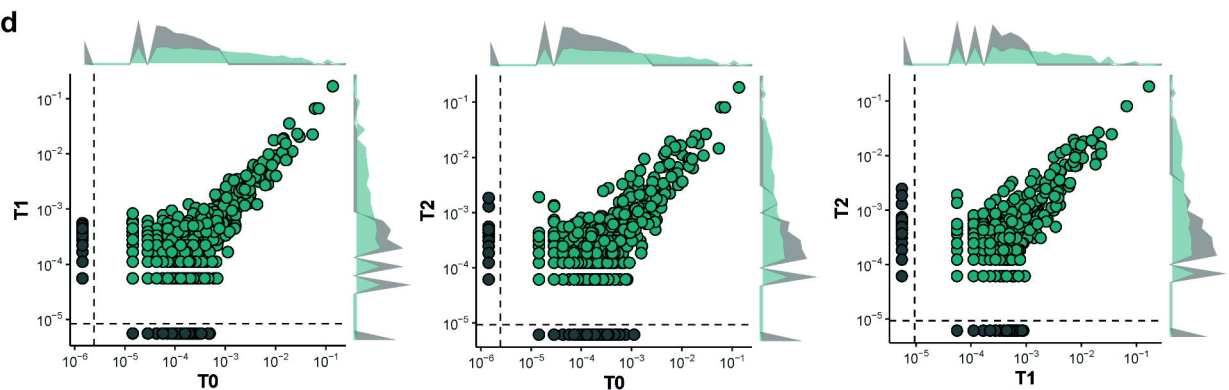
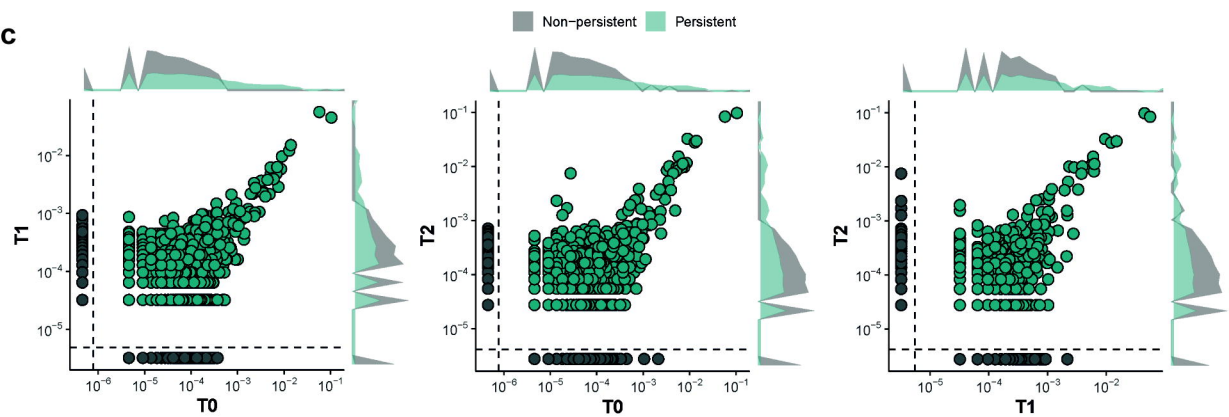
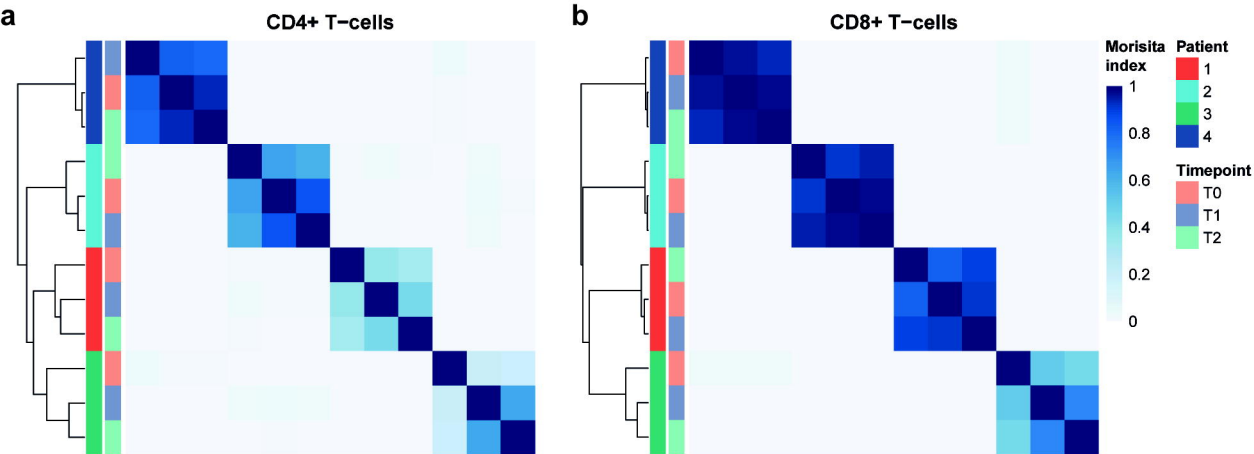
SSc 4	CASSYRENTGELFF	HLA-A*02	B2M	p65	CMV
SSc 4	CASSVGQAYEQYF	HLA-B*08:01	B2M	EBNA3A	EBV

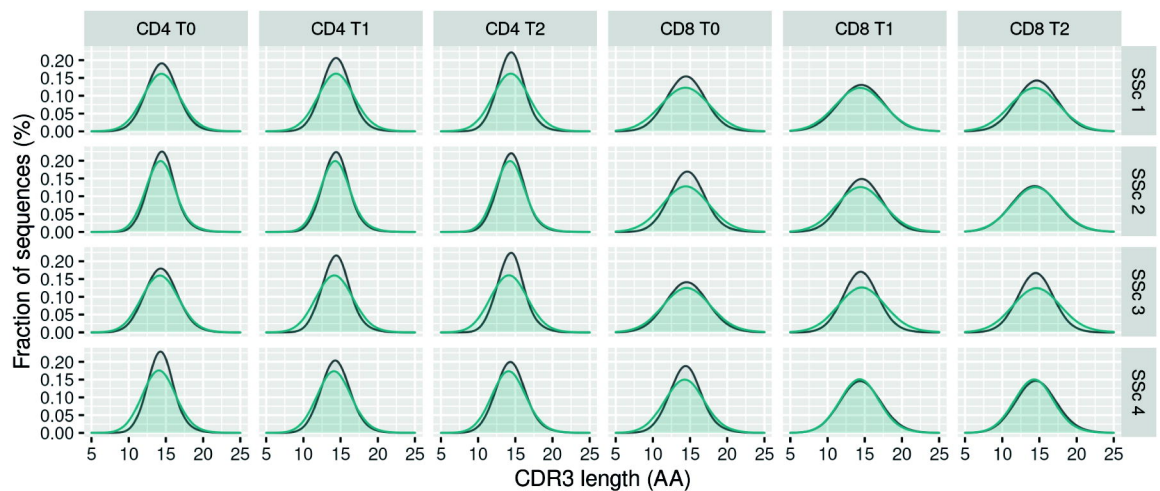
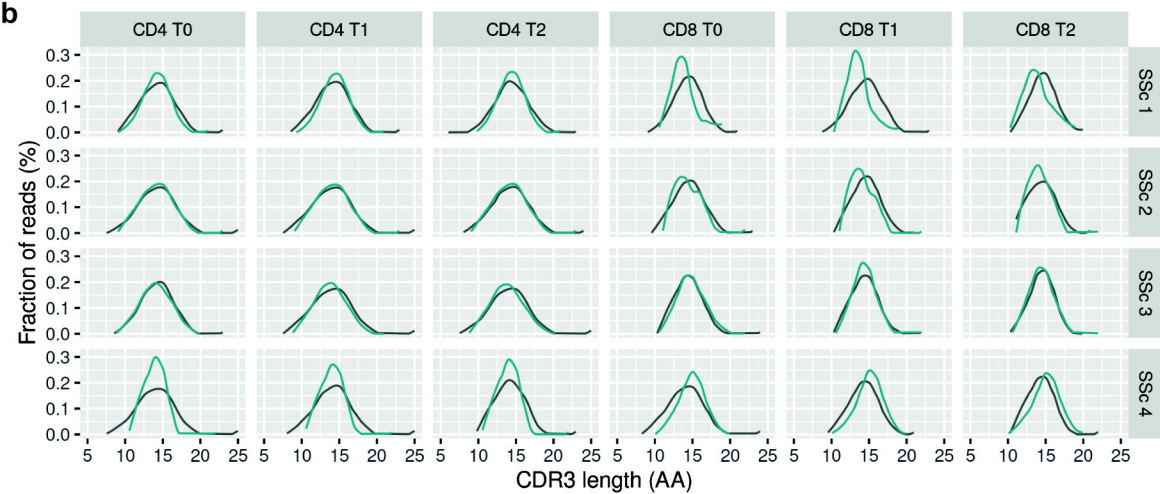
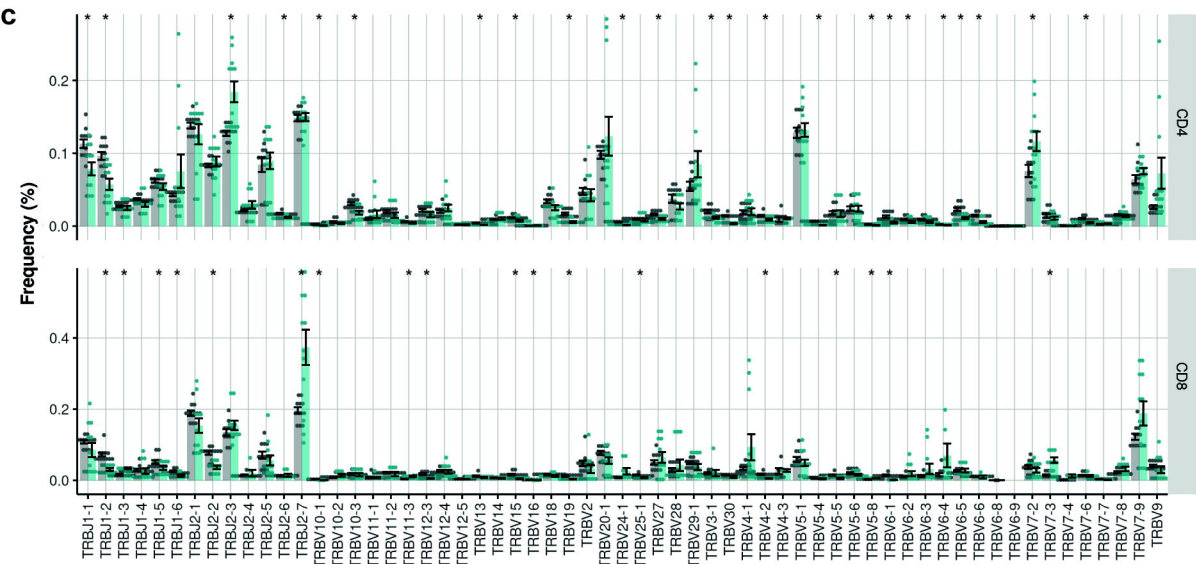
Table 2. Epitope specificity of persistent TCRβs (occurring in all three time points within a patient), according to VDJdb. For CD4+ TCRβ sequences, specificities for peptides presented on MHC II are given and for CD8+ TCRβ sequences, specificities for peptides presented on MHC I are given.

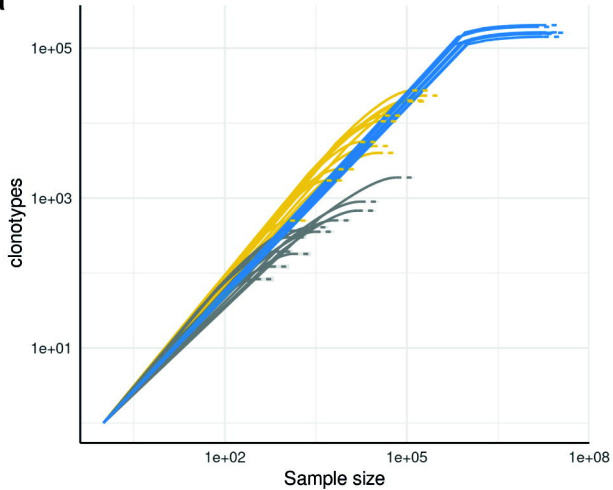
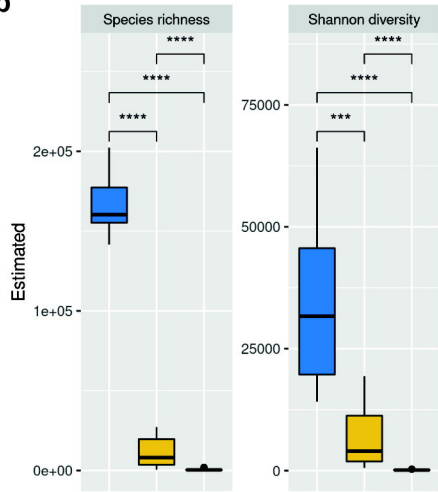
Cell type	Patient	Total clusters	Significant clusters
CD4	SSc 1	7800	204
	SSc 2	50365	3735
	SSc 3	34882	554
	SSc 4	22674	550
CD8	SSc 1	335	2
	SSc 2	534	23
	SSc 3	453	22
	SSc 4	1500	106

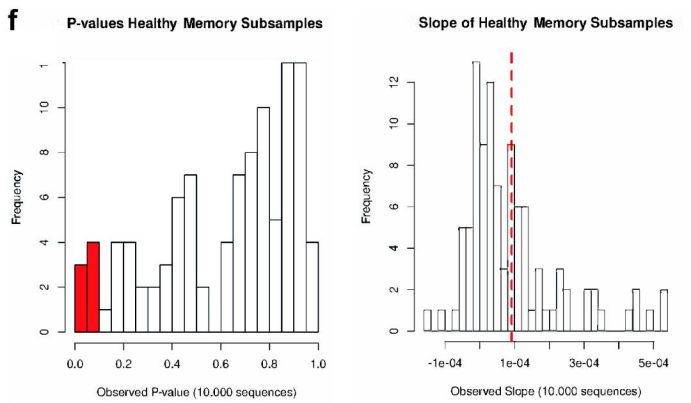
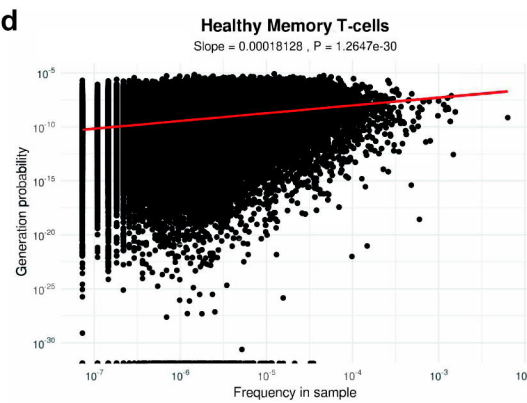
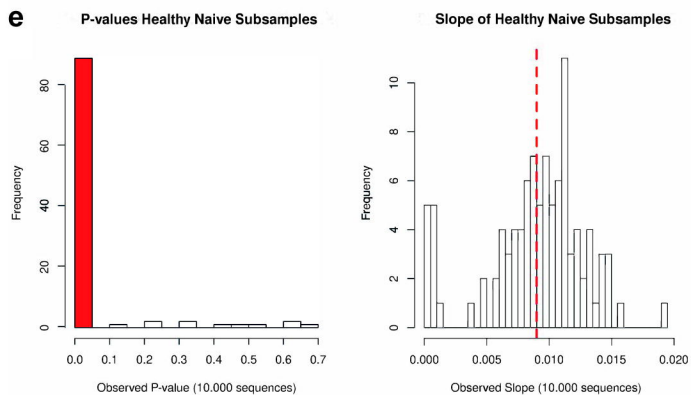
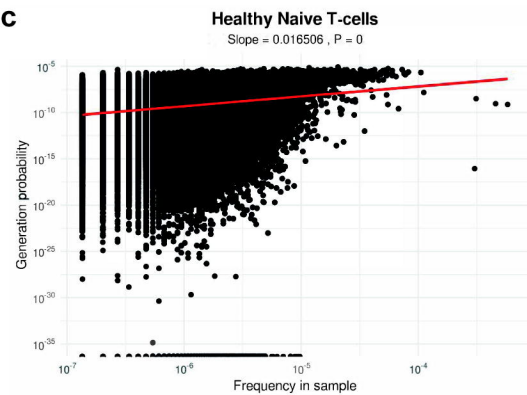
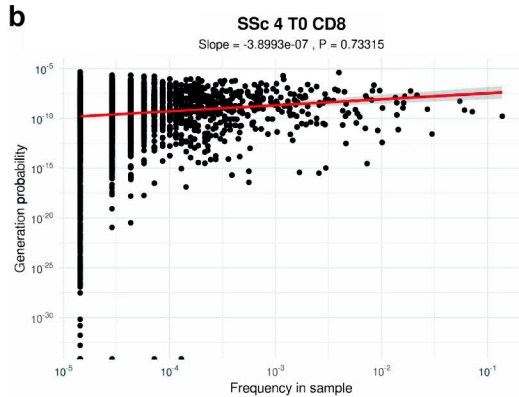
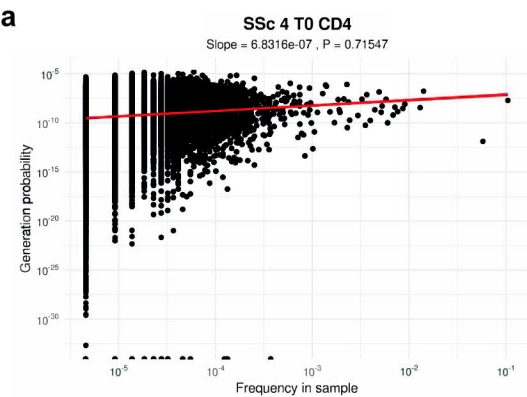
Table 3. Number of clusters identified by GLIPH2 within SSc patients. All sequences from all time points from each patient (for the two cell types) were used as input for GLIPH2. The total clusters column represents the total number of clusters identified by GLIPH2. The significant clusters column represents the number of significant clusters (consisting of sequences from all three time points, at least three unique CDR3s, length score <0.05, and Vβ score <0.05).

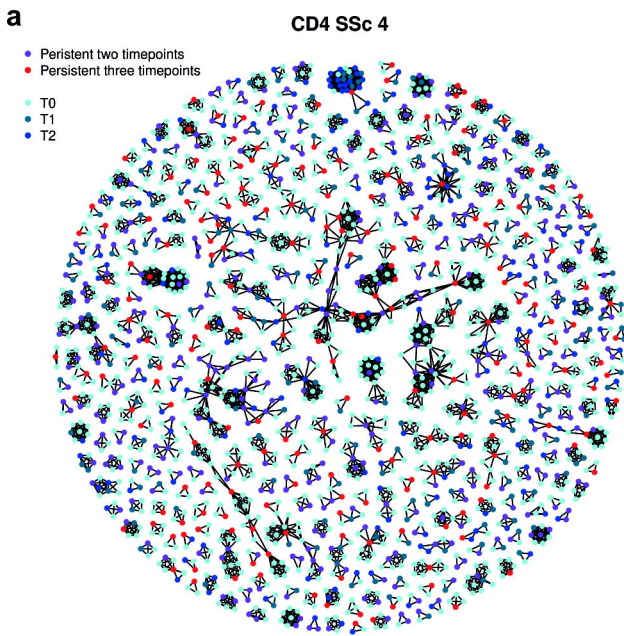




a
 Non-persistent
 Persistent
**b****c**

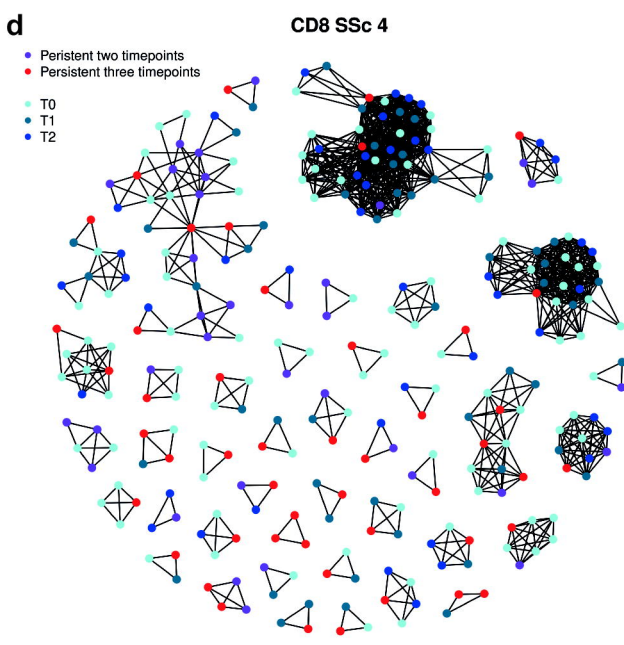
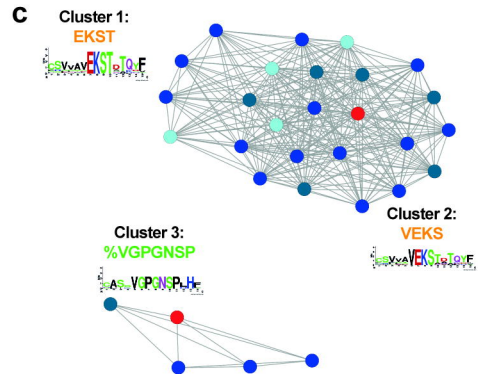
a**b**





b

Motif	Unique CDR3s	Final score	Type
EKST	21	1.90E-13	local
VEKS	18	1.40E-12	local
%VGPGNSP	5	1.50E-12	global
SVGPGNS%	4	2.30E-12	global
S%SGDQP	4	2.30E-12	global
SFGLA%ET	4	2.30E-12	global
%LRDT	4	2.30E-12	global
SYS%DQP	4	4.70E-12	global
RDAPRG%T	3	6.20E-12	global
VAVE%STD	3	6.20E-12	global



e

Motif	Unique CDR3s	Final score	Type
GTSP	13	7.70E-14	local
LEY	23	8.50E-14	local
LEYG	18	8.60E-14	local
EYGT	16	8.60E-14	local
SLEY	16	8.60E-14	local
TSPY	12	1.50E-13	local
PPVD	11	2.40E-13	local
VDTD	10	3.10E-13	local
PVD	18	3.40E-13	local
PVDT	12	7.70E-13	local

