ONLINE DATA SUPPLEMENT

T cell receptor-HLA-DRB1 associations suggest specific antigens in pulmonary sarcoidosis

Johan Grunewald[1*], Ylva Kaiser[1], Mahyar Ostadkarampour[1], Natalia V Rivera[1], Francesco Vezzi[2], Britta Lötstedt[3], Remi-André Olsen[2], Lina Sylwan[3], Sverker Lundin[4], Max Käller[4], Tatiana Sandalova[5], Kerstin Ahlgren[1], Jan Wahlström[1], Adnane Achour[5], Marcus Ronninger[1], Anders Eklund[1].

[1]Respiratory Medicine Unit, Department of Medicine Solna and CMM, Karolinska Institutet and Karolinska University Hospital, Solna, Sweden.

[2]Science for Life Laboratory (SciLifeLab), Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden

[3]Science for Life Laboratory (SciLifeLab), Department of Biosciences and Nutrition, Karolinska Institutet, Solna, Sweden

[4]Science for Life Laboratory (SciLifeLab), Royal Institute of Technology (KTH), Gene Technology, 171 65 Solna, Sweden,

[5]Science for Life Laboratory (SciLifeLab), Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden.

Supplementary Material and Methods

HLA typing

Genomic DNA was extracted from whole blood samples and HLA-DRB1 and DRB3 alleles were determined using the PCR-SSP (SSP = sequence-specific primers) technique (Olerup SSP DR low resolution kit, Saltsjöbaden, Sweden) (S1).

Flow cytometry

BAL fluid T cells were stained using the following antibodies: CD3-Pacific Blue, clone UCHT1 (BD Pharmingen, San Diego, CA, USA), CD4-APC-H7, clone SK3 (BD Biosciences, San Jose, CA, USA), Vα2.3-FITC, clone F1 (Thermo Scientific, Rockford, IL, USA) and Vβ22-PE, clone IMMU 546 (Beckman Coulter Immunotech, Marseille, France) (S2). For assessment of cell activation, we also included CD27-PerCP-Cy5.5, clone M-T271 and CD69-PE-Cy7, clone FN50 (BD Pharmingen). Flow cytometry was run on a BD FACS Canto II (Beckton Dickinson, San Jose, CA, USA) and results were analysed using BD FACSDiva v.8.0 and FlowJo X (TreeStar, Ashland, OR, USA) softwares.

mRNA extraction and cDNA generation

Total cellular mRNA was extracted either from FACS-sorted cells or from total BAL cells (3-10 x 10^6, depending on cell availability) using the RNeasy Plus Mini Kit (Qiagen, Hilden, Germany), according to manufacturer's protocol. Complementary DNA (cDNA) was generated from total RNA using the High-Capacity cDNA Reverse Transcription Kit, containing a dNTP mix, RT random primers and MultiScribe reverse transcriptase (Applied Biosystems, **Waltham, MA, USA**), following manufacturer's instructions.

TCR α and β gene amplification

PCR amplification of Vα2.3 and Vβ22 genomic regions, respectively, was performed using a specific variable (V) region oligonucleotide forward primer and a conserved constant (C) region reverse primer for each TCR chain. Sequences for the primers used were as follows: Forward Vα2.3 5'–GTGTTCCAGAGGGAGCCATTGCC–3', Reverse Cα 5'–AATAGGTCGACAGACTTGTCACTGGA–3'. Forward Vβ22 5'–AGGACCAGATGCCTGAGCTA–3', Reverse Cβ 5'–CTGGGTCCACTCGTCATTCT–3'. 12.5 µl REDTaq ReadyMix PCR Reaction Mix (Sigma-Aldrich, Saint Louis, MO, USA) containing standard Taq DNA polymerase, optimised buffer components and inert dye (for agarose gel visualisation of PCR products) was added to 8 µl cDNA and 10 µM primers to yield a final reaction volume of 25 µl. PCR reactions were carried out in a 2720 Thermal Cycler system (Applied Biosystems) under the following conditions for Vα2.3: initialisation 94°C 1 min; denaturation 94°C 1 min, annealing 55°C 1.5 min and elongation 72°C 1 min, repeated in 40 cycles; final elongation 72°C, 10 min. The same procedure was performed for Vβ22, except for an annealing temperature of 51°C and 42 cycles. PCR products were quality-assessed by loading 10 µl of each sample and 3 µl CoralRed ladder onto a 1.5% agarose gel stained with GelRed Nucleic Acid Gel Stain 10,000X (Biotium, Hayward, CA, USA) in 1X TAE buffer. Electrophoresis was performed at a constant voltage of 100V for 1.5 h, followed by visualisation under UV light.

*Three-dimensional molecular modelling of ternary TCR/DRB1*0301/peptide complexes*

Vα2.3/Vβ22 TCR sequences were retrieved from the IMGT database (S3), according to which Vα2.3/Vβ22 CDR-loops comprise five or six residues within the sequences NSASQS, SNHLY, VYSSGN and FYNNEI for CDR1α, CDR1β, CDR2α and CDR2β, respectively. The lengths of

these loops correspond to lengths observed in crystal structures of TCRs with 5-8 residues per CDR1/2 loop. While the CDR1α, CDR2α and CDR2β loops of Vα2.3/Vβ22 TCRs comprise six residues (NSASQS, VYSSGN, FYNNEI, respectively), the CDR1β loop consists of only five residues (SNHLY). All modelling was performed manually using the COOT program (S4). No ternary crystal structure of the DRB1*0301 allele in complex with a TCR is yet available. The crystal structure of the Ob.1A12 TCR in complex with HLA-DRB1*1501 and the myelin basic protein-derived peptide 85-98 (Protein Database Base (PDB) code 1YMM) (S5), was used as a template for modelling. The HLA molecule was replaced by the crystal structure of the DRA1/DRB1*0301/CLIP complex (PDB code 1A6A) (S6). The TCR CDR1/2 loops were changed to sequences corresponding to Vα2.3/Vβ22 classes (Table 3). Based on this template, TCR models were created with different CDR3 α/β chains. Adequate residues were mutated to correct sequences of targeted CDR3 loops and structure idealisation was thereafter performed in COOT using the "sphere regularisation" option in order to remove unfavourable contacts between residues.

Mate pair sequencing library preparation and sequencing

A common sequencing library for each sample was prepared by pooling the two PCR reactions and by the Illumina Nextera XT kit (Illumina, San Diego, CA, USA). The work was done following the protocol from the manufacturer, except the following steps that were adapted for automation on an Agilent NGS workstation (Agilent, Santa Clara, CA, USA): tagmentation (enzymatic fragmentation and adapter ligation), PCR amplification and finally amplicon purification using Dynabeads MyOne carboxylic acid beads (Thermo Fischer Scientific, Waltham, MA, USA) (S7). Libraries were normalised and pooled for MiSeq 2x250 sequencing.

By using this approach for library preparation, PCR-amplicons longer than the total read length (2x250 bp in this case) can be used since the tagmentation reaction randomly fragments the DNA amplicons and prepares the whole of them for DNA sequencing. This approach facilitate a counting of reads from individual TCR clone transcripts and thus measuring the relative abundance of the clones in each isolate without the need for, for instance, individual cloning and Sanger DNA sequencing.

De novo *transcript assembly*

In order to bioinformatically process the samples the following pipeline has been designed: (i) adaptor removal and trimming of low quality sequence ends using Trimmomatic (S8); (ii) discarding of read-pairs whose length was shorter than 150 base pairs; (iii) down-sampling of the survived sequences; (iv) *de novo* assembling of the subsampled set using Trinity (S9); (v) computing transcript abundance estimates using RSEM (S10).

Points (i), (ii), and (iii) are typical *de novo* assembly pre-processing steps. In particular, points (i) and (ii) allow us to exclude from downstream analysis shorter fragments which may stem from impurities. Point (iii), was implemented to level out all samples and run downstream analysis on a uniform set of data points. For this purpose, we decided to select 80.000 reads for each sample as this allowed us to have the same coverage across all samples. The transcript assembler Trinity was used in point (iv) to find the original spliced mRNA sequences, including the variable splice-site nucleotides, without using a reference sequence. It produced between 1 and 200 assembled transcripts for each sample. In point (v) RSEM has been used to perform abundance estimates in order to evaluate which is most likely to be the target amplicon. The most abundant sequences (*i.e.* has the most support in the sequencing data) have been selected for downstream analysis. In doing this we assume

that the less frequent sequences are likely to be the consequence of misassemblies or impurities.

DNA-sequence bioinformatic analysis

The resulting sequences were functionally analysed with tools made available by IMGT (S11), which include a comprehensive database of T cell receptor nucleotide sequences of human origin. IMGT/V-QUEST was used to identify T cell receptor domains and amino acid sequences and their positions in CDR3 regions. Comparison of amino acid frequency and sequences for all T cell CDR3 junctions of human origin was performed using IMGT/LIGM-DB, and subsequently used for statistical analysis by the Chi-square test; $p < 0.05$ was considered significant.

References

S1. Olerup O, Zetterquist H. HLA-DR typing by PCR amplification with sequence-specific primers (PCR-SSP) in 2 hours: an alternative to serological DR typing in clinical practice including donor-recipient matching in cadaveric transplantation. Tissue Antigens. 1992;39(5):225-35.

S2. Ahlgren KM, Ruckdeschel T, Eklund A, Wahlstrom J, Grunewald J. T cell receptor-Vbeta repertoires in lung and blood CD4+ and CD8+ T cells of pulmonary sarcoidosis patients. BMC Pulm Med. 2014;14:50.

S3. Ehrenmann F, Kaas Q, Lefranc MP. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. Nucleic Acids Res. 2010;38(Database issue):D301-7.

S4. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010;66(Pt 4):486-501.

S5. Hahn M, Nicholson MJ, Pyrdol J, Wucherpfennig KW. Unconventional topology of self peptide-major histocompatibility complex binding by a human autoimmune T cell receptor. Nat Immunol. 2005;6(5):490-6.

S6. Wahlstrom J, Dengjel J, Persson B, Duyar H, Rammensee HG, Stevanovic S, et al. Identification of HLA-DR-bound peptides presented by human bronchoalveolar lavage cells in sarcoidosis. J Clin Invest. 2007;117(11):3576-82.

S7. Lefranc MP. IMGT, The International ImMunoGeneTics Information System, http://imgt.cines.fr. Methods Mol Biol. 2004;248:27-49.

S8. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

S9. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644-52.

S10. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

S11. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. 2009;37(Database issue):D1006-12.

Table S1. Clinical characteristics of HLA-DRB1*03 negative patients, divided into DRB3*01 positive or DRB3*01 negative patients.

	Among HLA-DRB1*03 negative and DRB3*01 positive patients (n=6)	Among HLA-DRB1*03 negative and DRB3*01 negative patients (n=11)
Sex (male/female)	5/1	7/4
Age, years	35.0 (33.0-38.5)	59.0 (43.5-65.0)
LS	3	1
Smoking status (non-smoker/former/current)	3/1/2	5/6/0
[1]Chest radiographic stage 0/I/II/III/IV	0/1/5/0/0	0/4/4/2/1
[2]VC (% of predicted)	80.0 (75.0-92.0)	85.0 (78.0-93.0)
FEV1 (% of predicted)	79.0 (76.0-83.0)	73.0 (63.5-84.5)
DLCO (% of predicted)	84.0 (77.5-86.5)	84.0 (75.5-94.0)
% BAL recovery	61.0 (60.3-67.0)	60.0 (47.0-69.5)
[3]BAL cell concentration (10^6 cells/L)	296.3 (238.0-399.7)	198.0 (101.1-333.8)
% BALF macrophages	53.2 (45.1-74.2)	73.6 (51.8-79.7)
% BALF lymphocytes	45.8 (24.0-53.2)	25.6 (17.9-45.0)
% BALF neutrophils	1.5 (0.7-1.6)	1.2 (0.8-2.0)
% BALF eosinophils	0.2 (0.0-0.6)	0.2 (0.0-0.6)
BAL CD4/CD8 ratio	11.8 (7.3-15.5)	6.0 (2.7-8.1)

[1]Chest radiography staging as follows: stage 0 = normal chest radiography, stage I = enlarged lymph nodes, stage II = enlarged lymph nodes with parenchymal infiltrates, stage III = parenchymal infiltrates without enlarged lymph nodes and stage IV = signs of pulmonary fibrosis.

[2]VC= vital capacity, FEV1=forced expiratory volume in one second, DLCO= carbon monoxide diffusing capacity.

[3]BAL basophils and mast cells were excluded from the cell differential counts.

All percentage values are denoted as median (p25-p75).

Table S2. Summary of all BAL $CD4^+$ T lymphocytes that express Vα2.3, Vβ22, and Vα2.3 together with Vβ22. % of $Vα2.3^+$ $CD4^+$ BAL T cells that express Vβ22, and % of $Vβ22^+$ $CD4^+$ BAL T cells that express Vα2.3 are also stated.

All data are shown for HLA-DRB1*03 negative patients divided into DRB3*01 positive or DRB3*01 negative patients.

	Among HLA-DRB1*03 negative and DRB3*01 positive patients (n=6)	Among DRB1*03 negative and DRB3*01 negative patients (n=11)
% Vα2.3+ of BAL CD4+ T cells	18.1 (17.0-19.0)	4.5 (4.3-5.1)
% Vβ22+ of BAL CD4+ T cells	2.7 (2.3-3.2)	2.0 (1.3-3.3)
% Vα2.3+/Vβ22+ of BAL CD4+ T cells	0.8 (0.7-1.2)	0.2 (0.2-0.3)
% Vα2.3+ CD4+ BAL T cells that express Vβ22	5.4 (3.5-6.2)	3.8 (2.7-9.1)
% Vβ22+ CD4+ BAL T cells that express Vα2.3	24.1 (18.0-30.1)	7.6 (6.1-11.9)

All percentage values are denoted as median (p25-p75).

Table S3. Results of TCR sequencing for mRNA extracted from FACS-sorted Vα2.3+ Vβ22- T cells (left column) from patients 12, 13, 2 and 3 (same as in Table 3) and FACS-sorted Vα2.3-Vβ22+ T cells (right column), respectively, from patients 2 and 3.

		Vα2.3				**Vβ22**			
Sample ID	HLA-type	a.a. sequence	Freq. (%)	TRAJ	CDR3 length	a.a. sequence	Freq. (%)	TRBJ	CDR3 length
Patient 12	DRB1*03,04	**CVVNMAGNQFYF**	100	49*01 F	10				
Patient 13	DRB1*03,04	CVVNMAGGSQGNLIF	100	42*01 F	13				
Patient 2	DRB1*03,01	CVVTRYGGSQGNLIF	35.31	42*01 F	13	CASSGGTSGVSYNEQFF	23.32	2-1*01	15
		CVVNKAGGSYIPTF	64.69	6*01 F	12	CASSETVAGGAQFF	27.68	2-1*01	12
						CASSESIGSGNTIYF	48.34	1-3*01	13
Patient 3	DRB1*03,13	CVVNMVGGGSNYKLTF	13.33	53*01 F	14	CARGGSRDEQFF	57.03	2-1*01	10
		CVVNPGTGNQFYF	19.8	49*01 F	11	CASSRAPGTGPRETQYF	40.7	2-5*01	15
		CVVNGANAGKSTF	25.95	27*01 F	11				
		CVVTHNNARLMF	40.91	31*01 F	10				

a.a. refers to amino acid sequence.

Freq. (%) refers to the percentage of reads mapped to every given transcript (isoforms with > 10% frequency).

In cases where the total percentage is < 100%, the remaining sequences were either too short or did not align.

The designations **TRAJ** and **TRBJ** follow the IMGT TCR gene nomenclature.

Length of CDR3 region as derived from the IMGT database.

Vα2.3^{+} sequences (highlighted in grey/bold font) share the 49*01 J segment and identical or near-identical amino acid sequences.

Table S4. Amino acid and nucleotide sequences for Va2.3^{+} and Vb22^{+} chains of patients 1-7.

A

	Vα2.3	
Sample ID	a.a. sequence	Nucleotide sequence
Patient 1	CVVNTPGNTPLVF	TGT GTG AAC ACC CCA GGA AAC ACA CCT CTT GTC TTT
	CVVNMGNTGGFKTIF	TGT GTG GTG AAC ATG GGG AAT ACT GGA GGC TTC AAA ACT ATC TTT
Patient 2	CVVNIGYGNKLVF	TGT GTG GTG AAC ATC GGA TAT GGA AAC AAA CTG GTC TTT
	CVVSVQGAQKLVF	TGT GTG GTG AGC GTT CAG GGA GCC CAG AAG CTG GTA TTT
	CVVNGLNIGDSGGGADGLTF	TGT GTG GTG AAC GGT CTT AAT ATA GGC GAT TCA GGA GGA GGT GCT GAC GGA CTC ACC TTT
Patient 3	CVVNNYKLSF	TGT GTG GTG AAC AAC TAC AAG CTC AGC TTT
Patient 4	CVVNMGRGGSNYKLTF	TGT GTG GTG AAC ATG GGG CGT GGA GGT AGC AAC TAT AAA CTG ACA TTT
	CVVGINNRKLIW	TGT GTG GTG GGG ATC AAC AAC CGT AAG CTG ATT TGG
	CVVNVRPGNTPLVF	TGT GTG GTG AAC GTA CGA CCA GGA AAC ACA CCT CTT GTC TTT
Patient 5	**CVVNLAGNQFYF**	TGT GTG GTG AAC CTA GCC GGT AAC CAG TTC TAT TTT
	CVVNPLGGGSYIPTF	TGT GTG GTG AAC CCT TTA GGG GGA GGA AGC TAC ATA CCT ACA TTT
Patient 6	CVVKEGSYIPTF	TGT GTG GTG AAA GAA GGA AGC TAC ATA CCT ACA TTT
	CAVKSGNNRLAF	TGT GCC GTG AAA AGC GGG AAC AAC AGA CTC GCT TTT
	CVVNMEYGNKLVF	TGT GTG GTG AAC ATG GAA TAT GGA AAC AAA CTG GTC TTT
Patient 7	CVVIGSGGSQGNLIF	TGT GTG GTG ATA GGA AGT GGA GGA AGC CAA GGA AAT CTC ATC TTT
	CVVNLAGNQFYF	TGT GTG GTG AAC CTT GCC GGT AAC CAG TTC TAT TTT

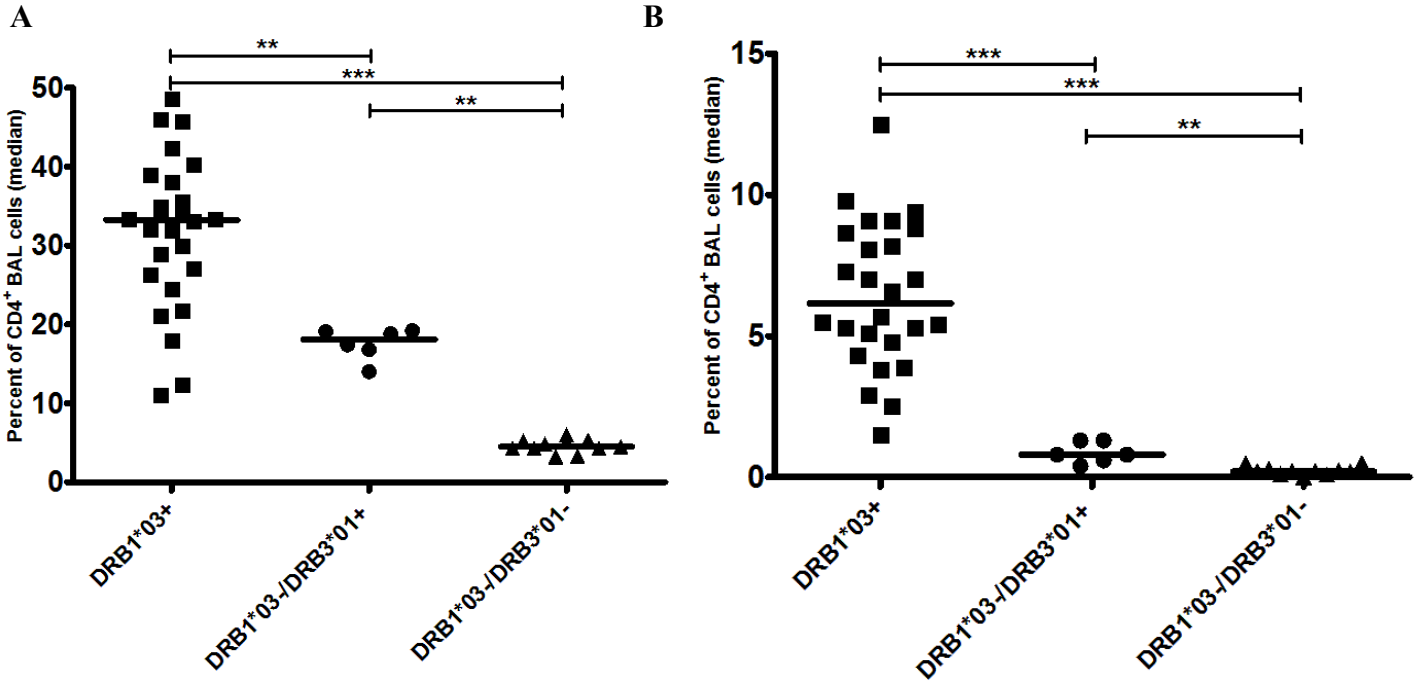
Note the different nucleotide sequences coding for the same amino acid (L) at the 5th amino acid position of the Vα2.3 chains in patients 5 and 7.

B

	Vβ22	
Sample ID	a.a. sequence	Nucleotide sequence
Patient 1	CASSEQGRGETQYF	TGT GCC AGC AGT GAA CAG GGG CGC GGG GAG ACC CAG TAC TTC
	CASSGTSVSTGELFF	TGT GCC AGC AGT GGG ACT AGC GTT TCC ACC GGG GAG CTG TTT TTT
Patient 2	CASSGPGGRTEAFF	TGT GCC AGC AGT GGT CCA GGG GGG AGA ACC GAA GCT TTC TTT
	CASSEMTRVVFHF	TGT GCC AGC AGT GAA ATG ACT CGG GTG GTC TTC CAC TTT
	CASSVITSGELFF	TGT GCC AGC AGT GTG ATC ACC TCC GGG GAG CTG TTT TTT
Patient 3	CASSGTGGAGTEAFF	TGT GCC AGC AGT GGC ACA GGG GGC GCC GGC ACT GAA GCT TTC TTT
	CASSEDVGRGAAFF	TGT GCC AGC AGT GAA GAC GTC GGT CGG GGG GCA GCT TTC TTT
	CASSGGFEQYF	TGT GCC AGC AGT GGG GGG TTC GAG CAG TAC TTC
Patient 4	CASSGGHGKGEQFF	TGT GCC AGC AGT GGC GGA CAC GGA AAG GGT GAG CAG TTC TTC
	CASSGAGGRGNEQFF	TGT GCC AGC AGT GGG GCA GGG GGC AGA GGC AAT GAG CAG TTC TTC
Patient 5	CASSVSTDTQYF	TGT GCC AGC AGT GTG AGC AGA GAT ACG CAG TAT TTT
	CASSEFGQGGHEQYF	TGT GCC AGC AGT GAG TTC GGA CAG GGG GGC CAC GAG CAG TAC TTC
Patient 6	CASSIDRSVGEKLFF	TGT GCC AGC AGT ATC GAC AGG AGT GTT GGT GAA AAA CTG TTT TTT
	CASSGTARNYGYTF	TGT GCC AGC AGT GGT ACG GCA AGG AAC TAT GGC TAC ACC TTC
Patient 7	CASSAITSNEKLFF	TGT GCC AGC AGT GCA ATT ACA TGT AAT GAA AAA CTG TTT TTT
	CASSAGSGQPQHF	TGT GCC AGC AGT GCA GGG TCG GGC CAG CCC CAG CAT TTT
	CASRPTSGRSDEQFF	TGT GCC AGC AGA CCA ACT AGC GGG CGT TCG GAT GAG CAG TTC TTC
	CASSVLGTAAVTF	TGT GCC AGC AGT GTT CTA GGG ACC GCG GCT GTA ACT TTC

Figure S1. Moderate Vα2.3^{+}Vβ22^{+} BAL T cell expansions in HLA-DRB1*03^{-}DRB3*01^{+} patients.

(A-B) Relative numbers of BAL T cells expressing TCR Vα2.3 (A) or TCR Vα2.3/Vβ22 (B) in DRB1*03^{+} (*n* = 26), DRB1*03^{-}/DRB3*01^{+} (*n* = 6), and DRB1*03^{-}/DRB3*01^{-} patients (*n* = 11), respectively.



(A) *** $p < 0.0001$, ** $p = 0.0048$, ** $p = 0.0032$

(B) *** $p < 0.0001$, *** $p = 0.0008$, ** $p = 0.0073$

Figure S2. Representative gating strategy for flow cytometry analysis of the percentage of Vα2.3^{+} CD4^{+} T cells expressing Vβ22, and vice versa.

Populations are expressed as percent of CD3^{+} CD4^{+} Vα2.3^{+} or Vβ22^{+} T cells in BAL fluid.

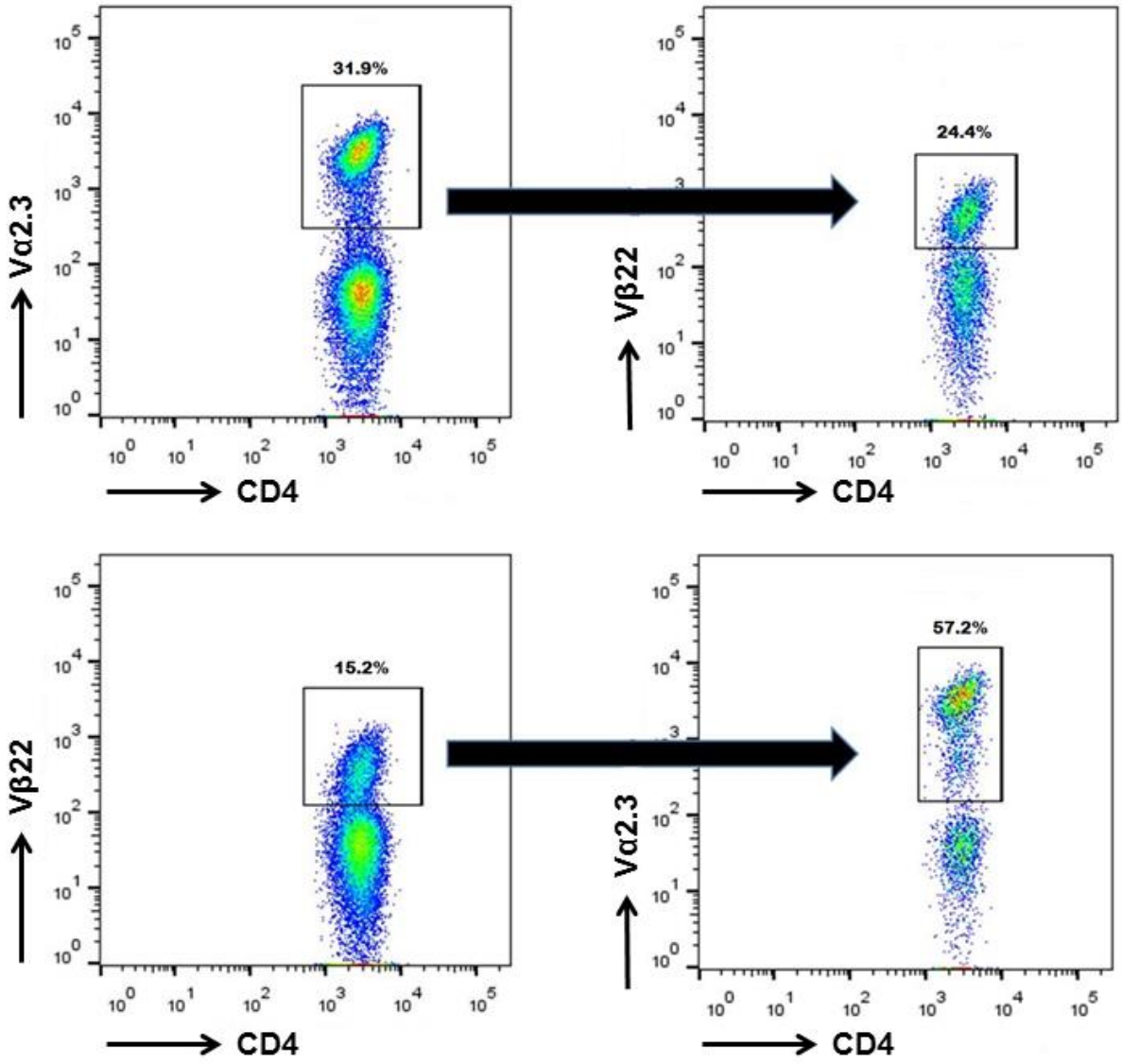


Figure S3. Amino acid frequencies per TCR β chain CDR3 position.

Amino acid frequencies are summarised for TCR β chain CDR3 sequences from IMGT reference and sarcoidosis groups (patients 1-7) and colour-coded by amino acid property. At position 112, the amino acid arginine occurs more frequently than expected in the sarcoidosis group in comparison with the reference group (p = 0.0003). Figures were generated with the aid of IMGT Junction Analysis Tool (S3, S7). For the reference group, positions between 111 and 112 have been omitted in the figure.