OPEN ACCESS

EXTENDED REPORT

A circulating reservoir of pathogenic-like CD4+ T cells shares a genetic and phenotypic signature with the inflamed synovial micro-environment

Roberto Spreafico,[1,2] Maura Rossetti,[1,2] Jorg van Loosdregt,[1] Carol A Wallace,[3] Margherita Massa,[4] Silvia Magni-Manzoni,[5] Marco Gattorno,[6] Alberto Martini,[6] Daniel J Lovell,[7,8] Salvatore Albani[1,2]

Handling editor Tore K Kvien

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/annrheumdis-2014-206226).

For numbered affiliations see end of article.

Correspondence to
Dr Roberto Spreafico, SingHealth Translational Immunology and Inflammation Centre, The Academia, 20 College Road, Discovery Tower Level 8, Singapore 169856, Singapore; roberto.spreafico@gmail.com

RS and MR contributed equally.

Received 7 July 2014
Revised 18 November 2014
Accepted 18 November 2014
Published Online First 12 December 2014

To cite: Spreafico R, Rossetti M, van Loosdregt J, *et al*. *Ann Rheum Dis* 2016;**75**:459–465.

ABSTRACT

Objectives Systemic immunological processes are profoundly shaped by the micro-environments where antigen recognition occurs. Identifying molecular signatures distinctive of such processes is pivotal to understand pathogenic immune responses and manipulate them for therapeutic purposes. Unfortunately, direct investigation of peripheral tissues, enriched in pathogenic T cells, is often impossible or imposingly invasive in humans. Conversely, blood is easily accessible, but pathogenic signatures are diluted systemically as a result of the strict compartmentalisation of immune responses. In this work, we aimed at defining immune mediators shared between the bloodstream and the synovial micro-environment, and relevant for disease activity in autoimmune arthritis.

Methods CD4+ T cells from blood and synovium of patients with juvenile idiopathic arthritis (JIA) were immunophenotyped by flow cytometry. The TCR repertoire of a circulating subset showing similarity with the synovium was analysed through next-generation sequencing of TCR β-chain CDR3 to confirm enrichment in synovial clonotypes. Finally, clinical relevance was established by monitoring the size of this subset in the blood of patients with JIA and rheumatoid arthritis (RA).

Results We identified a small subset of circulating CD4+ T cells replicating the phenotypical signature of lymphocytes infiltrating the inflamed synovium. These circulating pathogenic-like lymphocytes (CPLs) were enriched in synovial clonotypes and they exhibited strong production of pro-inflammatory cytokines. Importantly, CPLs were expanded in patients with JIA, who did not respond to therapy, and also correlated with disease activity in patients with RA.

Conclusions CPLs provide an accessible reservoir of pathogenic cells recirculating into the bloodstream and correlating with disease activity, to be exploited for diagnostic and research purposes.

INTRODUCTION

The blood is the most easily accessible human tissue, and as such, it has been heavily investigated for pathogenic immune signatures in a variety of T cell-mediated diseases, including juvenile and adult forms of autoimmune arthritis. Notwithstanding its value in blood-borne diseases, such as HIV, investigations in the blood have not been equally fruitful for autoimmune rheumatological diseases. This is likely due to the strict compartmentalisation of tissue-restricted immune responses, which are shaped by the peripheral micro-environment and become highly diluted in the bloodstream.[1–3] A good example of the immunological disconnect between the blood and the inflamed tissues is provided by past attempts at identifying pathogenic signatures in autoimmune arthritis by characterising the Vβ usage in synovial fluid versus blood.[4] These studies were inconclusive because they were hindered by the limitations of the technology available at the time and by the inherent difficulties in matching the synovial repertoire with the much more diverse clonal representation of blood T cells.

To get insights in disease pathogenesis, the scientific community is moving away from immunological studies in the blood to embrace tissue immunology. Unfortunately, normal and inflamed human tissues are accessible only through invasive procedures.[5] In addition, current therapies for autoimmune arthritis are often able to transiently reduce accumulation of synovial fluids at the affected joints, thereby reducing synovial sample availability and making analysis of the micro-environmental processes even more difficult.[6 7]

In this work, we first sought phenotypical patterns defining synovial T cells of patients affected by juvenile idiopathic arthritis (JIA) and adult rheumatoid arthritis (RA), and used this signature to select synovial-like T cells in the circulation. Then, we exploited next-generation sequencing to identify T cells based on their unique TCR, and used it as a barcode to demonstrate the identity between synovial T cells and their circulating counterparts. Finally, we proved that these circulating synovial-like T cells are associated with disease activity and responsiveness to therapy in human arthritis, expanding the clinical relevance of our discovery. Our work sheds new light on the relationship between micro-environmental and systemic immunity and might greatly facilitate the development of improved and targeted therapeutic approaches.

METHODS

TCR repertoire

TCRβ CDR3 sequencing was performed by Adaptive Biotechnologies.[8 9] Sorting was performed immediately after thawing. Dead cells were

excluded using Live/Dead Fixable Near-IR Stain. Circulating pathogenic-like lymphocytes (CPLs) were sorted as human leucocyte antigen (HLA)-DR$^+$CD14$^-$CD3$^+$CD4$^+$CD25$^{low/-}$ cells from total peripheral blood mononucleated cells.

The analysis of TCR species composition is conceptually identical to the investigation of biodiversity. Since ecological methods are well developed, they have been properly adapted to TCR repertoire analyses.[8–13] Each sample is interpreted as a distribution of individuals (T cell genomes) belonging to different species (TCR clonotypes), and can be interrogated for (i) its intrinsic diversity and (ii) its dissimilarity (ie, pairwise distance) from another sample. Diversity is determined by species richness (ie, the number of unique species relative to the population size) and distribution evenness: the higher the richness and the more evenly distributed the species, the higher the diversity. Likewise, the dissimilarity between samples is determined by the number of shared species, and their frequencies. We performed both types of analyses on productive nucleotide sequences (see online supplementary figure S4) and amino acid sequences derived by in silico translation (figure 3).

To properly address certain biological questions, cell subsets of varying size may be targeted for sampling.[12] In our study, CPLs represent only a small fraction of blood CD4$^+$ T cells; consequently, the number of T cell genomes obtained from CPL and CD4$^+$ blood samples will differ of at least an order of magnitude. Taking into account that sample size is critical, because the higher the number of individuals sampled, the higher the probability of detecting rare species. In other words, the larger the sample size, the larger the diversity.[12] This issue becomes apparent when two samples of different size are taken from the same population, the expectation being that they should display identical diversity, and no pairwise dissimilarity, except for the contribution of random sampling. By contrast, the larger sample will always show higher diversity than the smaller, and the two samples will appear more dissimilar than if they had the same size. This is because the larger samples will always detect rare species not comprised in the smaller sample, as a result of the imbalance in the number of examined individuals.[12] Therefore, to ensure fairness of comparisons, it is necessary to contrast samples of equal size. To this aim, the smaller sample could be upsized to the size of the larger; alternatively, the larger could be downsized to the size of the smaller. The former strategy requires diversity and dissimilarity be extrapolated as a function of size, and is not used in practice. Conversely, the latter strategy is the standard in the field. To make assessments robust, sub-sampling is performed repeatedly, and the median of the readout of interest across all sub-samples is reported.[10 12 13]

For analyses of diversity, the Renyi entropy index was used.[10 11] The larger the Renyi index, the larger the diversity. The Renyi index is modulated by the α parameter: a value of α above 1 puts more weight on abundant species, while a value below 1 puts more weight on rare species. Because, by definition, sampling of an entire population cannot be achieved, some TCR species will never be observed in the sample. Such under-sampling mostly affects rare TCRs. Therefore, the Renyi index must be evaluated across different values of α.[10 11] Differential species diversity between two samples is substantiated only if the Renyi index is consistently greater (or lower) across all values of α. Equal sample size was achieved by repeated (200 times) sub-sampling, and the median of the Renyi index across sub-samples at each value of α was then plotted.[10 12 13]

For analyses of dissimilarity, two independent methods were applied. Using the first method, the pairwise distance between samples was calculated using the Chao-modified Jaccard index, which corrects for the effect of under-sampling.[10 13] Using the second method, equal sample size was first achieved by repeated random sub-sampling.[10 12 13] For each sample, we selected 200 sampling sizes at equally spaced intervals, corresponding to 200 increasing T cell genome counts, from a T cell genome count of 1 to the T cell genome count of the whole sample. At each sampling size, the selected number of individuals was drawn 200 times without replacement from the whole sample, and the fraction of unique synovial clonotypes shared with their paired blood samples, weighted for their frequencies, was calculated for each draw. The median percentage of overlap for each sampling size was then plotted, and the percentage of overlap was calculated at the size of the smaller sample to avoid extrapolation.[10 12 13]

The R code used for the analyses of the TCR repertoire is available as online supplementary information.

RESULTS

Circulating CD4$^+$ T cells from patients with JIA with active disease do not reflect the synovial pro-inflammatory signature

Our work was first kindled by the observation that circulating lymphocytes do not mirror their counterparts from the inflamed synovial micro-environment, which are enriched in pathogenic clonotypes.[14–21] Figure 1A–D illustrates important discrepancies between blood and synovial T cells of patients with active JIA. Using flow cytometry, we inspected T cells for key processes involved in pathogenic T cell-mediated responses, that is, activation (CD69, CD25, HLA-DR), antigen recognition (Ki67, CD3), differentiation (CD45RA), migratory behaviour (CCR5, CCR6, CCR7) and maturation/exhaustion (CTLA-4, PD-1, LAG-3). Synovial T cells mainly comprised activated CD69$^+$CD45RA$^-$ memory T cells, and were enriched in CD25$^+$ and HLA-DR$^+$ cells, while circulating T cells showed a mixed naive/memory resting phenotype (figure 1A, B). Consistent with recent antigen recognition, a large fraction of synovial T cells expressed Ki67 (an evidence of recent in vivo proliferation, thus a surrogate for antigen encounter) and down-regulated CD3[22] (figure 1B). Moreover, they preferentially expressed the pro-inflammatory chemokine receptors CCR5 and CCR6 over the lymphoid organ-homing CCR7 (figure 1C). Finally, synovial T cells exhibited signs of maturation/exhaustion, as revealed by the expression of CTLA-4, LAG-3 and PD-1, consistent with sustained antigen exposure (figure 1D).

Altogether, these molecules compose a signature dramatically differentiating synovial and circulating T cells. Indeed, an unsupervised hierarchical clustering algorithm fed with this phenotypical data perfectly segregated the samples according to their tissue of origin (figure 1E).

From these data, it is evident that a much finer dissection is needed to distill complex pathogenic signatures from systemic compartments. However, the specific synovial T cell profile we identified might be used as a guide to identify pathogenic-like cells outside the inflamed synovium.

Circulating pathogenic-like lymphocytes (CPLs) from patients with active disease mirror the inflammatory synovial T cell signature

We hypothesised that if any fraction of pathogenic T cells recirculates out of the inflamed tissue, such population would have a phenotypical and functional profile similar to its synovial counterpart, enriched in arthritogenic T cells.[14–21] Interestingly, a very small subset of blood T cells still expressed markers highly upregulated on synovial T cells (figure 1), indicating that the

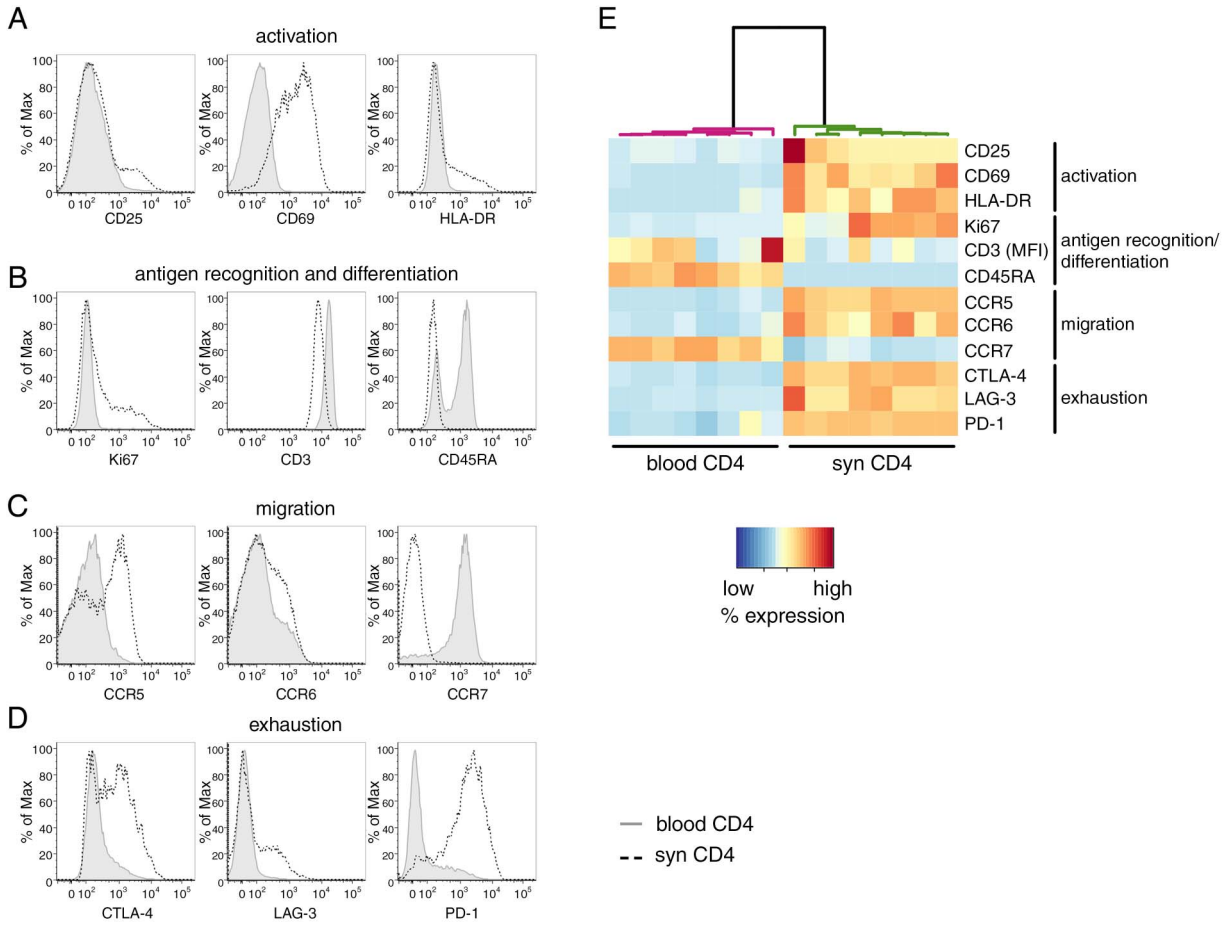


Figure 1 Circulating CD4$^+$ T cells from patients with active juvenile idiopathic arthritis (JIA) do not reflect the synovial pro-inflammatory signature. (A–D) Representative staining of synovial (black dotted lines) and peripheral blood (solid grey lines) samples of patients with active JIA, segregated by function. T cells were gated as CD14$^-$CD3$^+$CD4$^+$CD25$^{low/-}$FOXP3$^-$. n=8. (E). Phenotypical results (MFI, median fluorescence intensity, for CD3, % positive for all other markers) were fed to an unsupervised hierarchical clustering algorithm and colour-coded upon row-wise normalisation. Each column represents an individual sample.

same signature is detectable also in the blood, only much more diluted than in the synovium. In the blood, we observed high pairwise correlations between most of the markers (figure 2A, B). Some notable exceptions were CD69 and LAG-3, whose expression did not correlate with that of any other molecule, and CD25, which correlated with some but not all. By contrast, another activation marker, HLA-DR, was one of the most correlated (figure 2A, B). These findings are in line with the differential expression kinetics of activation markers; indeed, its late upregulation when compared with CD69[23 24] enables HLA-DR to still remain detectable when activated T cells reach the blood. These data indicate that different T cell activation markers have unique features; as such, they cannot be interchangeably used as proxies for peripheral inflammation.

An important practical consequence of the high intercorrelations between most markers composing the signature of synovial-like T cells is that any of these proteins might be used to pinpoint this population. This provides the opportunity to isolate it with a single marker, thus reducing technical complexity in view of the potential clinical application of our approach. Among the molecules tested, HLA-DR was particularly appealing as a sorting marker, because (1) it is a surface protein, thereby allowing the isolation of living cells and downstream functional analyses, in contrast to other relevant intracellular antigens that require fixation, such as Ki67 and CTLA-4; (2) differently from other proteins, including PD-1 and chemokine receptors, it is seemingly devoid of signalling activity in T cells, enabling the sorting of unperturbed T cells. Several studies reported conflicting results on HLA-DR signalling capabilities.[25–32] In our hands, cross-linking of HLA-DR in T cell lines did not elicit calcium flux (see online supplementary figure S3A), a key signalling event downstream of HLA-DR triggering in dendritic cells.[32] In addition, HLA-DR$^+$ T cell lines did not induce proliferation of allogeneic CD4$^+$ T cells (see online supplementary figure S3B). Altogether, these data demonstrate that HLA-DR does not exhibit any back-signalling activity and does not endow T cells with antigen-presenting capabilities. Therefore, HLA-DR can be used to sort viable and unperturbed T cells.

We sought to confirm that HLA-DR could indeed be used to identify circulating T cells similar to synovial T cells. Differently from the bulk of blood CD4$^+$ T cells, HLA-DR-gated T cells were enriched in activated, memory-like, antigen-experienced Ki67$^+$ cells; in addition, they might have been exposed to autoantigens in affected joints, as they preferentially expressed CCR5 and CCR6 over CCR7, and they also expressed high levels of maturation/exhaustion markers (figure 2C). Importantly, when these phenotypical data were analysed through unsupervised clustering, the algorithm partitioned HLA-DR-gated cells apart from blood T cells, but together with synovial T cells (figure 2C).

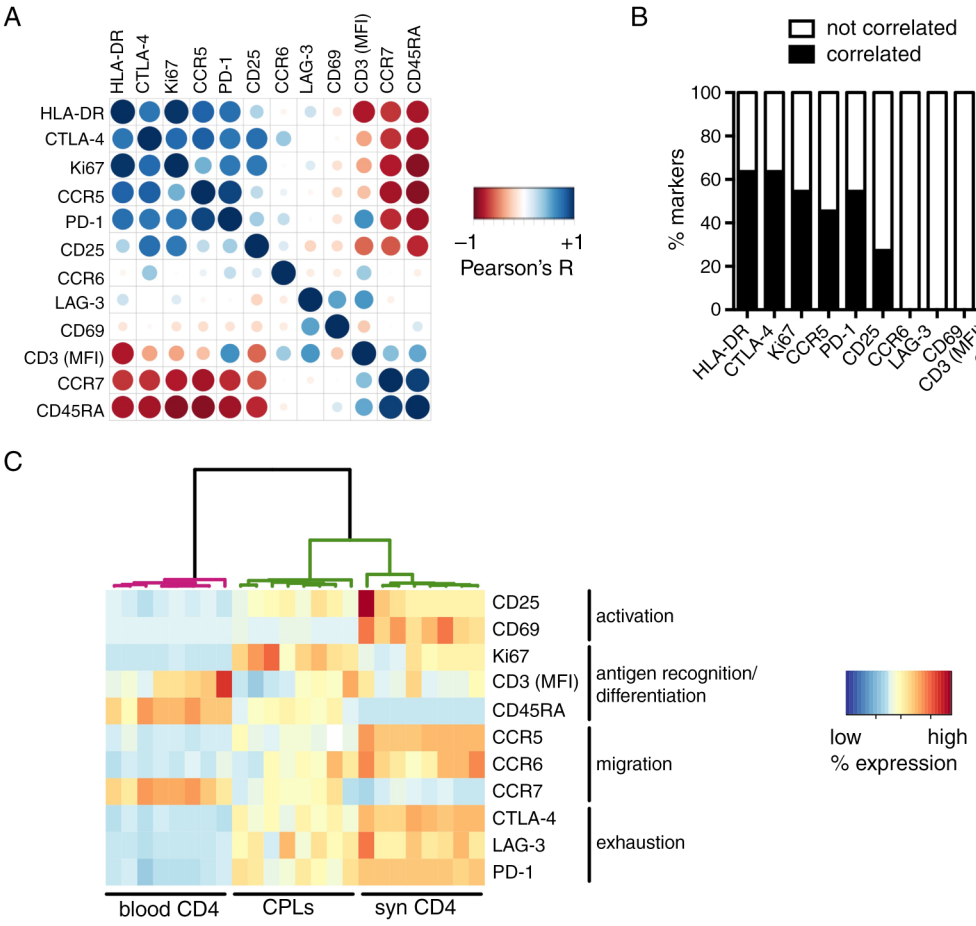


Figure 2 Circulating pathogenic-like lymphocytes (CPLs) from patients with active disease mirror the inflammatory synovial T cell signature. (A) Pairwise correlations (Pearson's R) were calculated for the indicated markers (median fluorescence intensity (MFI) for CD3, % positive for all other markers) measured on circulating $CD4^+$ T cells of patients with active juvenile idiopathic arthritis (JIA), and then colour-coded as indicated in the legend and re-ordered according to the aggregate level of correlation (first principal component). Larger circles indicate higher |R|. n=32. (B) For each marker from panel A, the percentage of pairwise comparisons with all other markers from the panel ensuing in correlation (|R| ≥0.7), or no correlation (|R|<0.7), is reported. (C) Phenotypical results (MFI for CD3, % positive for all other markers) within gated CPLs and the bulk of blood or synovial $CD4^+$ T cells from patients with JIA with active disease were fed to a hierarchical clustering algorithm and colour-coded upon row-wise normalisation. A white cell indicates data not available. n=8 per group.

Altogether, our data demonstrate that HLA-DR can be used to identify a small population of 'circulating pathogenic-like lymphocytes' (CPLs) expressing the typical synovial T cell signature. CPLs comprise activated, antigen-experienced T cells able to recirculate through inflamed sites.

CPLs are enriched in clonotypes of synovial T cells

To get a compelling proof of the parallel between CPLs and synovial T cells, we compared their TCR repertoires through next-generation sequencing of TCRβ CDR3, the main determinant of TCR specificity. This approach enabled us to verify T cell identity at clonal level (ie, at the level of a single CDR3 sequence). We were able to sequence an average number of 71 500 blood T cells, 5000 CPLs and 15 000 synovial T cells.

First, we evaluated the TCR repertoire diversity by calculating the Renyi index.[10 11] In accordance with previous reports of synovial T cell oligoclonality,[33] the TCR diversity in the synovium was lower than in the blood. Importantly, also CPLs were less diverse compared with the bulk of blood T cells—and in some instances as oligoclonal or even more oligoclonal than the synovium—at the level of both amino acid and nucleotide sequences (figure 3A, B and see online supplementary figure S4A,B, respectively). These results demonstrate that CPLs display reduced TCR diversity, similarly to synovial T cells.[33]

Next, we investigated how close the TCR repertoire of CPLs and synovial T cells are at clonal level. Upon clustering based on the dissimilarity of distribution of the TCR sequences,[13] CPLs segregated together with synovial T cells and apart from blood T cells (figure 3C, D and see online supplementary figure S4C, D). Accordingly, CPLs shared a substantially higher fraction of CDR3 sequences with synovial TCRs than blood T cells (figure 3E, F and see online supplementary figure S4E, F, respectively). Importantly, the conspicuous difference observed between CPLs and the total peripheral blood T cells was completely overlooked in a traditional analysis of Vβ family usage (see online supplementary figure S4G), which lacks the resolution necessary to finely dissect the clonal composition of the TCR repertoire.

In summary, these data demonstrate that CPLs comprise a large fraction of synovial clonotypes.

CPLs retain their pro-inflammatory potential

As CPLs are highly enriched in T cells recirculating through the inflamed synovial micro-environment, which is, in turn, enriched in arthritogenic T cells,[14–21] we investigated whether they would retain their pro-inflammatory characteristics while

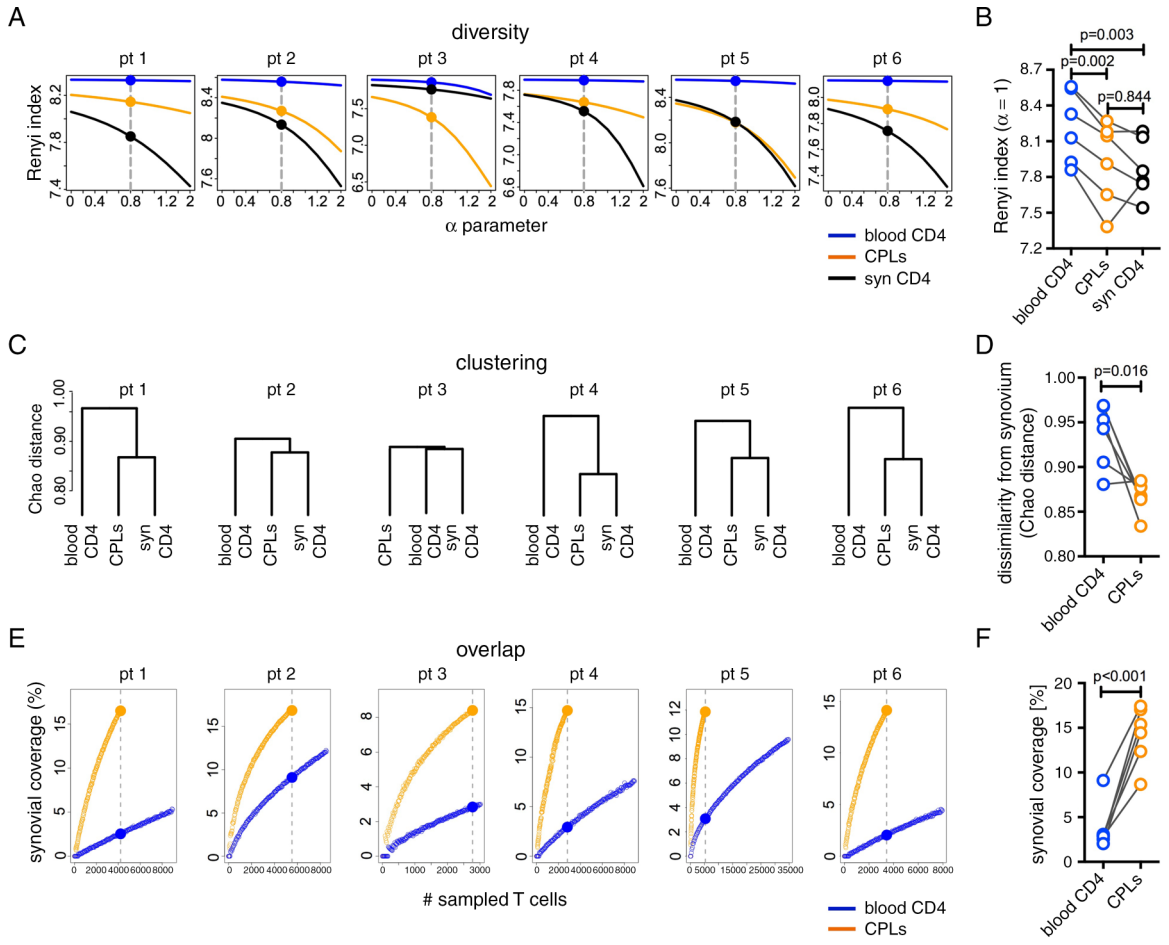


Figure 3 Circulating pathogenic-like lymphocytes (CPLs) are enriched in clonotypes of synovial T cells. Next-generation sequencing of TCRβ CDR3 sequences was performed on blood or synovial CD4$^+$ T cells of patients with active juvenile idiopathic arthritis (JIA). All data refer to amino acid sequences derived by in silico translation. (A and B) Renyi diversity indices for different values of α (A) and summary at α=1 (B; grey dashed line in panel A). Differential species diversity is substantiated only if the Renyi index is consistently greater (or lower) across all values of the α parameter. (C) Clustering based on TCR repertoire distances, calculated as 1—Chao-modified Jaccard index (Chao distance). (D) Summary of the TCR repertoire pairwise Chao distances between CPL or blood samples and their paired synovial sample. (E) Overlap of the TCR repertoire from CPL and blood samples with their matched synovial sample. Each open circle is the median overlap of 200 subsamples of the indicated size from either CPL or blood samples with 200 subsamples of the same size from their paired synovial sample. (F) Summary statistics for the overlap data, as measured at the size of the smaller sample (grey dashed line in panel E). In A, C and E, each panel represents an individual patient.

recirculating in peripheral blood. Importantly, CPLs produced much higher levels of the pro-inflammatory cytokines interferon-γ, interleukin-17 and tumour necrosis factor-α than the bulk of blood CD4$^+$ T cells (figure 4), demonstrating that they are able to actively fuel inflammation.

CPLs correlate with unresponsiveness to therapy and disease activity in both juvenile and adult autoimmune arthritis

Our discovery of a circulating pool of pathogenic-like T cells might have profound clinical relevance. Indeed, it might provide the missing link between the clinical manifestations of the disease in the periphery and the seemingly normal systemic immunity. Under this framework, the frequency of CPLs in the periphery should correlate with disease activity. We tested this hypothesis on peripheral blood samples from patients with JIA, collected before (T0) and after (Tend) therapy,[34] and stratified for responsiveness to therapy based on whether they reached inactive disease (ID)[35] or not (NO ID) at Tend. CPLs slightly declined in prospective ID patients, but substantially increased in NO ID patients, resulting in NO ID patients harbouring a

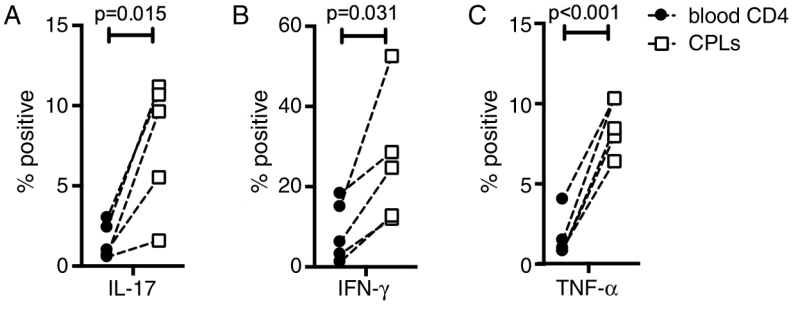


Figure 4 CPLs retain their pro-inflammatory potential. Cytokine production from circulating CD4$^+$ T cells of patients with active JIA upon PMA/ionomycin stimulation. Each line corresponds to an individual patient. CPLs, circulating pathogenic-like lymphocytes; JIA, juvenile idiopathic arthritis; IFN, interferon; IL, interleukin; PMA, phorbol 12-myristate 13-acetate; TNF, tumour necrosis factor.

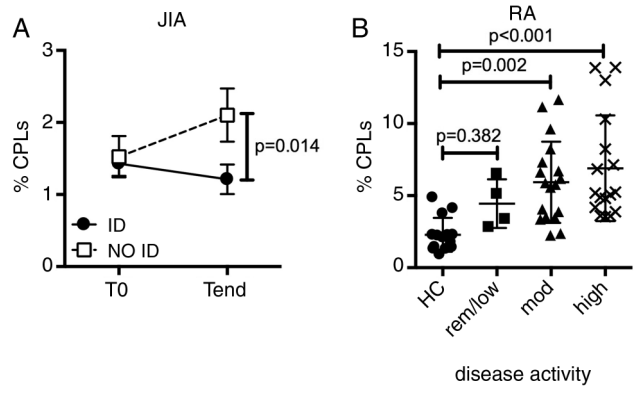


Figure 5 Circulating pathogenic-like lymphocytes (CPLs) correlate with unresponsiveness to therapy and disease activity in both juvenile and adult autoimmune arthritis. (A) Percentage of CPLs over time in blood CD4$^+$ T cells from patients with juvenile idiopathic arthritis (JIA). All patients were in active disease at T0 and then segregated based on their clinical activity at Tend. ID=inactive disease; NO ID: not inactive disease. Vertical lines represent SEM. n=10–15 per group, per time point. (B) CPL representation in blood CD4$^+$ T cells from patients with rheumatoid arthritis (RA), segregated by disease activity score (DAS-28 3) and compared with healthy controls (HC). low/rem: low/remission (DAS≤3.2); mod: moderate (3.2<DAS≤5.1); high: DAS>5.1. Each dot represents an individual patient.

doubled frequency of CPLs at Tend when compared with ID patients (figure 5A). Thus, the increase in CPL frequency mirrors the persistent or intensified synovial inflammation clinically observed in NO ID patients.

We then asked whether this finding, unequivocal in JIA, might be extended to other autoimmune diseases. Therefore, we examined adult RA[36] for the presence of CPLs. Importantly, CPLs were expanded in patients with RA when compared with healthy controls, and substantially increased with worsening activity score (figure 5B). This observation confers a broader clinical significance to CPLs beyond paediatric autoimmunity.

DISCUSSION

In this work, we identified a small population of pathogenic-like T cells that is (1) reproducing the synovial T cell signature in the circulation; (2) able to recirculate through the site of autoimmune reaction; (3) highly enriched in synovial clonotypes; and (4) correlating with disease activity in both juvenile and adult autoimmune arthritis. Our discovery exposes an easily accessible pool of pathogenic-like cells, which may be used to substantially advance our understanding of aberrant self-recognition, as well as to design targeted diagnostic and therapeutic tools.

Our claims are supported by the functional profile of CPLs, and then formally substantiated by their TCR repertoire. Indeed, CPLs recapitulate the synovial T cell signature, being enriched in proliferating Ki67$^+$ cells and preferentially expressing pro-inflammatory chemokine receptors and exhaustion markers, in the absence of overt infection. In addition, CPLs display lower levels of CD3 compared with the bulk of blood T cells, a sign of TCR-mediated activation, as opposed to cytokine-mediated bystander activation.[22] The analysis of their TCR repertoire provides direct evidence in support of this claim: indeed, CPLs are oligoclonal and share a consistent fraction of clonotypes with the synovium, where arthritogenic T cells reside.[14–21] Importantly, the overlap increases as a function of the number of sequenced T cells. Therefore, at the level of the entire CPL repertoire, the percentage of clonotypes shared with the synovium is even higher than the one we measured on a sample (5000 CPLs on average from less than 10 mL of blood). Overall, these findings indicate that CPLs include T cells identical to those residing in the inflamed micro-environment, and still retain their pro-inflammatory features while recirculating.

CPL features are consistent with the hypothesis that they escape from the site of autoimmune reaction. Alternatively, it is possible that CPLs recirculate in the blood after being primed/reactivated in the lymph nodes draining auto-reactive sites, and are thus migrating to, rather than from, the inflamed joints. There is a third scenario, which would reconcile both alternatives. Chronic autoimmunity is often accompanied by the appearance of tertiary lymphoid structures within autoreactive sites.[37] These hybrid structures show elements of both peripheral tissues and lymph nodes, and support direct antigen presentation and T cell activation. In vivo cell tracking would be needed to distinguish between these possibilities. Unfortunately, class II MHCs are only expressed by antigen-presenting cells but not by T cells in mice, due to a mutation in the class II transactivator gene.[38] This limitation underscores the need for exquisitely human studies, such as our own, to avoid overlooking important aspects of disease pathogenesis due to a disconnect between the human disease and animal models.

Although CPLs are present in paediatric ID patients and adult healthy individuals, consistent with a basal level of possibly self-reactive clonotypes even in the absence of disease,[39] they are selectively expanded in NO ID patients and patients with active RA. Thus, CPLs provide a correlate and suggest a potential interpretation for unresponsiveness to therapy and disease activity for future mechanistic studies. Indeed, the increase in CPL frequency may result from the synovial expansion of arthritogenic T cells resistant to suppression and therapy, which boost synovial inflammation, ultimately fostering the clinical manifestations of the disease. Given their ability to actively produce inflammatory cytokines, CPLs might also contribute directly to the spreading of the disease to unaffected joints and to the exacerbation of systemic symptoms.

CPLs are defined by their synovial-like T cell signature, but they can be isolated with a single marker, substantially simplifying their investigation and isolation in clinical settings. Our strategy for the identification of arthritogenic T cells bypasses the limitations of current approaches based on tetramers or restimulation, namely: (1) the need for prior knowledge of antigen specificity, a major hurdle in autoimmune diseases such as JIA, for which most auto-antigens have yet to be identified;[5 40] and (2) the ability to detect only TCR clones with high affinity for any given autoantigen.[41] In addition, the small amount of blood withdrawn from children affected by JIA imposes serious additional constraints on the number of antigens that can be tested on a single sample. These issues motivated our search for alternative and more effective strategies to identify pathogenic T cells in the circulation. We first moved from the consideration that synovial T cells are enriched in arthritogenic T cells.[14–21] Thus, we took advantage of recent developments in TCR repertoire profiling to identify synovial clonotypes in the bloodstream. Differently from traditional methods, the next-generation sequencing of TCR CDR3 regions provides the opportunity to perform investigations at clonal level, enabling us to establish the genetic overlap between CPLs and synovial T cells.

Differently from blood-borne diseases, where it is easy to envision that pathogenic T cells can be easily isolated from the blood, uncovering a small but detectable pool of circulating pathogenic-like T cells in an autoimmune disease was unexpected. Given the small size of this pool, it is easy to appreciate why the search of arthritogenic signatures in whole blood has been disappointing. Indeed, the majority of T cells (~90%–98%) display negligible similarity to their synovial counterparts,

in contrast to the tiny fraction (~2%–10%) of CPLs, which also correlate with disease activity. Therefore, when blood is investigated as a whole, most molecular events underlying the mechanisms of disease will be too diluted to emerge from noise. Our study provides researchers with the right target population, thereby removing from measurements that irrelevant background ultimately responsible for holding back progress in the understanding of autoimmune arthritis.

We first discovered CPLs in JIA, and then confirmed their presence and clinical relevance also in RA. An important direction for future studies is investigating the role of CPLs in other autoimmune diseases. These studies may prove particularly useful for conditions with difficult or no access to the site of inflammation, such as multiple sclerosis.

Author affiliations
[1]Translational Research Unit, Sanford-Burnham Medical Research Institute, San Diego, California, USA
[2]SingHealth Translational Immunology and Inflammation Centre, Singhealth and Duke-NUS Graduate Medical School, Singapore, Singapore
[3]Seattle Children's Hospital and Research Institute, Seattle, Washington, USA
[4]Lab Biotecnologie, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy
[5]Pediatric Rheumatology Unit, IRCCS Ospedale Pediatrico Bambino Gesù, Rome, Italy
[6]Second Pediatrics Division, University of Genoa and G Gaslini Institute, Genova, Italy
[7]Division of Rheumatology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[8]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA

Contributors RS and MR equally contributed to study design, experiments, data collection, analysis and interpretation, statistical analyses and writing of the manuscript. JvL contributed to some experiments and writing of the manuscript. CAW, MM, SM-M, MG, AM and DJL contributed to sample collection, data interpretation and writing of the manuscript. SA contributed to study design and oversight, data interpretation and writing of the manuscript.

Funding Work was supported by NIAMS (1R01 AR049762), the Bartman Foundation, the NMRC (NMRC/STaR/020/2013; Ministry of Health, Singapore) and Singhealth. JvL was supported by the Dutch Arthritis Foundation.

Competing interests None.

Ethics approval IRB of the Sanford-Burnham Medical Research Institute.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Maldonado L, Teague JE, Morrow MP, *et al*. Intramuscular therapeutic vaccination targeting HPV16 induces T cell responses that localize in mucosal lesions. *Sci Transl Med* 2014;6:221ra213.
2 Ferraro A, Socci C, Stabilini A, *et al*. Expansion of Th17 cells and functional defects in T regulatory cells are key features of the pancreatic lymph nodes in patients with type 1 diabetes. *Diabetes* 2011;60:2903–13.
3 Sathaliyawala T, Kubota M, Yudanin N, *et al*. Distribution and compartmentalization of human circulating and tissue-resident memory T cell subsets. *Immunity* 2013;38:187–97.
4 Grom AA, Thompson SD, Luyrink L, *et al*. Dominant T-cell-receptor beta chain variable region V beta 14+ clones in juvenile rheumatoid arthritis. *Proc Natl Acad Sci USA* 1993;90:11104–8.
5 Dornmair K, Meinl E, Hohlfeld R. Novel approaches for identifying target antigens of autoreactive human B and T cells. *Semin Immunopathol* 2009;31:467–77.
6 Prakken B, Albani S, Martini A. Juvenile idiopathic arthritis. *Lancet* 2011;377:2138–49.
7 Scott DL, Wolfe F, Huizinga TW. Rheumatoid arthritis. *Lancet* 2010;376:1094–108.
8 Robins HS, Campregher PV, Srivastava SK, *et al*. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 2009;114:4099–107.
9 Carlson CS, Emerson RO, Sherwood AM, *et al*. Using synthetic templates to design an unbiased multiplex PCR assay. *Nat Commun* 2013;4:2680.
10 Rempala GA, Seweryn M. Methods for diversity and overlap analysis in T-cell receptor populations. *J Math Biol* 2013;67:1339–68.
11 Cebula A, Seweryn M, Rempala GA, *et al*. Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota. *Nature* 2013;497:258–62.
12 Venturi V, Kedzierska K, Turner SJ, *et al*. Methods for comparing the diversity of samples of the T cell receptor repertoire. *J Immunol Methods* 2007;321:182–95.
13 Chao A, Chazdon RL, Colwell RK, *et al*. A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* 2005;8:148–59.
14 Maurice MM, Res PC, Leow A, *et al*. Joint-derived T cells in rheumatoid arthritis proliferate to antigens present in autologous synovial fluid. *Scand J Rheumatol Suppl* 1995;101:169–77.
15 De Graeff-Meeder ER, van der Zee R, Rijkers GT, *et al*. Recognition of human 60 kD heat shock protein by mononuclear cells from patients with juvenile chronic arthritis. *Lancet* 1991;337:1368–72.
16 Melchers I, Jooss-Rudiger J, Peter HH. Reactivity patterns of synovial T-cell lines derived from a patient with rheumatoid arthritis. I. Reactions with defined antigens and auto-antigens suggest the existence of multireactive T-cell clones. *Scand J Immunol* 1997;46:187–94.
17 Blass S, Schumann F, Hain NA, *et al*. p205 is a major target of autoreactive T cells in rheumatoid arthritis. *Arthritis Rheum* 1999;42:971–80.
18 Fang Q, Sun YY, Cai W, *et al*. Cartilage-reactive T cells in rheumatoid synovium. *Int Immunol* 2000;12:659–69.
19 Zhu L, Ji F, Wang Y, *et al*. Synovial autoreactive T cells in rheumatoid arthritis resist IDO-mediated inhibition. *J Immunol* 2006;177:8226–33.
20 Snir O, Backlund J, Bostrom J, *et al*. Multifunctional T cell reactivity with native and glycosylated type II collagen in rheumatoid arthritis. *Arthritis Rheum* 2012;64:2482–8.
21 Fritsch R, Eselbock D, Skriner K, *et al*. Characterization of autoreactive T cells to the autoantigens heterogeneous nuclear ribonucleoprotein A2 (RA33) and filaggrin in patients with rheumatoid arthritis. *J Immunol* 2002;169:1068–76.
22 Bangs SC, Baban D, Cattan HJ, *et al*. Human CD4+ memory T cells are preferential targets for bystander activation and apoptosis. *J Immunol* 2009;182:1962–71.
23 DeWolf WC, Schlossman SF, Yunis EJ. DRw antisera react with activated T cells. *J Immunol* 1979;122:1780–4.
24 Ko HS, Fu SM, Winchester RJ, *et al*. Ia determinants on stimulated human T lymphocytes. Occurrence on mitogen- and antigen-activated T cells. *J Exp Med* 1979;150:246–55.
25 Hewitt CR, Lamb JR, Hayball J, *et al*. Major histocompatibility complex independent clonal T cell anergy by direct interaction of *Staphylococcus aureus* enterotoxin B with the T cell antigen receptor. *J Exp Med* 1992;175:1493–9.
26 Hollsberg P, Wucherpfennig KW, Ausubel LJ, *et al*. Characterization of HTLV-I in vivo infected T cell clones. IL-2-independent growth of nontransformed T cells. *J Immunol* 1992;148:3256–63.
27 Kudo H, Matsuoka T, Mitsuya H, *et al*. Cross-linking HLA-DR molecules on Th1 cells induces energy in association with increased level of cyclin-dependent kinase inhibitor p27(Kip1). *Immunol Lett* 2002;81:149–55.
28 Lamb JR, Fledmann M. A human suppressor T cell clone which recognizes an autologous helper T cell clone. *Nature* 1982;300:456–8.
29 Lanzavecchia A, Roosnek E, Gregory T, *et al*. T cells can present antigens such as HIV gp120 targeted to their own surface molecules. *Nature* 1988;334:530–2.
30 LaSalle JM, Ota K, Hafler DA. Presentation of autoantigen by human T cells. *J Immunol* 1991;147:774–80.
31 LaSalle JM, Tolentino PJ, Freeman GJ, *et al*. Early signaling defects in human T cells anergized by T cell presentation of autoantigen. *J Exp Med* 1992;176:177–86.
32 Odum N, Martin PJ, Schieven GL, *et al*. Signal transduction by HLA class II antigens expressed on activated T cells. *Eur J Immunol* 1991;21:123–9.
33 Chini L, Bardare M, Cancrini C, *et al*. Evidence of clonotypic pattern of T-cell repertoire in synovial fluid of children with juvenile rheumatoid arthritis at the onset of the disease. *Scand J Immunol* 2002;56:512–7.
34 Wallace CA, Giannini EH, Spalding SJ, *et al*. Trial of early aggressive therapy in polyarticular juvenile idiopathic arthritis. *Arthritis Rheum* 2012;64:2012–21.
35 Wallace CA, Ruperto N, Giannini E. Preliminary criteria for clinical remission for select categories of juvenile idiopathic arthritis. *J Rheumatol* 2004;31:2290–4.
36 Koffeman EC, Genovese M, Amox D, *et al*. Epitope-specific immunotherapy of rheumatoid arthritis: clinical responsiveness occurs with immune deviation and relies on the expression of a cluster of molecules associated with T cell tolerance in a double-blind, placebo-controlled, pilot phase II trial. *Arthritis Rheum* 2009;60:3207–16.
37 Pitzalis C, Jones GW, Bombardieri M, *et al*. Ectopic lymphoid-like structures in infection, cancer and autoimmunity. *Nat Rev Immunol* 2014;14:447–62.
38 Chang CH, Hong SC, Hughes CC, *et al*. CIITA activates the expression of MHC class II genes in mouse T cells. *Int Immunol* 1995;7:1515–18.
39 Danke NA, Koelle DM, Yee C, *et al*. Autoreactive T cells in healthy individuals. *J Immunol* 2004;172:5967–72.
40 Vanderlugt CL, Miller SD. Epitope spreading in immune-mediated diseases: implications for immunotherapy. *Nat Rev Immunol* 2002;2:85–95.
41 Sabatino JJ Jr, Huang J, Zhu C, *et al*. High prevalence of low affinity peptide-MHC II tetramer-negative effectors during polyclonal CD4+ T cell responses. *J Exp Med* 2011;208:81–90.

Supplementary Materials and Methods

Samples

EDTA-anticoagulated blood and synovial fluid from 14 polyarticular and extended oligoarticular JIA patients with active disease were collected at the G. Gaslini Institute, Genoa, and IRCCS Policlinico S. Matteo Foundation, Pavia (Italy). Blood from 32 JIA patients before (T0) and after (Tend) at least 6 months of treatment with methotrexate, etanercept and prednisolone (tapered to zero in 4 months) was collected from participants in the TREAT study[34]. Blood from 39 RA patients was collected as part of the DNAJP1 study[36]. Blood from 14 adult healthy controls was collected at the TSRI Normal Blood Donor Service (La Jolla, CA). All samples were collected upon informed consent. All experiments have been conducted according to the principles expressed in the Declaration of Helsinki and IRB approved. Mononuclear cells were separated by density gradient with Histopaque-1077 (Sigma-Aldrich) and frozen in freezing medium (90% FCS, 10% DMSO).

Flow cytometry

Ex vivo immunophenotyping was performed immediately after thawing. Dead cells were excluded using Live/Dead Fixable Near-IR Stain (Life Technologies). Fc receptors (FcR) were blocked using FcR blocking reagent (Miltenyi Biotec) to avoid non-specific antibody binding. Intracellular staining was performed with the FOXP3 buffer set (eBioscience). To measure cytokine production, PBMCs were thawed and allowed to rest overnight before stimulation with 20 nM PMA and 200 nM ionomycin (Fisher Scientific) for 6 hours. GolgiPlug and GolgiStop (BD Biosciences) were added to cell culture 2 hours after PMA/ionomycin stimulation. Samples were acquired with an LSRFortessa (BD Biosciences). Antibodies were from Biolegend, BD Biosciences and eBioscience. Analysis was performed with FlowJo (Treestar). The gating strategy is shown in **Figure S1**. The stability of HLA-DR expression upon PMA/ionomycin stimulation is shown in **Figure S2**.

Cell culture

Complete medium was prepared as follows: RPMI-1640 (HyClone), 10% FBS (Fisher Scientific) for cell lines or 5% heat-inactivated human serum AB (GemCell) for fresh cells, 2 mM L-Glutamine (Gibco) and 100 U/ml penicillin and 100 µg/ml streptomycin (Gibco). T cell lines were established by stimulating freshly sorted HLA-DR$^-$CD14$^-$CD4$^+$ T cells with anti-CD3/CD28-coated beads (Life Technologies) at 1:5 ratio. Complete medium

supplemented with 20 U/ml rhIL-2 (Life Technologies) was added every 2-3 days. After 7 days, dead cells were removed with the Dead Cell Removal kit (Miltenyi Biotec) and viable cells were allowed to rest for 40 hours in complete medium. Before experiments, T cell lines were routinely tested by flow cytometry to confirm that at least 50% of cells were positive for HLA-DR.

Mixed lymphocyte reaction

$CD14^{-}CD4^{+}CD25^{low/-}$ T cells were sorted from thawed PBMCs after resting overnight in complete medium with 20 IU/ml rhIL-2. T cells were labeled with Cell Trace Violet (Life Technologies). Allogeneic HLA-DR-expressing T cell lines were labeled with CFDA-SE and used to stimulate sorted T cells at 1:1 ratio in the presence of rhIL-2 (50 IU/ml) and, where indicated, SEB (10 ng/ml, Sigma-Aldrich) to ensure HLA-DR was still expressed during the assay. SEB bridges specific TCR Vβ domains to MHC class II molecules, leading to nonspecific T cell activation[42]. As a control, allogeneic monocyte-derived dendritic cells activated with LPS (50 ng/ml, Sigma-Aldrich) were added to fresh T cells at 1:1 ratio. Dendritic cells were generated by incubating $CD14^{+}$ monocytes in complete medium with rhGM-CSF (50 ng/ml) and rhIL-4 (10 ng/ml, both cytokines from R&D Systems) for 6 days.

Calcium flux

HLA-DR-expressing T cell lines were loaded with 2 μM Fluo-4 AM (Life Technologies) for 30 min at RT, then placed in Tyrode's solution containing Sytox Red (Life Technologies) as a viability dye. Cells were pre-warmed at 37 degrees before analysis. Calcium flux was measured at 37 degrees for 30 seconds (baseline) + 1 min (addition of mouse primary Ab, 8 μg/ml) + 5 min (addition of anti-mouse IgM/G, 16 μg/ml). Ionomycin was used as a positive control at 2 μg/ml.

Statistical analyses

t-tests were performed for single pairwise comparisons, while ANOVA with post hoc tests was used for multiple comparisons (Tukey’s when all possible pairs were assessed, or Dunnett’s when each of a number of conditions was compared with a single control). Paired t-tests/repeated measures ANOVA were used in comparisons of matched blood, CPLs and synovium, while unequal variance unpaired t-tests/regular ANOVA were used in all other comparisons.

Hierarchical clustering of phenotypic data was performed using the euclidian distance and Ward linkage. Hierarchical clustering of the TCR repertoire was performed using the Chao-modified Jaccard index[12, 15] and single linkage.

Supplementary References

42. Li H, Llera A, Malchiodi EL, Mariuzza RA. The structural basis of T cell activation by superantigens. Annual review of immunology. 1999; 17:435-466.

Supplementary figure legends

Supplementary Figure 1. Gating strategy for immunophenotyping of HLA-DR$^+$ T cells. Total PBMCs were thawed and immediately stained with a viability dye, CD14, CD4, CD3, CD25, FOXP3 and HLA-DR, and gated as shown.

Supplementary Figure 2. Stability of HLA-DR upon PMA/ionomycin stimulation. Total PBMCs were allowed to rest overnight in complete medium and then stained with CD4, CD3, CD25, FOXP3 and HLA-DR, either immediately or after 6-hour stimulation with PMA/ionomycin. A representative donor out of 4 is shown.

Supplementary Figure 3. HLA-DR is not endowed with signaling activity in T cells. T cell lines (> 50% HLA-DR$^+$) were obtained by stimulating sorted HLA-DR$^-$CD14$^-$CD4$^+$ T cells from healthy controls for 7 days with anti-CD3/CD28-coated beads. A representative experiment out of 4 is shown. **A.** Calcium flux in Fluo-4-loaded T cell lines. The stimulus (either cross-linking secondary antibody or ionomycin) was added at the time indicated by the arrow. **B.** Antigen-presenting capacity of T cell lines in 5-day co-cultures with fresh autologous T cells labeled with CFSE. In selected conditions, SEB was added as a control that HLA-DR was still expressed during the assay. Allogeneic monocyte-derived dendritic cells activated with LPS were used as a positive control of antigen-presenting cells.

Supplementary Figure 4. CPLs are enriched in clonotypes of synovial T cells. Next-generation sequencing of TCRβ CDR3 sequences was performed on blood or synovial CD4$^+$ T cells of JIA patients with active disease. All data refer to in-frame nucleotide sequences. **A-B.** Renyi diversity indices for different values of α (**A**) and summary at α=1

(**B**; grey dashed line in panel A). Differential species diversity is substantiated only if the Renyi index is consistently greater (or lower) across all values of the α parameter. **C.** Clustering based on TCR repertoire distances, calculated as 1 - Chao-modified Jaccard index (Chao distance). **D.** Summary of the TCR repertoire pairwise Chao distances between CPL or blood samples and their paired synovial sample. **E**. Overlap of the TCR repertoire from CPL and blood samples with their matched synovial sample. Each open circle is the median overlap of 200 subsamples of the indicated size from either CPL or blood sample with 200 subsamples of the same size from their paired synovial sample. **F.** Summary for the overlap data as measured at the size of the smaller sample (grey dashed line in panel E). **G.** Vβ gene family usage. In A, C, E and G, each panel represents an individual patient.

Figure S1

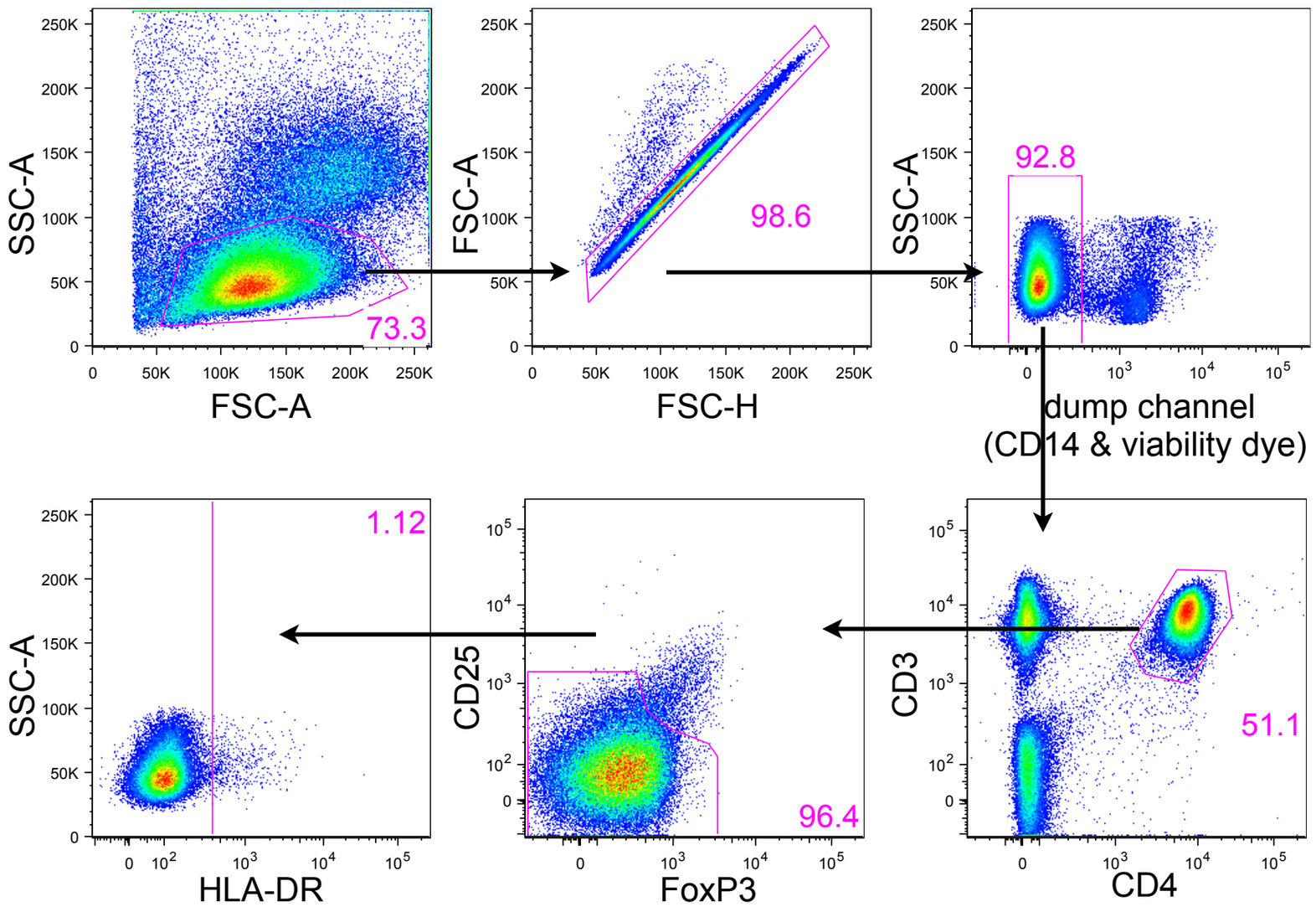

SSC-A
FSC-A
73.3
FSC-A
FSC-H
98.6
SSC-A
92.8
dump channel
(CD14 & viability dye)
SSC-A
HLA-DR
1.12
CD25
FoxP3
96.4
CD3
CD4
51.1

Figure S2

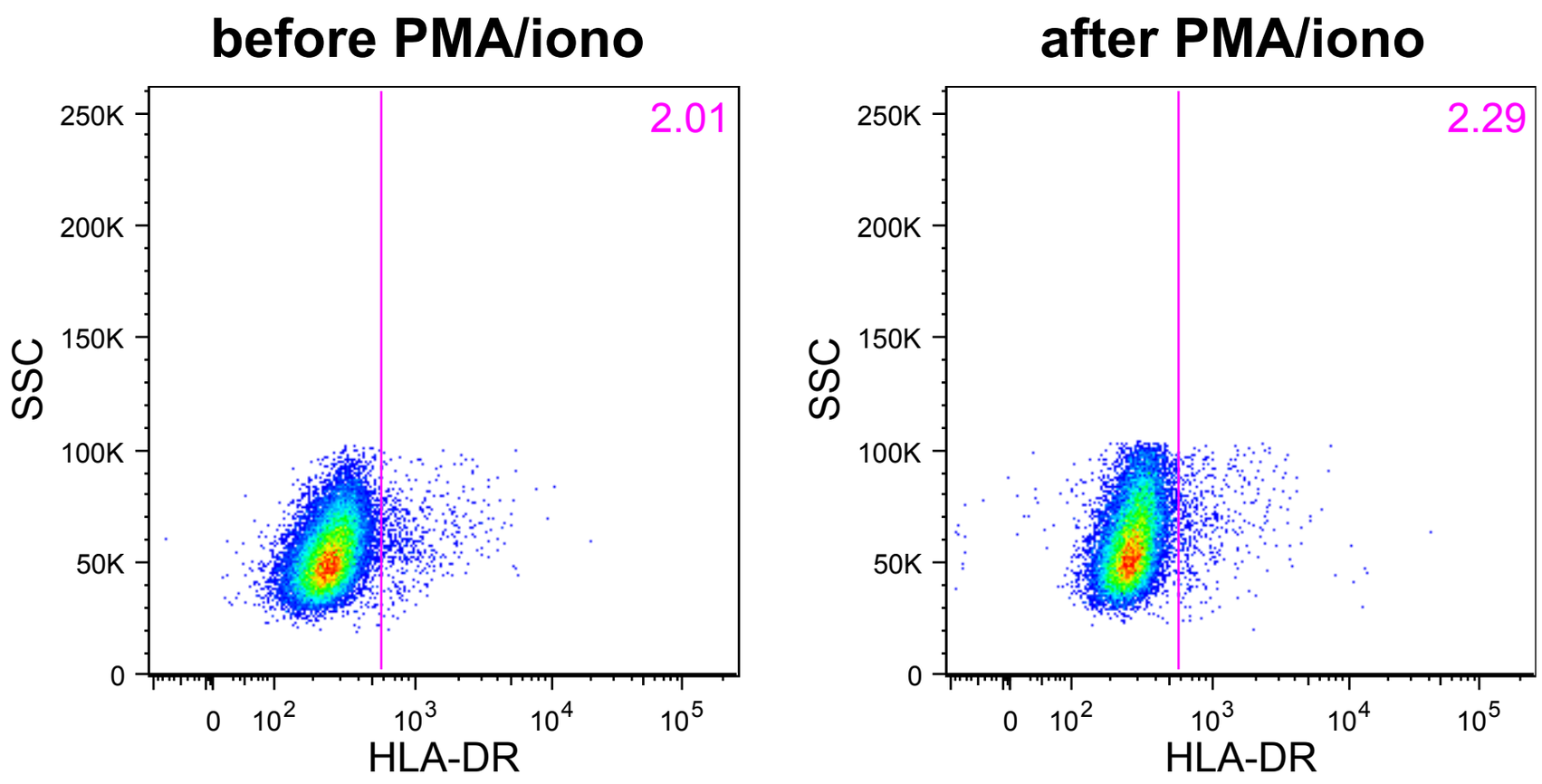

before PMA/iono
2.01
250K
200K
150K
100K
50K
0
SSC
0
10^2
10^3
10^4
10^5
HLA-DR
after PMA/iono
2.29
250K
200K
150K
100K
50K
0
SSC
0
10^2
10^3
10^4
10^5
HLA-DR

Figure S3

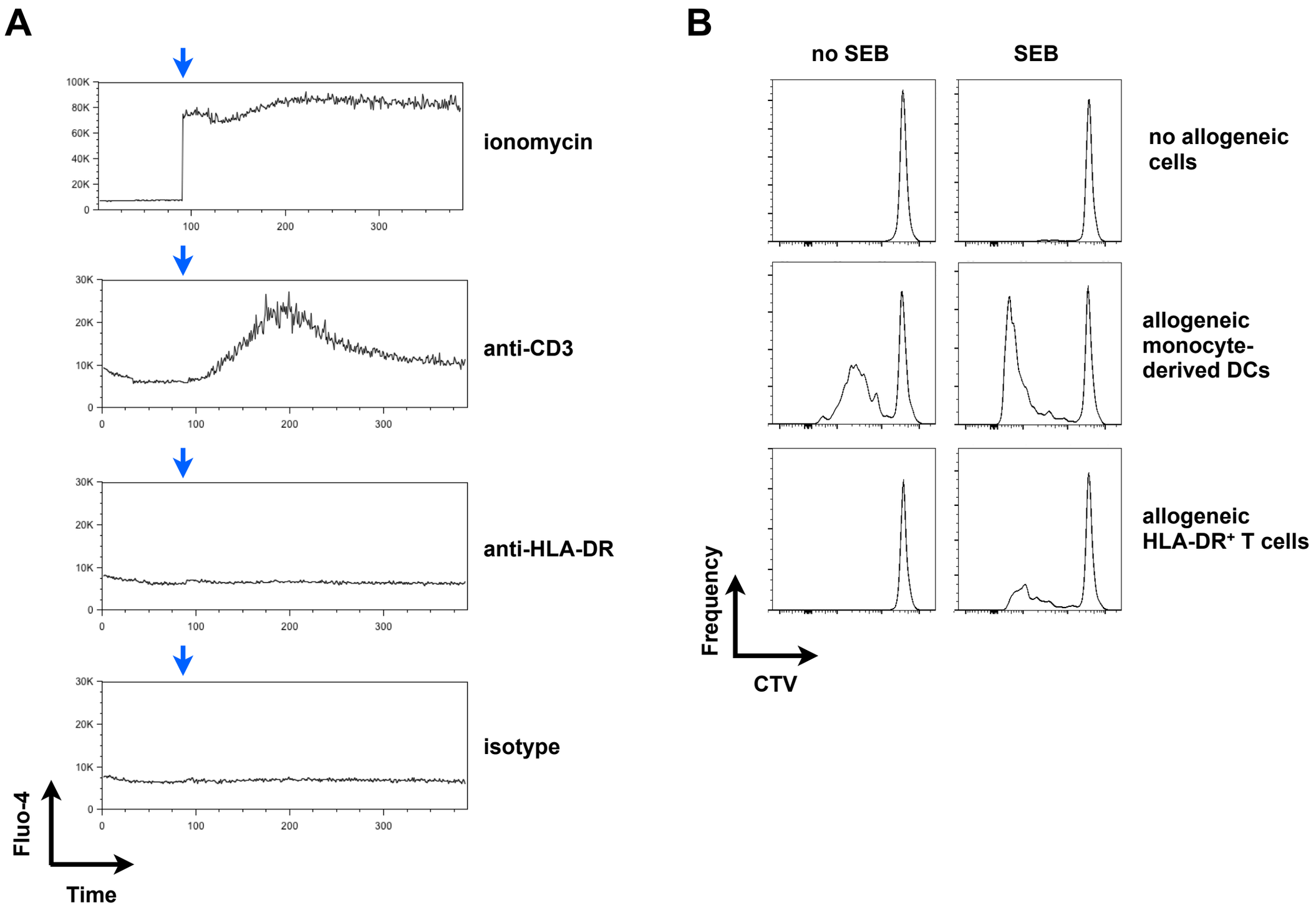
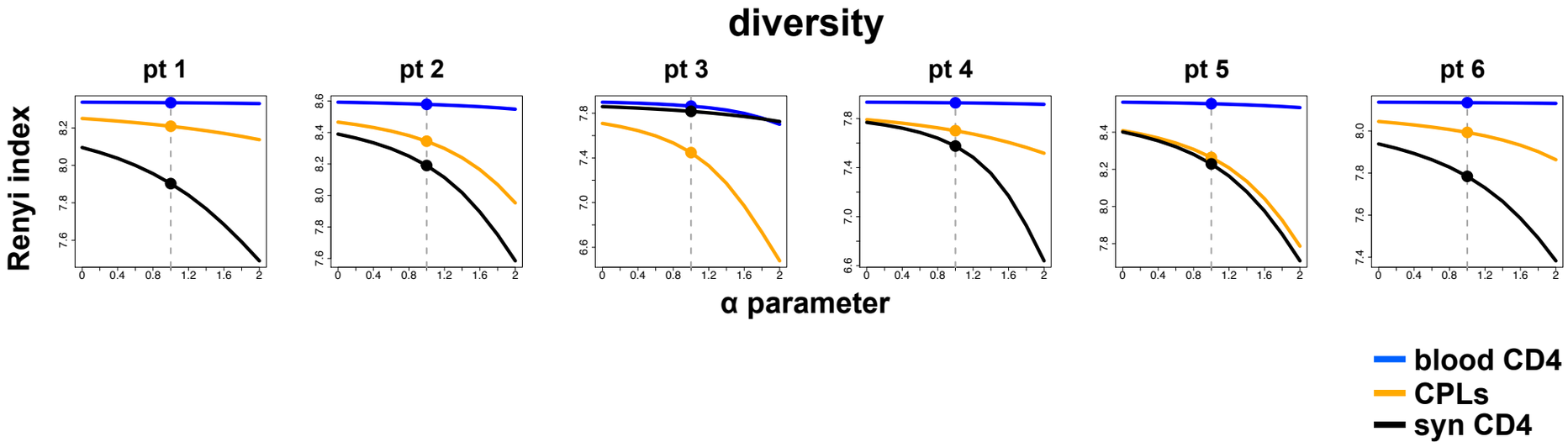
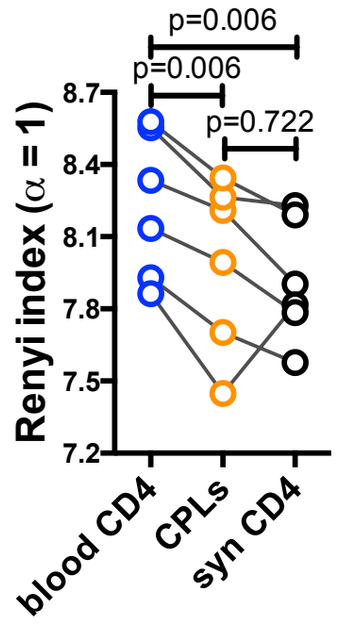
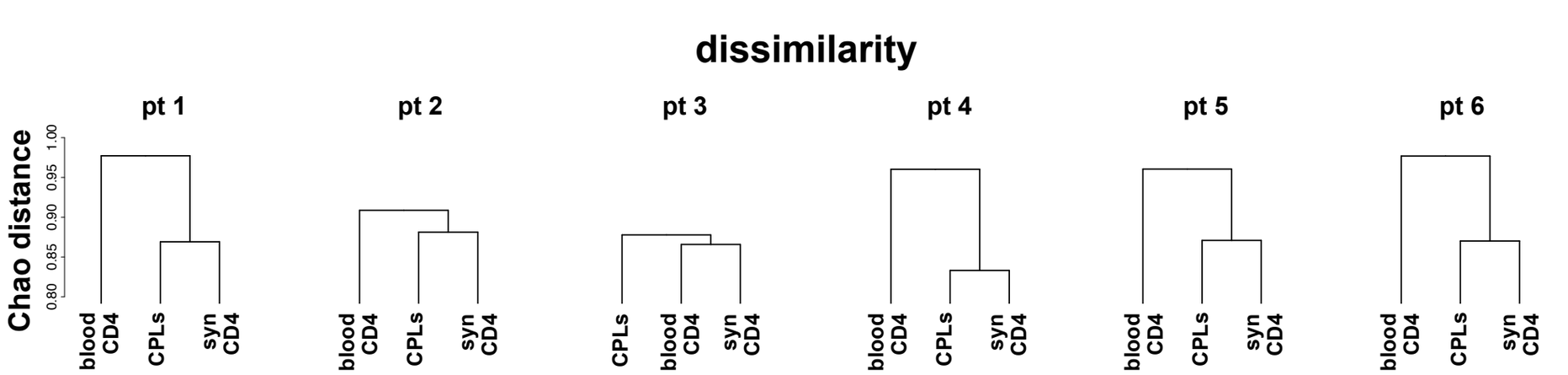
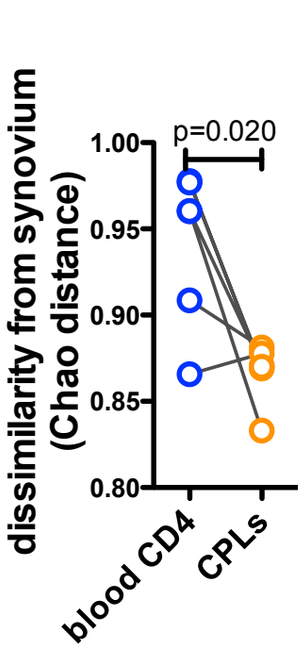

A
100K
80K
60K
40K
20K
0
100
200
300
ionomycin
30K
20K
10K
0
anti-CD3
anti-HLA-DR
isotype
Fluo-4
Time
B
no SEB
SEB
no allogeneic cells
allogeneic monocyte-derived DCs
allogeneic HLA-DR+ T cells
Frequency
CTV

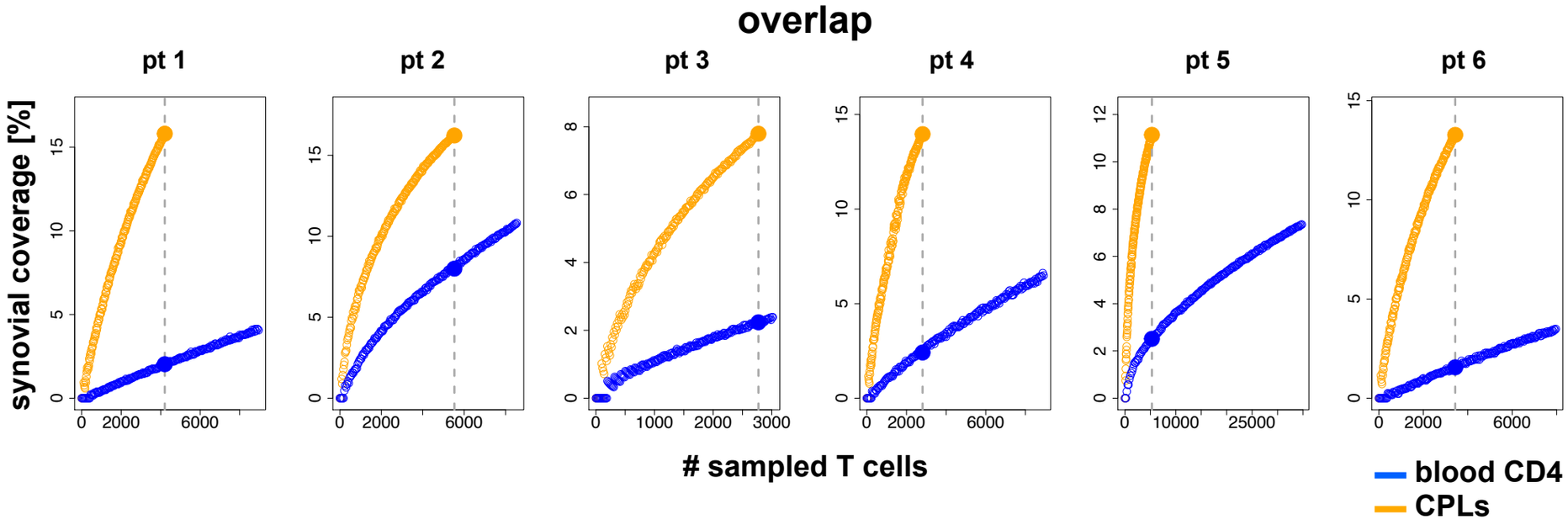
Figure S4

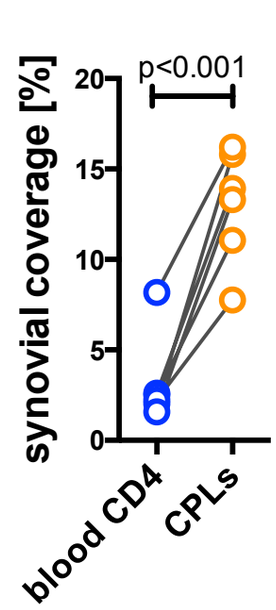

A
diversity
pt 1
pt 2
pt 3
pt 4
pt 5
pt 6
Renyi index
α parameter
blood CD4
CPLs
syn CD4

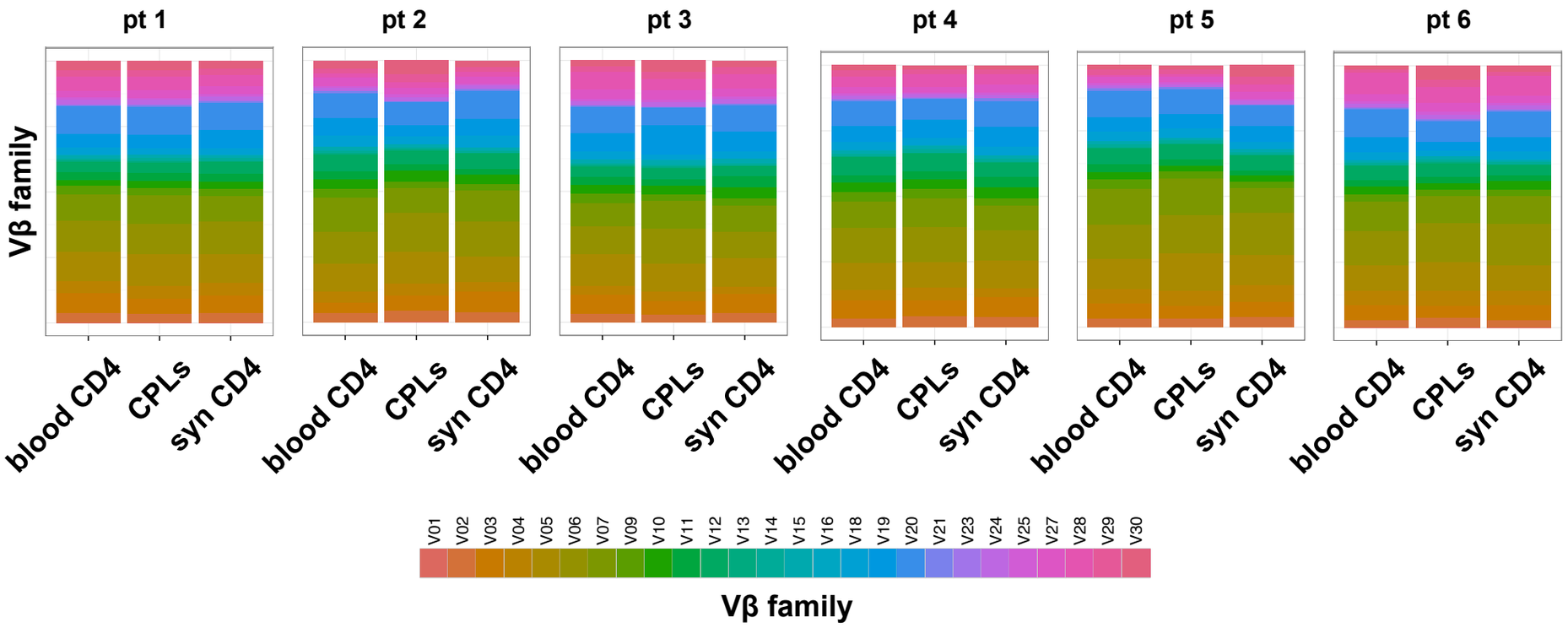

B
p=0.006
p=0.006
p=0.722
Renyi index (α = 1)
blood CD4
CPLs
syn CD4


C
dissimilarity
pt 1
pt 2
pt 3
pt 4
pt 5
pt 6
Chao distance
blood CD4
CPLs
syn CD4


D
dissimilarity from synovium
(Chao distance)
p=0.020
blood CD4
CPLs


E
overlap
pt 1
pt 2
pt 3
pt 4
pt 5
pt 6
synovial coverage [%]
sampled T cells
blood CD4
CPLs


F
synovial coverage [%]
p<0.001
blood CD4
CPLs


G
pt 1
pt 2
pt 3
pt 4
pt 5
pt 6
Vβ family
blood CD4
CPLs
syn CD4
V01
V02
V03
V04
V05
V06
V07
V09
V10
V11
V12
V13
V14
V15
V16
V18
V19
V20
V21
V23
V24
V25
V27
V28
V29
V30
Vβ family

```
# ====================================
#
# This script analyzes TCRB sequencing data as returned by
Adaptive Technologies and made available online
#
# - Each tab-delimited file is a single sample: blood, CPL or
synovium of a patient. File naming convention is
Patient_X_Blood|CPL|Synovium_TCRB.tsv
# - Each row is a nucleotide sequence (column "nucleotide"),
with its in-silico translation (column "aminoAcid"), in-frame
information (column "sequenceStatus"), T cell genome count
(column "estimatedNumberGenomes") and V Family Name (column
"vFamilyName")
#
# ====================================

# === Analysis options - BEGIN
seq.type <- "aminoAcid"  # set whether analysis is on
"aminoAcid" or "nucleotide" sequences
sampling.replace <- F  # set whether sampling is with (T) or
without (F) replacement
n.subs <- 200  # set number of subsamples
n.intervals <- 200  # set number of sample size intervals for
analysis of TCR repertoire overlap
# === Analysis options - END

# === Load required packages - BEGIN
library(plyr)
library(dplyr)
library(vegan)
library(ggplot2)
library(Biobase)
library(foreach)
library(doMC)
library(iterators)
# === Load required packages - END

registerDoMC(cores=detectCores())
mcoptions <- list(preschedule=T, set.seed=F)

samples <- c("Blood", "CPL", "Synovium")
chart.colors <- c("blue", "orange", "black")

# === Load data - BEGIN
```

```
dir.create(file.path(".", "results"), showWarnings=F)
index <- list.files(path=".", pattern="tsv")  # assumes this
script runs from a folder containing all data files
cat("Analysis in progress...\n")
all.data <- llply(index, read.delim)
names(all.data) <- gsub("_TCRB.tsv", "", index)
all.data <- lapply(all.data, tbl_df)
all.data <- lapply(all.data, function(x) dplyr::filter(x,
sequenceStatus == "In"))  # retain productive TCR
rearrangements
all.data <- lapply(all.data, function(x) dplyr::select(x,
nucleotide, aminoAcid, estimatedNumberGenomes, vFamilyName))
# === Load data - END

# === Vb family analysis - BEGIN
v.counts <- lapply(all.data, function(x)
aggregate(estimatedNumberGenomes ~ vFamilyName, FUN=sum,
data=x))
v.counts <- lapply(v.counts, function(x) x[x$vFamilyName != ""
& x$vFamilyName != "TCRBVA", ])
for(i in 1:length(v.counts)) v.counts[[i]] <-
cbind(v.counts[[i]], dataset=names(v.counts[i]))
v.tot <- sapply(v.counts, function(x)
sum(x$estimatedNumberGenomes))
for(i in 1:length(v.counts))
v.counts[[i]]$estimatedNumberGenomes <-
v.counts[[i]]$estimatedNumberGenomes / v.tot[i]
all.v <- data.frame()
for(i in 1:length(v.counts)) all.v <- rbind(all.v,
v.counts[[i]])
all.v <- droplevels(all.v)
all.v$vFamilyName <- factor(all.v$vFamilyName,
levels=levels(all.v$vFamilyName)
[order(levels(all.v$vFamilyName))])
levels(all.v$vFamilyName) <- sub("TCRB", "",
levels(all.v$vFamilyName))
all.v$Group <- substr(all.v$dataset, 1, 9)
all.v <- all.v[!is.na(all.v$Group), ]
all.v$Group <- factor(all.v$Group)
all.v$dataset <- as.character(all.v$dataset); all.v$dataset <-
substr(all.v$dataset, 11, nchar(all.v$dataset))
all.v$dataset <- factor(all.v$dataset)

# plot results
v.chart <- qplot(data=all.v, geom="bar", position="stack",
```

```
x=dataset, y=estimatedNumberGenomes, fill=vFamilyName,
stat="identity") + facet_wrap(~ Group, drop=T, scales="free_x",
ncol=3) + theme_bw() + scale_fill_hue(l=58, c=90) +
theme(legend.position="right", legend.title=element_blank(),
axis.title.x=element_blank(), axis.title.y=element_blank(),
axis.ticks.y=element_blank())
ggsave(file.path(".", "results", "V_Family.pdf"), v.chart,
width=10, height=8)
rm(list=c("index", "all.v", "v.counts", "v.tot", "v.chart",
"i"))
# === Vb family analysis - END

# === Reshape data - BEGIN
subset.data <- lapply(all.data, function(x) x[, c(seq.type,
"estimatedNumberGenomes")])
for(z in 1:length(subset.data)) colnames(subset.data[[z]]) <-
c("seq", "n")
subset.data <- lapply(subset.data, function(x) x %>%
dplyr::filter(seq != "") %>% group_by(seq) %>%
dplyr::summarise(n=sum(n)))
for(z in 1:length(subset.data)) subset.data[[z]]$seq <-
as.character(subset.data[[z]]$seq)
patients <- unique(substr(names(subset.data), 1, 9))
patient.data <- vector("list", length(patients))
names(patient.data) <- patients; rm(patients)
for(i in 1:length(patient.data)) {patient.data[[i]] <-
vector("list", 3); names(patient.data[[i]]) <-
paste(names(patient.data[i]), samples, sep="_")}
for(i in 1:length(patient.data))
    for(z in 1:length(samples))
        patient.data[[i]][[z]] <-
subset.data[[which(grepl(names(patient.data[[i]][z]),
names(subset.data)))]]
rm(list=c("all.data", "subset.data", "seq.type", "i", "z"))
# === Reshape data - END

foreach(p=iter(patient.data, .inorder=F)) %dopar% {
    sink(file.path(".", "results", paste(substr(names(p[1]), 1,
9), ".txt", sep="")))

    # === Chao distance - BEGIN
    chao.data <- Map(function(x, i) setNames(x, ifelse(names(x)
%in% c("seq"), names(x), sprintf('%s.%d', names(x), i))), p,
seq_along(p))
```

```
    chao.data <- Reduce(function(x, y) merge(x, y, sort=F,
all=T, by="seq"), chao.data, accumulate=F)
    chao.data[is.na(chao.data)] <- 0
    chao.data <- t(chao.data[, -1])
    rownames(chao.data) <- samples

    # calculate the index
    chao.dist <- vegdist(chao.data, method="chao")

    # plot results
    pdf(file.path(".", "results", paste(substr(names(p[1]), 1,
9), "ChaoDistance.pdf", sep="_")), height=8, width=5,
onefile=T)
    par(lwd=4); par(cex.axis=2.5)
    plot(hclust(chao.dist, method="single"), main="", xlab="",
sub="")
    plot(as.dendrogram(hclust(chao.dist, method="single")),
main="", ylim=c(0.8, 1))
    dev.off()

    cat("\nChao Distance:\n"); print(chao.dist)  # outputs the
Chao distances to allow calculating summary statistics for all
patients

    rm(list=c("chao.data", "chao.dist"))
    # === Chao distance - END

    # convert data to one-row-per-sequence format
    samples.data <- lapply(p, function(y)
unlist(lapply(1:nrow(y), FUN=function(x) rep(y$seq[x],
times=y$n[x]))))
    names(samples.data) <- samples

    # === Renyi index - BEGIN

    # calculate the index
    samples.renyi <- lapply(lapply(samples.data, function(y)
lapply(1:n.subs, function(x) table(sample(y,
min(sapply(samples.data, length)),
replace=sampling.replace)))), function(w)
rowMedians(simplify2array(lapply(1:n.subs, function(s)
renyi(w[[s]], scales=seq(0, 2, length.out=11))))))

    # plot results
    pdf(file.path(".", "results", paste(substr(names(p[1]), 1,
```

```
9), "RenyiIndex.pdf", sep="_")), height=8, width=8)
    plot(c(1:11), type="n", xaxt="n", xlab="alpha", ylab="Renyi
diversity index", ylim=c(min(sapply(samples.renyi, min)),
max(sapply(samples.renyi, max))), cex.axis=2, cex.lab=1.5)
    axis(1, at=1:11, labels=seq(0, 2, length.out=11),
cex.axis=2)
    for(t in 1:length(samples.renyi)) lines(samples.renyi[[t]],
col=chart.colors[t], lwd=10)
    abline(v=6, lty="dashed", col="darkgrey", lwd=6)
    for(t in 1:length(samples.renyi)) points(6,
samples.renyi[[t]][6], cex=4, pch=19, col=chart.colors[t])
    legend("bottomleft", legend=samples, col=chart.colors,
lty=1, lwd=3)
    dev.off()

    cat("\nRenyi Index at alpha=1:\n"); for(t in
1:length(samples.renyi)) cat(names(samples.renyi[t]),
samples.renyi[[t]][6], "\n")  # outputs the Renyi indexes to
allow calculating summary statistics for all patients

    rm(list=c("samples.renyi", "t"))
    # === Renyi index - END

    # === Pairwise overlap - BEGIN
    pos.x1 <- grep("Blood", samples)
    pos.x2 <- grep("CPL", samples)
    pos.ref <- grep("Synovium", samples)
    min.x1 <- min(length(samples.data[[pos.x1]]),
length(samples.data[[pos.ref]]))
    min.x2 <- min(length(samples.data[[pos.x2]]),
length(samples.data[[pos.ref]]))
    min.xs <- min(min.x1, min.x2)
    acc.x1 <- acc.x2 <- vector("numeric", length=n.intervals)

    compute.overlap <- function(pair, sample.size) {
        pair.table <- lapply(pair, function(y) lapply(1:n.subs,
function(x) data.frame(table(sample(y, sample.size,
replace=sampling.replace)))))
        return(median(simplify2array(lapply(1:n.subs,
function(y) {
            d.ref <- pair.table[[2]][[y]]
            common.set <- intersect(pair.table[[1]][[y]][, 1],
d.ref[, 1])
            return(sum(d.ref[d.ref[, 1] %in% common.set, 2]) /
```

```
sum(d.ref[, 2]))
        })))
    }

    ovp.x1 <- compute.overlap(samples.data[c(pos.x1, pos.ref)],
min.xs) * 100
    ovp.x2 <- compute.overlap(samples.data[c(pos.x2, pos.ref)],
min.xs) * 100

    cat("\nOverlap with Synovium reference by
subsampling:\nBlood:", ovp.x1, "\nCPL:", ovp.x2)  # outputs the
overlap by subsampling to allow calculating summary statistics
for all patients

    ind.x1 <- round(seq.int(1, min.x1, length.out=n.intervals))
    for(t in 1:length(ind.x1))
        acc.x1[[t]] <- compute.overlap(samples.data[c(pos.x1,
pos.ref)], ind.x1[t])
    acc.x1 <- acc.x1 * 100

    ind.x2 <- round(seq.int(1, min.x2, length.out=n.intervals))
    for(t in 1:length(ind.x2))
        acc.x2[[t]] <- compute.overlap(samples.data[c(pos.x2,
pos.ref)], ind.x2[t])
    acc.x2 <- acc.x2 * 100

    save(ovp.x1, ovp.x2, min.xs, acc.x1, acc.x2, ind.x1,
ind.x2, file=file.path(".", "results",
paste(substr(names(p[1]), 1, 9), "OverlapBySubsampling.rda",
sep="_")))

    # plot results
    pdf(file.path(".", "results", paste(substr(names(p[1]), 1,
9), "OverlapBySubsampling.pdf", sep="_")), width=5, height=8)
    plot(ind.x2, acc.x2, xlim=c(0, max(ind.x1, ind.x2)),
ylim=c(0, max(acc.x1, acc.x2)) * 1.1, col=chart.colors[2],
xlab="# sampled T cells", ylab="Synovial Coverage [%]", cex=2,
cex.lab=1.5)
    points(ind.x1, acc.x1, col=chart.colors[1], cex=2)
    abline(v=min.xs, lty="dashed", col="darkgrey", lwd=4)
    points(min.xs, ovp.x1, pch=19, col=chart.colors[1], cex=4)
    points(min.xs, ovp.x2, pch=19, col=chart.colors[2], cex=4)
    legend("topright", legend=c("Blood ", "CPL "),
col=c(chart.colors[1], chart.colors[2]), lty=c(1, 1), lwd=4,
pch=c(NA, NA))
```

```
    dev.off()

    rm(list=c("pos.x1", "pos.x2", "pos.ref", "min.x1",
"min.x2", "min.xs", "ovp.x1", "ovp.x2", "t", "ind.x1",
"ind.x2", "acc.x1", "acc.x2", "compute.overlap"))
    # === Pairwise overlap - END

    sink()
    return(paste(substr(names(p[1]), 1, 9), "completed!", sep="
"))
}
```