SUPPLEMENTARY

Quantitative tracking of T cell clones after hematopoietic stem cell transplantation

Mamedov I.Z.[1]*, Britanova O.V.[1]*, Bolotin D.A. [1]*, Chkalina A.V.[1], Staroverov D.B.[1], Zvyagin I.V.[1], Kotlobay A.A.[1], Turchaninova M.A.[1], Fedorenko D.A.[2], Novik, A. A.[2], Sharonov G. V.[3], Lukyanov S.[1], Chudakov D.M. [1], Lebedev Y.B.[1]**

[1]Shemiakin-Ovchinnikov Institute of Bioorganic Chemistry, RAS, Miklukho-Maklaya 16/10, 117997, Moscow, Russia

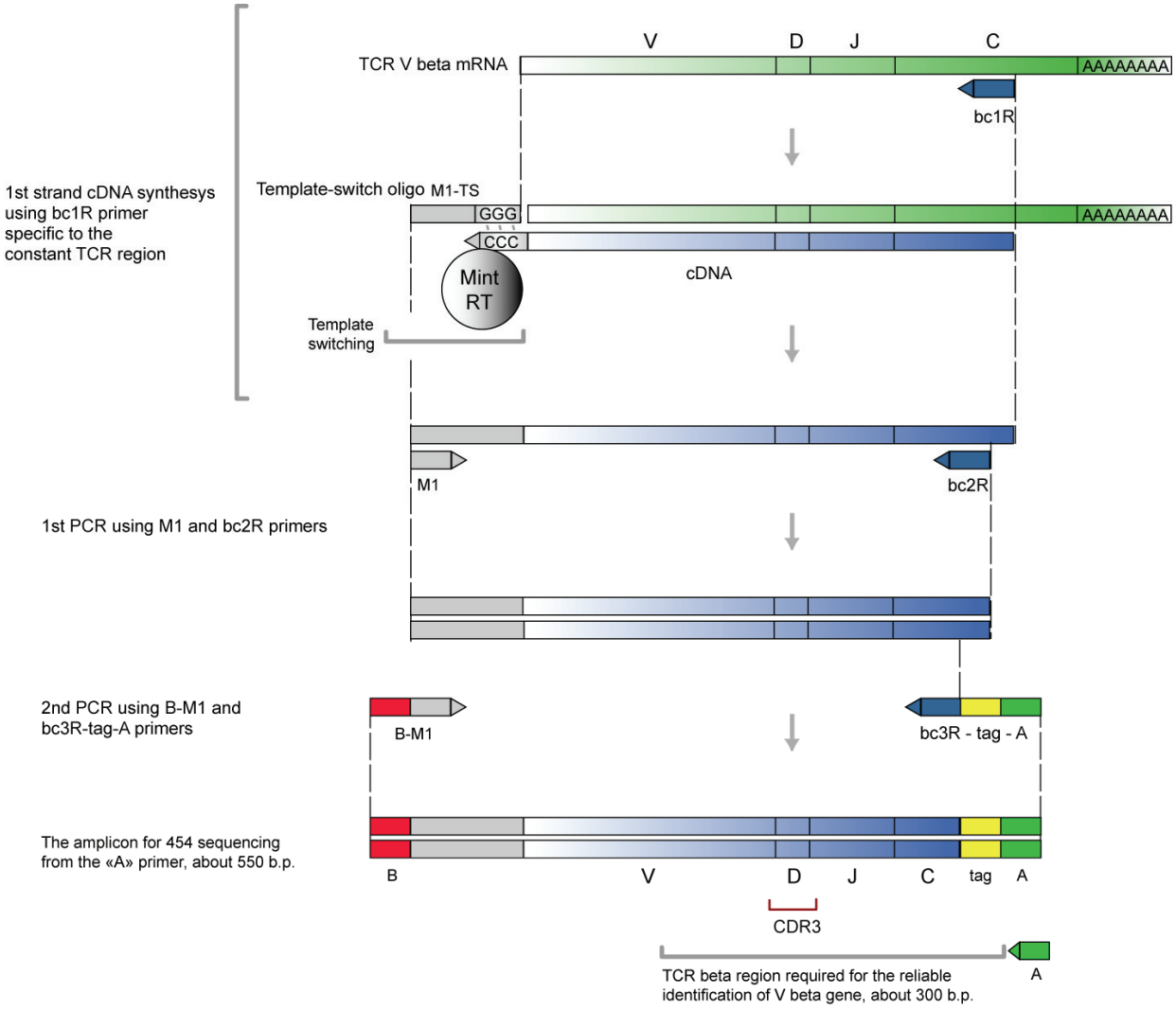
[2]Pirogov National Medical Surgical Center, Department of Haematology and Cellular Therapy, Moscow, Russia

[3]Faculty of Medicine, Lomonosov Moscow State University, Lomonosovsky Ave. 31/5, 119192, Moscow, Russia.

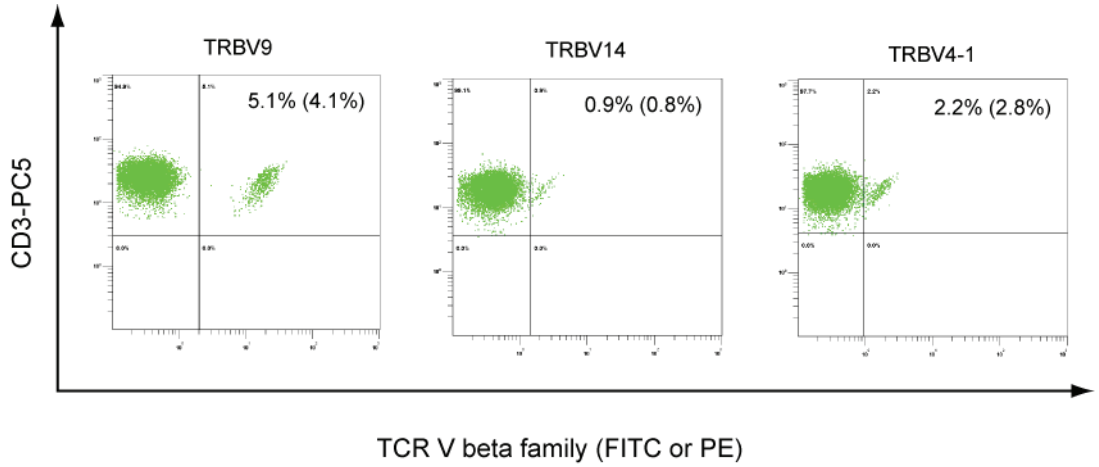
*These authors contributed equally to this work.

**Corresponding author. E-mail: ChudakovDM@mail.ru

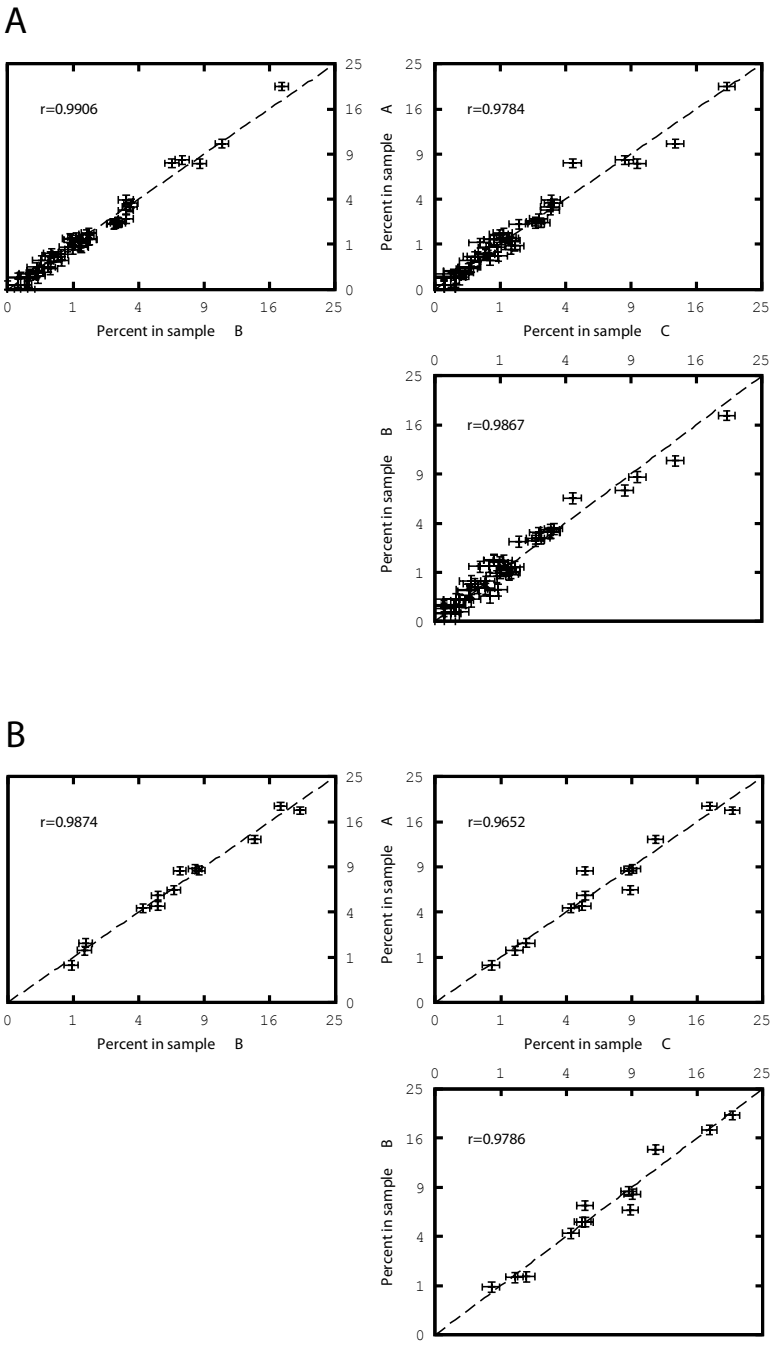
Supplementary Figure 1.	**Outline of the amplification technique.**
Supplementary Figure 2.	**Flow cytometric analysis using TCR V beta gene family-specific antibodies.**
Supplementary Figure 3.	**Reproducibility of the mass sequencing approach measured as TCR V beta and TCR J beta gene segments usage in three independent samples.**
Supplementary Figure 4.	**Block diagram of TCRbase modules.**
Supplementary Figure 5.	**Screenshot of TCRbase Web Client**
Supplementary Figure 6.	**Relative TCR beta diversity before, 4 months after, and 10 months after HSCT**
Supplementary Figure 7.	**Changes in TCR V beta gene usage after HSCT**
Supplementary Figure 8.	**Percentage of naive T cells observed before and after HSCT**
Supplementary Figure 9.	**Flow cytometric analysis using CMV-specific MHC tetramer**
Supplementary Table 1.	**Relative abundance of T cell clones determined using TCR V beta family-nested real-time PCR or total mass sequencing**
Supplementary Table 2.	**Key oligonucleotides used for cDNA synthesis and PCR amplification**
Supplementary Table 3.	**Fate of selected T cell clones after HSCT**
Supplementary Table 4.	**Fate of the group of homologous clones of TCR V beta gene family 20-1**
Supplementary Notes.	**Optimization of TCR beta cDNA amplification technique** **Checking for potential bias between and within TCR V beta genes** **Development of software for deep analysis** **Software algorithms** **HSCT and patient details**
Supplementary Data 1.	**List of the clustered clonal sequences for the samples before, 4 months after, and 10 months after HSCT.**
Supplementary Data 2.	**100% identical TCR beta CDR3 variants referenced in the NCBI protein database.**



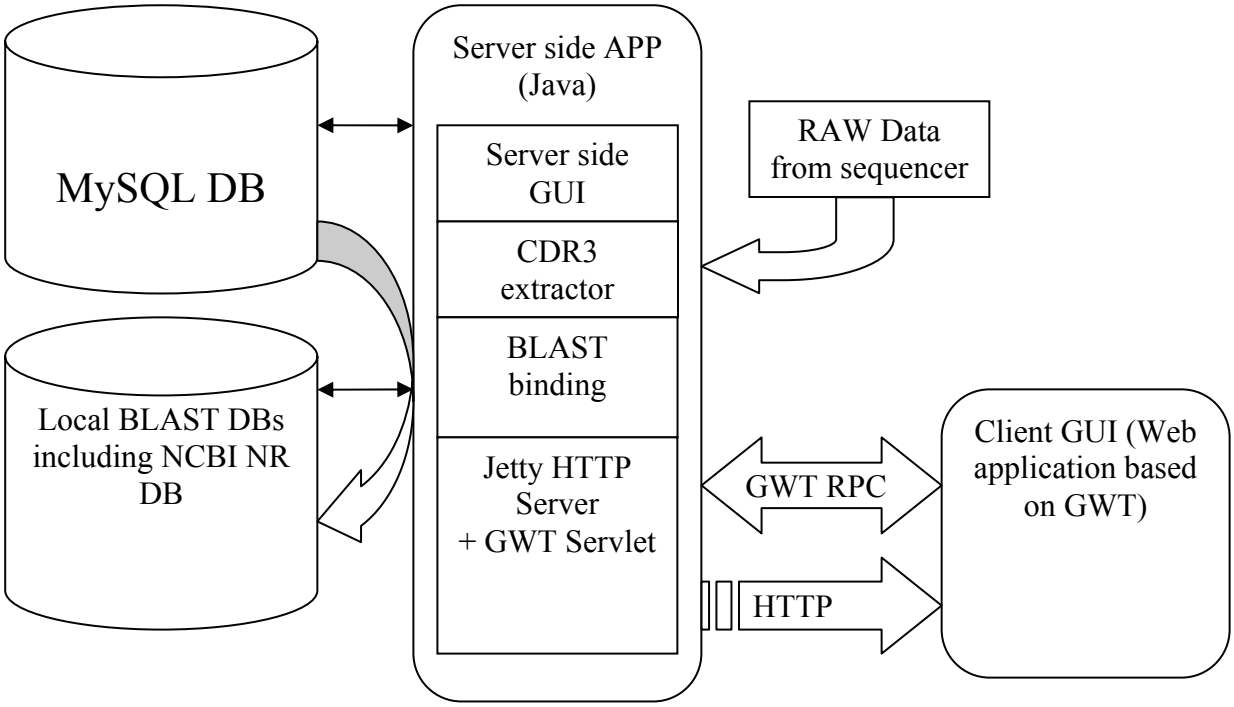
Supplementary Figure 1. Outline of the amplification technique.



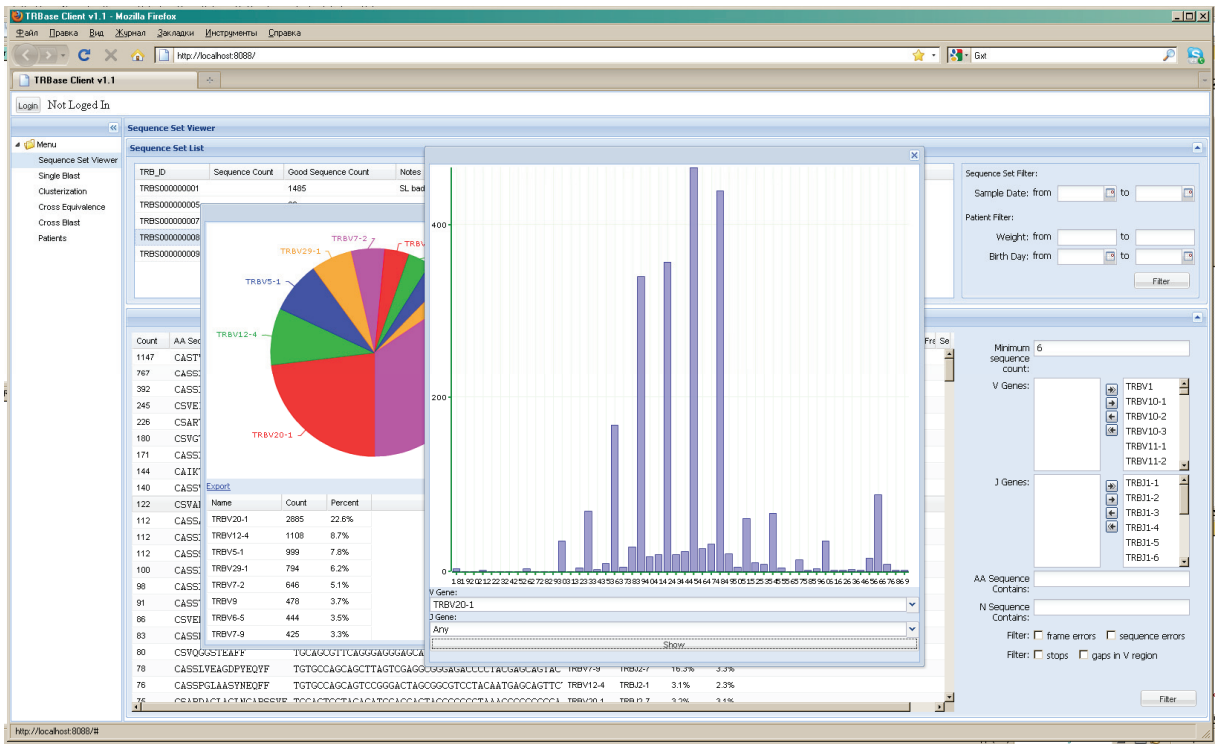
Supplementary Figure 2. Flow cytometric analysis using TCR V beta gene family-specific antibodies. Gated CD3+ T cells are shown. Numbers shown are percentage of T cells that correspond to a specific TCR V beta family member, based on flow cytomeric and mass sequencing results (mass sequencing results in parentheses).



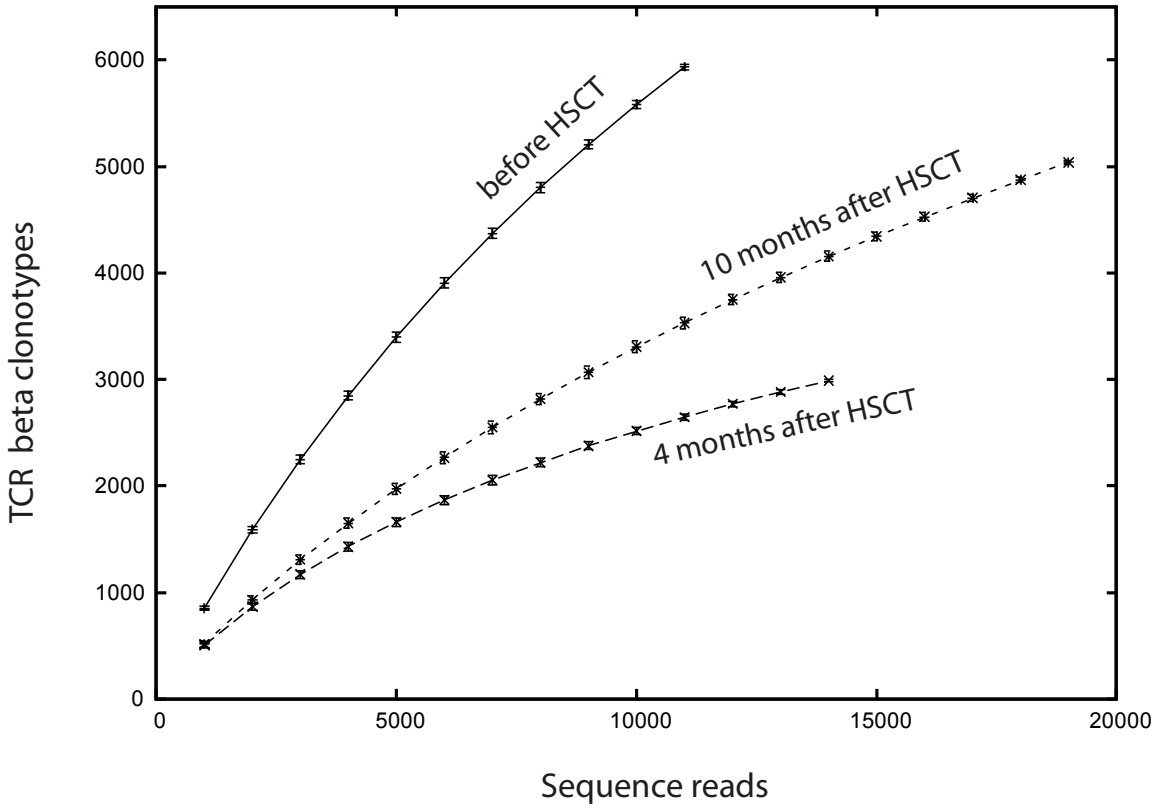
Supplementary Figure 3. Reproducibility of the mass sequencing approach measured as TCR V beta and TCR J beta gene segments usage in three independent samples. Graphs show correlation of V beta (A) and J beta (B) gene segments frequencies. Data presented with 95% confidence interval calculated from beta posterior distribution of binomial distribution parameter (Gelman et al, 1997). "r" is correlation coefficient for corresponding samples pair.



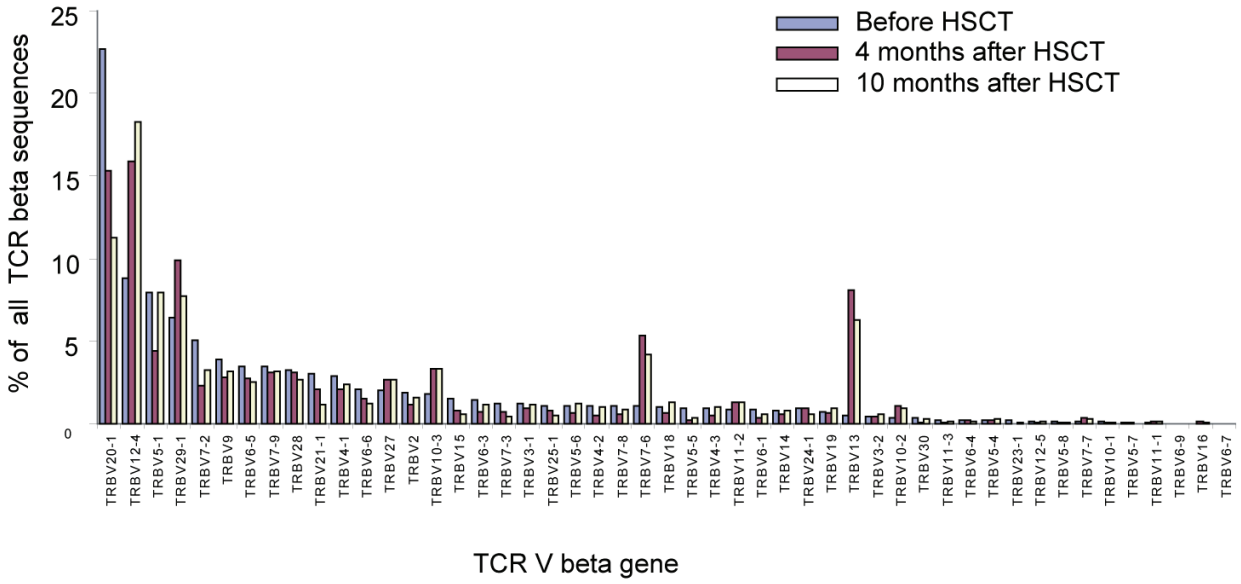
Supplementary Figure 4. **Block diagram of TCRbase modules.**



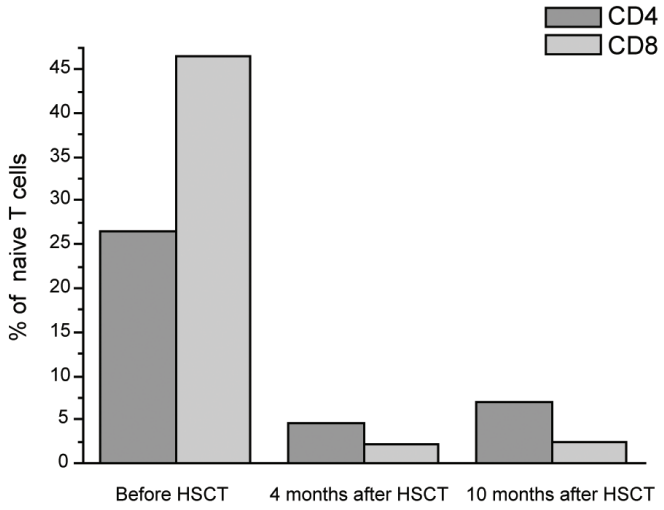
Supplementary Figure 5. Screenshot of TCRbase Web Client.



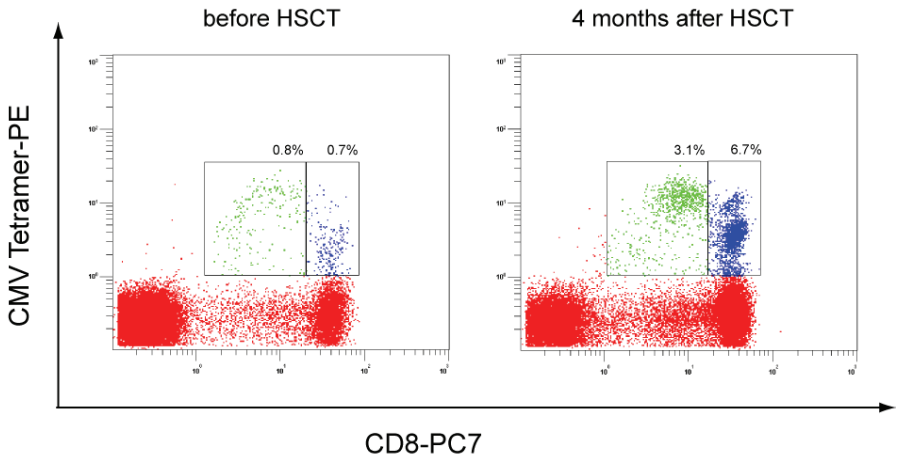
Supplementary Figure 6. Relative TCR beta diversity before, 4 months after, and 10 months after HSCT. To compare TCR beta diversity in different data sets we plotted diversity curve for each of them. The diversity curve is the dependence of number of unique clones on the size of random subset taken form the given data set. The number of clonotypes is plotted as a function of the number of reads, i.e. each point with error bar on the plot represents mean and standard deviation of the number of unique clones in corresponding data subset.



Supplementary Figure 7. **Changes in TCR V beta gene usage after HSCT.** Relative abundance of TCR V beta genes (IMGT gene nomenclature) is shown before, 4 months after and 10 months after HSCT. TCR V beta genes are ranged according to their relative abundance before HSCT.



Supplementary Figure 8. Percentage of naive T cells observed before and after HSCT. Naive T cells were identified using flow cytometry analysis as $CD45RA^{+}/CD27^{high}$ subpopulations of $CD4^{+}$ or $CD8^{+}$ T cells. Percentages of $CD45RA^{+}/CD27^{high}/CD4^{+}$ cells of all $CD4^{+}$ T cells and percentage of $CD8^{+}/CD45RA^{+}/CD27^{high}$ cells of all $CD8^{+}$ T cells are shown.



Supplementary Figure 9. Flow cytometric analysis using CMV-specific MHC tetramer. Numbers indicate percentage of PBLs that are NLVPMVATV-specific $CD8^+$ T cells (blue). Interestingly, NLVPMVATV-specific $CD8^{low}$ cells, which are likely NKT cells (displayed in green), are also abundant. Note the extremely high percentage of NLVPMVATV-specific cells 4 months after HSCT.

Supplementary Table 1. Relative abundance of T cell clones determined using TCR V beta family-nested real-time PCR or total mass sequencing.

T cell clone (beta chain CDR3)	% within a corresponding TCR V beta family	
	Nested real-time PCR	Total pyrosequencing
CASSVALGLNYEQYF	56.5 +/- 4.0	50.3
CSARGDTGTGYEQYF	3.14 +/- 0.7	3.1
CSVRGSEDTQYF	3.9+/-1.4	2.6
CASSPTTGTIANYGYTF	3.02+/-1.0	3.4
CASRWGSRADTQYF	6.35+/-1.6	6.6
CASSQEDRGTLYGYTF	5.02+/-0.97	5.8
CASSRRNYGYTF	6.48+/-0.58	6.4

Supplementary Table 2. Key oligonucleotides used for cDNA synthesis and PCR amplification.

Name	Sequence	Application
BC1R	CAGTATCTGGAGTCATTGA	1st strand synthesis
BC2R	TGCTTCTGATGGCTCAAACAC	1st PCR
M1	AAGCAGTGGTATCAACGCAGAGT	1st PCR
B-M1	GCCTTGCCAGCCCGCTCAGAAGCAGTGGTATCAACGCAGAGT	2nd PCR
A-MID1-BC3R	GCCTCCCTCGCGCCATCAGACGAGTGCGTCGACCTCGGGTGGGAACA	2nd PCR
A-MID2-BC3R	GCCTCCCTCGCGCCATCAGACGCTCGACACGACCTCGGGTGGGAACA	2nd PCR
B-MID3-BC3R	GCCTTGCCAGCCCGCTCAGAGACGCACTCCGACCTCGGGTGGGAACA	2nd PCR
B-MID5-BC3R	GCCTTGCCAGCCCGCTCAGATCAGACACGCGACCTCGGGTGGGAACA	2nd PCR
B-MID6-BC3R	GCCTTGCCAGCCCGCTCAGATATCGCGAGCGACCTCGGGTGGGAACA	2nd PCR

Supplementary Table 3. Fate of selected T cell clones after HSCT. Status of clones in blood samples before, 4 months after, and 10 months after HSCT is shown: present: <0.1% of all sequences (**+**); abundant: 0.1%<X<1% (**+ +**); hyper-expanded: >1% (**+ + +**), absent (**–**). Stably expanded clones tracked in the same patient during the 3 years before HSCT (Mamedov et al, 2009) are shown bold. CMV-specific clones CASSLAPGATNEKLFF-1 and CASSLAPGATNEKLFF-2 identified by MHC tetramer assay are shown italic.

T cell clone (beta chain CDR3)	Status before HSCT	Status 4 months after HSCT	Status 10 months after HSCT
CASSLSGGAGELFF	+ + +	+ + +	+ + +
CASSVALGLNYEQYF	+ + +	+ +	+ +
CASSLAPGATNEKLFF-1	+ +	+ + +	+ + +
CASSLAPGATNEKLFF-2	+	+ +	–
CSARTTYGTDIISQHF	+ +	+ + +	+ + +
CSARGDTGTGYEQYF	+ +	+ +	+ +
CASSPGLAASYNEQFF	+ +	+ +	+ +
CASSFKGADTQYF	+ +	+ +	+ +
CSARVVPGVQETQYF	+ +	+ +	–
CASRWGSRADTQYF	+ +	+	+ +
CSARADSGNGYEQYF	+ +	+	+
CSARADRGEGYEQYF	+ +	+	+
CSVRGSEDTQYF	+ +	+	+
CASSPTTGTIANYGYTF	+ +	+	–
CSARGDRGHGYEQYF	+ +	–	+
CSVVQNYEQYF	+ +	–	–
CASSIGDGAQYF	+ +	–	–
CASSLGTDSYEQYF	+ +	–	–
CSAGEFVGPLNQPQHF	+ +	–	–
CASSQLLAGTDTQYF	+ +	–	–
CSASSASGQPNYGYTF	+ +	–	–
CSASLAGGPGQYF	+ +	–	–
CASTVDSLDTEAFF	+	+ + +	+ + +
CASSLGENIQYF	+	+ + +	+ + +
CSVEIWDSSYNEQFF	+	+ + +	+ + +
CSVGTSEAYEQYF	+	+ + +	+ + +
CAIKTTSGIVDEQFF	+	+ + +	+ + +
CASSLWEGREQFF	–	+ + +	+ +
CASSRRNYGYTF	+	+	+
CSARGDRGRGYEQYF	+	+	–
CASSQEDRGTLYGYTF	+	+	–

Supplementary Table 4. Fate of the group of homologous clones of TCR V beta gene family 20-1. Identical amino acid residues are shown grey. Stably expanded clones tracked in the same patient during the 3 years before HSCT (Mamedov et al, 2009) are shown bold. After transplantation, the expansion of some of these homologous clones was suppressed, while four new highly homologous variants were identified. This suggests that T cell clones that have common antigen specificities can be differentially reprogrammed, and that new clones can be recruited after HSCT in response to the persistence of certain antigens.

CDR3 (V beta 20-1, J beta 2-7)	**Number of sequences (% of total)**		
	Sample taken 1 week before HSCT	**Sample taken 4 months after HSCT**	**Sample taken 10 months after HSCT**
CSARGDRGH**GYEQYF**	29 (0.21%)	0	4 (0.01%)
CSARA**DRG**E**GYEQYF**	25 (0.19%)	12 (0.08%)	8 (0.03%)
CSARGDRGEGYEQYF	11 (0.09%)	17 (0.11%)	20 (0.08%)
CSARGDRGR**GYEQYF**	2 (0.02%)	4 (0.03%)	0
CSARGDRGQGYEQYF-1	0	7 (0.04%)	13 (0.05%)
CSARGDRGQGYEQYF-2	0	5 (0.03%)	0
CSARGDRGFGYEQYF	0	4 (0.03%)	6 (0.02%)
CSARADRGDGYEQYF	0	4 (0.03%)	0

Supplementary Notes

Optimization of TCR beta cDNA amplification technique

The hypervariable complementary determining region 3 (CDR3) of the TCR beta chain is largely responsible for determining T cell specificity. The nucleotide sequence of CDR3 is unique for each T cell clone and thus can be used to track it and to quantify its abundance in human blood (Rufer, 2005).

The first step necessary for the individual in-depth TCR repertoire analysis is selective amplification of the TCR beta fragment starting either from the DNA or mRNA template obtained from a blood sample. The mRNA template is preferable since a significant portion of T cells carry a second (rearranged but nonfunctional) TCR beta gene (Bhalla et al, 2009; Wang et al, 2002), and such genes cannot be excluded from a DNA-based library. At the same time, nonfunctional TCR beta mRNA transcripts that carry premature termination codons are strongly downregulated by the nonsense-mediated mRNA decay mechanism (Bhalla et al, 2009; Wang et al, 2002). Therefore, an mRNA-based library will contain only a small number of non-functional rearranged TCR beta genes.

For quantitative analysis, unbiased and highly representative TCR beta amplification is crucial at this first step, without distorting the natural level TCR clonal sequence abundances that closely reflect the abundances of the corresponding T cell clones. To maintain natural abundance levels, primers that are universal for all possible TCR beta chains are required, since multiplex PCR, as proposed by others (Robins et al, 2009), leads to an inevitably high bias towards specific TCR V beta gene segments. Indeed, in multiplex PCR there is always preferential amplification of one target sequence over another, while many target sequences may not be amplified at all (Elnifro et al, 2000; Markoulatos et al, 2002).

Universal primers can be designed for the constant region of the TCR beta chain located downstream of the CDR3 region, in which one out of two highly homologous TCR C beta gene segments is presented (this also requires mRNA template, due to the introns located between the J and C segments). However, each functional TCR beta chain comprises a V beta gene segment that has been randomly selected from more than 50 different genes (http://imgt.cines.fr/). Therefore, it is impossible to design a universal primer located upstream of TCR V beta CDR3.

Lastly, for large-scale analysis, it is critical to maximize the specificity and efficiency of amplification in order to minimize the number of PCR amplification cycles, thereby reducing accumulated PCR errors and amplification of non-target sequences.

To address these problems, we have developed and implemented several technical solutions (**Supplementary Fig. 1**).

First, we employed gene-specific priming for TCR beta cDNA synthesis. For this, we have designed a set of TCR C beta-specific primers and selected one that provided the most specific and efficient first strand synthesis (data not shown). The use of this optimized specific primer for cDNA synthesis was absolutely necessary for further successful amplification. Other methods of cDNA priming, such as oligo-dT or random hexamers, led to a dramatic loss in PCR specificity and did not allow for the TCR beta amplification within a reasonable number of PCR cycles.

Second, we exploited the reverse transcriptase template switching effect (Douek et al, 2002; Matz et al, 1999) to generate a universal primer at the 5'end of TCR beta chain. Thus, the efficiency of further amplification was independent on the sequence of V beta gene segment, since all TCR beta sequences were amplified using universal primer pairs (see verification procedures below).

Third, two nested primers that bound the constant region of the TCR beta chain gene were sequentially used to further increase amplification specificity.

Fourth, to suppress the potential products of non-specific annealing of the switching primer, we used step-out PCR (Matz et al, 1999) accompanied by the PCR-suppression effect (Siebert et al, 1995).

Finally, to overcome potential inefficiency of DNA adapter-ligation, oligonucleotides for subsequent sequencing and DNA barcodes enabling the analysis of multiple samples in the same sequencing run were introduced in the course of the last amplification step.

The combination of these solutions allowed us to specifically amplify the ready-for-sequencing, rearranged TCR beta gene fragment within 21-23 PCR cycles when starting with approximately 2 μg of total RNA obtained from fresh PBLs purified from a 2 ml blood sample.

Checking for potential bias between and within TCR V beta genes

To confirm that our approach of analyzing the TCR beta repertoire is unbiased with respect to the relative abundance of TCR V beta genes, we performed flow cytometric analysis of PBLs from a blood sample obtained before HSCT. We stained PBLs with TCR V beta family-specific antibodies and compared the percentage of corresponding T cells with those determined by pyrosequencing. Despite potentially variable TCR expression levels and incomplete knowledge with respect to the specificity of some antibody clones, flow cytometry data were quite consistent with data from the mass sequencing experiments (**Supplementary**

Fig. 2), indicating that the sequenced repertoire represents an unbiased quantitative replica of the real T cell clonal diversity.

To verify the potential bias between the amplification of clonal sequences carrying the same TCR V beta genes, we have employed our earlier experience of tracking the stably-expanded T cell clones in the same patient using CDR3-specific real-time PCR (Chkalina et al, 2010; Mamedov et al, 2009). We compared the pyrosequencing data obtained for the total TCR beta amplicon with those obtained using CDR3-specific real-time nested-PCR from the amplicons of the corresponding TCR V beta genes. Importantly, we used the same pre-HSCT blood sample, but cDNA synthesis was performed independently and using different protocols: a specific primer was used for cDNA synthesis in pyrosequencing as described above, while the oligo-T primer was employed for cDNA synthesis in CDR3-specific real-time PCR experiments, according to the previously reported protocol (Chkalina et al, 2010; Mamedov et al, 2009). Data from these two experimental approaches were generally quite consistent with each other (**Supplementary Table 1**), indicating that the relative abundance of clones carrying the same TCR V beta gene was likely not distorted by TCR V beta-nested real-time PCR or by the total pyrosequencing method.

Finally, to verify the reproducibility of the whole approach, we have compared usage of TCR V beta and TCR J beta gene segments between the tree subsets of the same sequence run (blood sample obtained 10 months after HSCT), that were amplified independently from three cDNA samples and were differentiated by internal barcodes. The verification showed very good reproducibility (**Supplementary Fig. 3**).

Development of software for deep analysis

We developed a software program, named TCRbase, to process and analyze data obtained from the sequencer. This software was designed to perform two main tasks: (i) efficiently extract sets of CDR3 sequences from raw data obtained from the 454 sequencer, and (ii) subsequently analyze the extracted sets through the user-friendly interface.

First, TCRbase analyzes each sequence based on human TCR V beta and J beta gene sequences from the IMGT database (http://imgt.cines.fr/). To identify TCR V beta genes, TCRbase uses a BLAST algorithm that maximizes the extraction efficiency for valid sequences despite potential PCR or sequencing errors within the TCR V beta gene segment. The first nucleotide of CDR3, a conserved cysteine, is determined based on the alignment with the identified V beta gene. This process is repeated to identify the J beta gene and the last nucleotide of CDR3, a conserved TCR J beta phenylalanine in the amino acid motif FGXG.

In the next step, TCRbase clusters nucleotide sequences carrying identical CDR3, TCR V beta, and J beta (i.e. clonal TCR beta sequences). The relative amount of each clonal sequence corresponds to the actual abundance of the T cell clone in the patient's blood sample. Thus, each resulting group of clonal sequences is further referred to as a T cell clone, which is characterized by its TCR V beta and J beta genes, CDR3 nucleotide sequence, translated amino acid CDR3 sequence, and the number of such clonal sequences. TCRbase can optionally exclude out-of-frame CDR3 sequences and/or CDR3 sequences carrying stop codons from further analysis.

TCRbase can be used to perform several types of clustered data analysis. Specifically, BLAST searches can be performed to identify nucleotide or amino acid CDR3 sequences that are homologous or identical to the sequence of interest. Besides, the entire dataset can be analyzed by BLAST to search for sequence matches across the NCBI database. Groups of clones with particular TCR V beta and/or J beta gene segments can be extracted and used for both manual and BLAST searches. TCRbase can also generate *in silico* CDR3 length spectratypes, which are displayed as histograms showing the number of sequences sorted according to the CDR3 lengths for each group of clones sharing identical V beta or J beta gene segments, or both (see **Supplementary Fig. 5**). This simplifies analysis and allows for the comparison of data obtained through this method with results from a classic spectratyping approach (Rufer, 2005).

The key point in terms of HSCT impact analysis is the ability to track individual T cell clones that survived transplantation, as well as to identify new clones that have arisen. For this purpose, TCRbase can perform cross-identity and cross-BLAST analysis of multiple datasets. This enables identification of common sequences present before and after HSCT, as well as at multiple time points or in samples obtained from many patients. Both cross-identity and cross-BLAST tests can be used for nucleotide and amino acid sequence comparison to identify T cell clones with the same antigen specificity (i.e., clones with the same or very similar amino acid CDR3 sequences) but of a different origin (i.e., clones with different nucleotide CDR3 sequences).

Non-commercial and academic users are granted free access to the TCRbase implementation described in this paper. Source code is available at no charge: http://sourceforge.net/projects/tcrbase/files/v0.1/ .

Software algorithms

Input data

Input data to analysis presented as a single file with sequences in FASTA format.

Reference sequences of human V beta and J beta

Reference sequences of V beta and J beta genes were obtained from IMGT/GENE-DB database. TRBV with IMGT gaps and TRBJ sequences were downloaded from (<http://www.imgt.org/IMGT_GENE-DB/GENElect?query=7.1+TRBV&species=Homo+sapiens>) and (<http://www.imgt.org/IMGT_GENE-DB/GENElect?query=7.5+TRBJ&species=Homo+sapiens>), respectively. TRBV sequences with IMGT gaps were used to determine the position of the first nucleotide of the codon that codes for 2nd-CYS (according to the IMGT definition (<http://www.imgt.org/textes/IMGTScientificChart/Nomenclature/IMGT-FRCDRdefinition.html>), CDR3 begins immediately following the second cysteine [2nd-CYS] in the 104th amino acid sequence position of the T-cell receptor beta chain). IMGT inserts additional gaps (marked with dots) to shift the 2nd-CYS to this position. Because the V gene is the first gene in a mature T cell receptor beta chain (V-D-J-C), the 2nd-CYS is also in position 104 in the amino acid sequence and in position 310 in the nucleotide sequence.

Example of TRBV reference sequence from IMGT:

```
gatgctgaaatcacccagagcccaagacacaagatcacagagacaggaaggcaggtgacc
ttggcgtgtcaccagacttggaaccac.....................aacaatatgttc
tggtatcgacaagacctgggacatgggctgaggctgatccattactcatatggt......
......gttcaagacactaacaaaggagaagtctca...gatggctacagtgtctctaga
tca...aacacagaggacctccccctcactctggagtctgctgcctcctcccagacatct
gtatatttctgcgccagcagtgagtc
|         |         |
300       310       320
```

The gaps were added to the J gene sequences to correctly position J-PHE/J-TRP, which are the amino acids after the end of CDR3 according to the IMGT definition (<http://www.imgt.org/textes/IMGTScientificChart/Nomenclature/IMGT-FRCDRdefinition.html>), similar to the V genes. Adding the gaps eased the subsequent loading

of this sequence into TCRbase. All orphans from chromosome 9 were excluded from the reference dataset.

Extraction of CDR3 sequences from raw data

Definition

Our definition of CDR3 is similar to the IMGT definition, except that we included 2nd-CYS and J-PHE/J-TRP into the sequence.

Overview

To extract the CDR3 sequence, we determined the 2^{nd}-CYS and J-PHE/J-TRP positions in the raw sequence. Identification of these positions and of the closest V and J genes is based on global pairwise alignment performed by stand-alone BLAST.

Why BLAST?

We use BLAST rather than IMGT/V-QUEST (Brochet et al, 2008), which uses the simplest pairwise alignment without insertions or deletions, because BLAST will identify sequences that have PCR- or sequence-based errors in the TCR V beta or J beta gene segments. Thus, using BLAST over the simple pairwise alignment algorithm increases the output of mass sequencing.

Algorithm

BLAST databases containing V beta and J beta nucleotide reference sequences were prepared using the *makeblastdb* utility from the stand-alone BLAST package.

The following steps are performed to analyze each sequence compared with the original data set:

1. Using the *blastn* program, the raw sequence is searched to identify matches in V beta database with an e-value threshold of 10^{-8}.
2. Using the *blastn* program, the raw sequence is searched to identify matches in the J beta database with an e-value threshold of 10^{-6}.
3. In the case where matches are found in both the above searches, we take best matches as V and J genes for the sequence.
4. The 2^{nd}-CYS and J-PHE/J-TRP positions are determined using the resulting alignments.
5. Error tests are performed (*see below*).

Error tests

To improve the quality of CDR3 sequence data, the software can perform optional filtering based on the following requirements:

1. The first nucleotides of 2nd-CYS and J-PHE/J-TRP codons must be in the same reading frame.
2. The alignment of the V and J genes must have no gaps or deletions in the CDR3 region.
3. The CDR3 sequence must not have stop codons.

Factorization of CDR3 sequences

The program then factorizes the set of analyzed sequences based on the data obtained by the equivalence of V beta and J beta genes types, and a CDR3 nucleotide sequence. In other words, the program categorizes the original set into subsets of equivalent sequences. We refer to the equivalence classes obtained as "clones", and characterize each with a V beta and J beta gene type, CDR3 nucleotide sequence, and power (number of sequences). The set of "clones" is the concentrated result of processing raw data at this stage.

Analysis

Cross-Identity

An important type of analysis is the search for identical "clones" in a number of different samples for their subsequent comparison. Thus, this would allow for analyzing dynamics of representation degrees of various "clones" in samples from the same patient under a specific treatment regimen. To do this, we identify clones that are equivalent in terms of CDR3 nucleotide sequence and V beta and J beta gene types in different factorized sequence sets and output them side by side in a table with information about copy count and relative contribution to the V beta and J beta families in corresponding sets.

Implementation

All algorithms and analyses were implemented in a Java™-based system called TCRbase (T-Cell Receptor database). The block diagram of TCRbase is shown in **Supplementary Fig. 4**.

Core (server side application)

The main part of TCRbase is a Java application. The application includes:

- Data object-relational mapping (ORM) based on the Hibernate 3.3.2 library, which allows for communication with the MySQL server.

- A module that performs extraction and factorization of CDR3 sequences in a fully automatic manner.
- To facilitate the above procedure, the core application has a binding module for the BLAST program, which allows for creating and indexing BLAST databases and running BLAST programs (like blastp, blastn, etc.), and the subsequent parsing of the results.
- A server-side GUI to monitor client connections and to manage datasets. For example, the GUI allows for the loading of new raw sequence sets and extracting and factorizing CDR3 sequences.
- A web server based on a Jetty 6.1.21 server to facilitate connection from remote clients.
- Analytical modules that perform different analyses of factorized data on request by a remote user.

All algorithms are optimized for multiprocessor systems using Java threading facilities.

Database server

All data in the system, including nucleotide and amino acid sequences, are stored on a MySQL 5.1.37 server.

Client web application

The main user interface (UI) for TCRbase is a web-based application called TRWeb. It was written using the Google Web Toolkit (GWT) 2.0.3 and Ext GWT (GXT) 2.1.1. TRWeb allows for the efficient manipulation of analytical tools and examination of factorized data and analysis results. A screenshot of UI is shown in **Supplementary Fig. 5**.

HSCT and patient details

We had the unique chance to track the fate of T cell clones for one of the first cases of autologous HSCT used to treat a patient with ankylosing spondylitis (AS). AS is an autoimmune disease characterized by chronic inflammation of the spine and the sacroiliac joints, strongly associated with the MHC class I allele HLA-B27 (Brewerton et al, 1973) and characterized by stable oligoclonal T cell expansions (Duchmann et al, 2001; Mamedov et al, 2009). Therefore, T cell clones apparently play an important role in AS pathogenesis. Although manifestations of AS can be quite severe, this disease is usually not life-threatening and HSCT has thus not yet become the therapy of choice for AS. However, this type of therapy of AS can

enter clinical practice in the coming years, and thus it was of particular interest to track the fate of T cell clones in one of the first AS patients undergoing HSCT.

The patient, a 46-year-old, HLA-B27-positive man, was diagnosed with AS in 1984. During 1995-2005, the disease activity increased and in 2005, the patient started to receive monthly infusions of Remicade (Infliximab). During the first year of anti-TNF-alpha therapy, the patient reported significant functional improvement after each administration of Remicade. However, the remission period shortened and there was a complete absence of relief following the last injection (March 2008). In June 2009, high-dose chemotherapy (HDCT) and autologous HSCT procedures were performed in accordance with the European Group for Bone Marrow Transplantation (EBMT) protocol. It is notable that complete remission was observed for at least 12 months following HSCT, the time of manuscript submission.

Samples of peripheral blood were collected 1 week before, 4 months after transplantation (at which time point the population of white blood cells expanded to pre-HSCT level), and 10 months after transplantation.

References:

<http://www.imgt.org/IMGT_GENE-DB/GENElect?query=7.1+TRBV&species=Homo+sapiens>.

<http://www.imgt.org/IMGT_GENE-DB/GENElect?query=7.5+TRBJ&species=Homo+sapiens>

<http://www.imgt.org/textes/IMGTScientificChart/Nomenclature/IMGT-FRCDRdefinition.html>

Bhalla AD, Gudikote JP, Wang J, Chan WK, Chang YF, Olivas OR, Wilkinson MF (2009) Nonsense codons trigger an RNA partitioning shift. J Biol Chem 284**:** 4062-4072

Brewerton DA, Hart FD, Nicholls A, Caffrey M, James DC, Sturrock RD (1973) Ankylosing spondylitis and HL-A 27. Lancet 1**:** 904-907

Brochet X, Lefranc MP, Giudicelli V (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res 36**:** W503-508

Chkalina AV, Zvyagin IV, Mamedov IZ, Britanova OV, Staroverov DB, Lebedev YB (2010) The Oligoclonal Expansion of T Cells: the Investigation of Its Stability over Time. Russian Journal of Bioorganic Chemistry 36**:** 191-198

Douek DC, Betts MR, Brenchley JM, Hill BJ, Ambrozak DR, Ngai KL, Karandikar NJ, Casazza JP, Koup RA (2002) A novel approach to the analysis of specificity, clonality, and

frequency of HIV-specific T cell responses reveals a potential mechanism for control of viral escape. J Immunol 168**:** 3099-3104

Duchmann R, Lambert C, May E, Hohler T, Marker-Hermann E (2001) CD4+ and CD8+ clonal T cell expansions indicate a role of antigens in ankylosing spondylitis; a study in HLA-B27+ monozygotic twins. Clin Exp Immunol 123**:** 315-322

Elnifro EM, Ashshi AM, Cooper RJ, Klapper PE (2000) Multiplex PCR: optimization and application in diagnostic virology. Clin Microbiol Rev 13**:** 559-570

Gelman A, Carlin JB, Stern HS, Rubin DB (1997) Bayesian Data Analysis. London: Chapman and Hall

Mamedov IZ, Britanova OV, Chkalina AV, Staroverov DB, Amosova AL, Mishin AS, Kurnikova MA, Zvyagin IV, Mutovina ZY, Gordeev AV et al (2009) Individual characterization of stably expanded T cell clones in ankylosing spondylitis patients. Autoimmunity 42**:** 525-536

Markoulatos P, Siafakas N, Moncany M (2002) Multiplex polymerase chain reaction: a practical approach. J Clin Lab Anal 16**:** 47-51

Matz M, Shagin D, Bogdanova E, Britanova O, Lukyanov S, Diatchenko L, Chenchik A (1999) Amplification of cDNA ends based on template-switching effect and step-out PCR. Nucleic Acids Res 27**:** 1558-1560

Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114**:** 4099-4107

Rufer N (2005) Molecular tracking of antigen-specific T-cell clones during immune responses. Curr Opin Immunol 17**:** 441-447

Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA, Lukyanov SA (1995) An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res 23**:** 1087-1088

Wang J, Vock VM, Li S, Olivas OR, Wilkinson MF (2002) A quality control pathway that down-regulates aberrant T-cell receptor (TCR) transcripts by a mechanism requiring UPF2 and translation. J Biol Chem 277**:** 18489-18493