



# TCR Profiling Bioinformatics Analysis Report

---

Service Type:	TCR Profiling Stage II (level 1)
Order Reference (Evrogen) No.	0000000
Order PO No.	0000000
Orderer Contact Name:	Dr. John Doe
Orderer Contact E-mail:	j.doe@aaaa.aa
Orderer Institution:	Aaaa
Input Filename:	example_dataset.fasta
Output Filename:	0000000_Report.xls
Report Date:	DD.MM.YYYY

Evrogen Lab  
Miklukho-Maklaya str, 16/10,  
117997, Moscow, Russia  
Tel: +7(495) 988 4086  
Fax: +7(495) 988 4085  
service@evrogen.com  
[www.evrogen.com](http://www.evrogen.com)

## Contents

I. Overview	3
II. Input Data Quality Statistics	3
III. Output Description	4
a. Sequences Table	4
b. Spectratypes	5
c. Genes Abundance	7
IV. TRBase algorithm at a glance	8

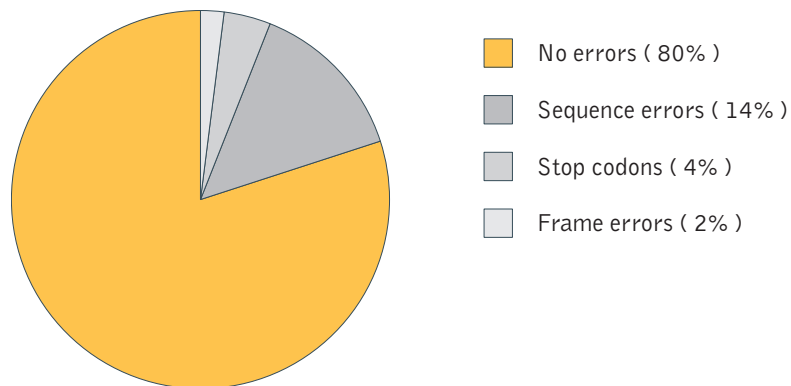
## I. Overview

Received input file `example_dataset.fasta` was analyzed by Evrogen's TRBase software. The following steps were performed:

1. Multistep sequence error removal (see [Input Data Quality Statistics](#))
2. CDR3 extraction from RAW FASTA data (see [Sequences Table](#)).
3. Referring each sequence read to a particular V and J family employing BLAST algorithm (see [Sequences Table](#)).
4. Clusterisation of individual clones (sequence reads with identical V, J and CDR3). The relative amount of each clonal sequence corresponds to the actual abundance of the T cell clone (see [Sequences Table](#)).
5. Virtual CDR3 length spectratyping for each of the V and J gene families (see [Spectratypes](#)).
6. Computing a relative abundance of V and J gene families (see [Genes Abundance](#)).

## II. Input Data Quality Statistics

Input data quality, for `example_dataset.fasta`, was rated overall as good. Several error tests were performed to filter final clone sets (see [TRBase algorithm at a glance](#)). Only clones without any errors underwent further analysis. However, sequences with detected errors were not discarded but marked with flags (see [Sequences Table](#)). The pie-chart below represents error type percentage:



### III. Output Description

TRBase output is located in the attached *CC129531\_Report.xls* file. Here you can find a brief description of available data.

#### a. Sequences Table

	A	B	C	D	E	F	G	H	I	J	K
1	TRB_ID	Count	AA CDR3 Sequence	N CDR3 Sequence	V Percent	V Gene	J Percent	J Gene	Has Stops	Has Sequence Errors	Has Frame Errors
2	TC00002426F	25	CASSAGLLGEQYF	TGTGCCAGCAGTGC GGGGCTTTGGGGGAGCAGTACTTC	5,97%	TRBV6-5	1,04%	TRBJ2-7			
3	TC000024270	4	CAIQMPAYEQYF	TGTGCCATTGATGTACCCAGCCTACGAGCAGTACTTC	0,78%	TRBV10-3	0,17%	TRBJ2-7			
4	TC000024271	14	CASSYRGGQGSYEQFF	TGTGCCAGCAGTTACTCAAGGGGACAGGGGCTCTATGAGCAGTTCTTC	3,34%	TRBV6-5	0,42%	TRBJ2-1			
5	TC000024272	2	CASRPGQKSMSSS	TGTGCCAGCAGACCCGGGACAGAAATCAATGAGCAGTTCTTC	0,91%	TRBV6-6	0,06%	TRBJ2-1			*
6	TC000024273	1	CASSYSGVLATVDTGELF	TGTGCCAGCAGTTACTCGGGGCTCTTGCACCGTGGACACCGGGGAGCTGTTTT	0,24%	TRBV6-5	0,09%	TRBJ2-2			*
7	TC000024274	49	CASSKARTNKQ*AVL	TGTGCCAGCAGCAAAGCAAGGACCAATAAACAATGAGCAGTTCTTC	14,37%	TRBV21-1	1,49%	TRBJ2-1	*		*
8	TC000024275	1	CSARALLSTEAF	TGCAGTGCTAGAGCTTTATTGAGCACTGAAGCTTTCTTT	0,04%	TRBV20-1	0,04%	TRBJ1-1			
9	TC000024276	8	CASSLDSSSYNEQFF	TGTGCTAGCAGCTTGGACTCGAGTTCTTACAATGAGCAGTTCTTC	17,78%	TRBV7-7	0,24%	TRBJ2-1			
10	TC000024277	2	CSVEGPTYNEQFF	TGCAGCGTTGAAGGGCCAACGTTGTACAATGAGCAGTTCTTC	0,13%	TRBV29-1	0,06%	TRBJ2-1			
11	TC000024278	1	CASSRPGAGTDTQYF	TGTGCCAGCAGCCGACCCGGGACGGGCACAGATACGCAAGTATTTT	0,23%	TRBV9	0,08%	TRBJ2-3			
12	TC000024279	1	CASRSLTGGGAETQYF	TGTGCCAGCAGGAGCCTCACTGGTGGGGGGCGGAGACCCAGTACTTC	0,20%	TRBV28	0,15%	TRBJ2-5			
13	TC00002427A	1	CSASASYNPLHF	TGCAGTGCTAGCGCCAGTTATAATTCAACCCCTCCACTTT	0,04%	TRBV20-1	0,51%	TRBJ1-6			
14	TC00002427B	2	CSAREGGTPGSCF	TGCAGTGCTAGAGAGGGGGAACACCGGGGAGCTGTTTTTTT	0,08%	TRBV20-1	0,18%	TRBJ2-2			*
15	TC00002427C	86	CSVEEWASRYNEQFF	TGCAGCGTTGAAGAGTGGGCTAGCAGATACAATGAGCAGTTCTTC	5,59%	TRBV29-1	2,61%	TRBJ2-1			
16	TC00002427D	1146	CASTVDSLDTAEFF	TGTGCCAGCACCGTGGACAGTCTGGACACTGAAGCTTTCTTT	46,08%	TRBV12-4	43,76%	TRBJ1-1			

Column A contains internal ID of the clone.

Column B contains sequence reads count – calculated number of equivalent sequences comprising the clone.

Column C contains amino-acid sequence of the clone derived from the nucleotide sequence.

Column D contains nucleotide sequence of the clone.

Column E contains percentage of the clone in the V gene family. IMGT gene nomenclature is used.

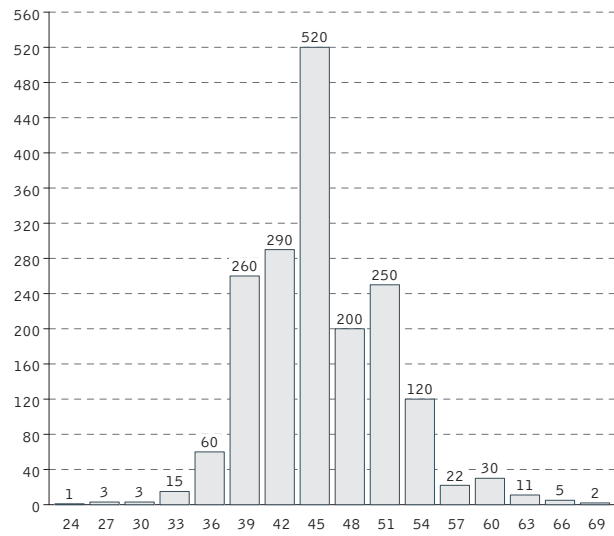
Column F contains V beta gene segment name identified for the clone.

Column G contains percentage of the clone in the J gene family. IMGT gene nomenclature is used.

Column H contains J gene family name identified for the clone.

Columns I-K contain flags that mark error type detected in the clone (if any).

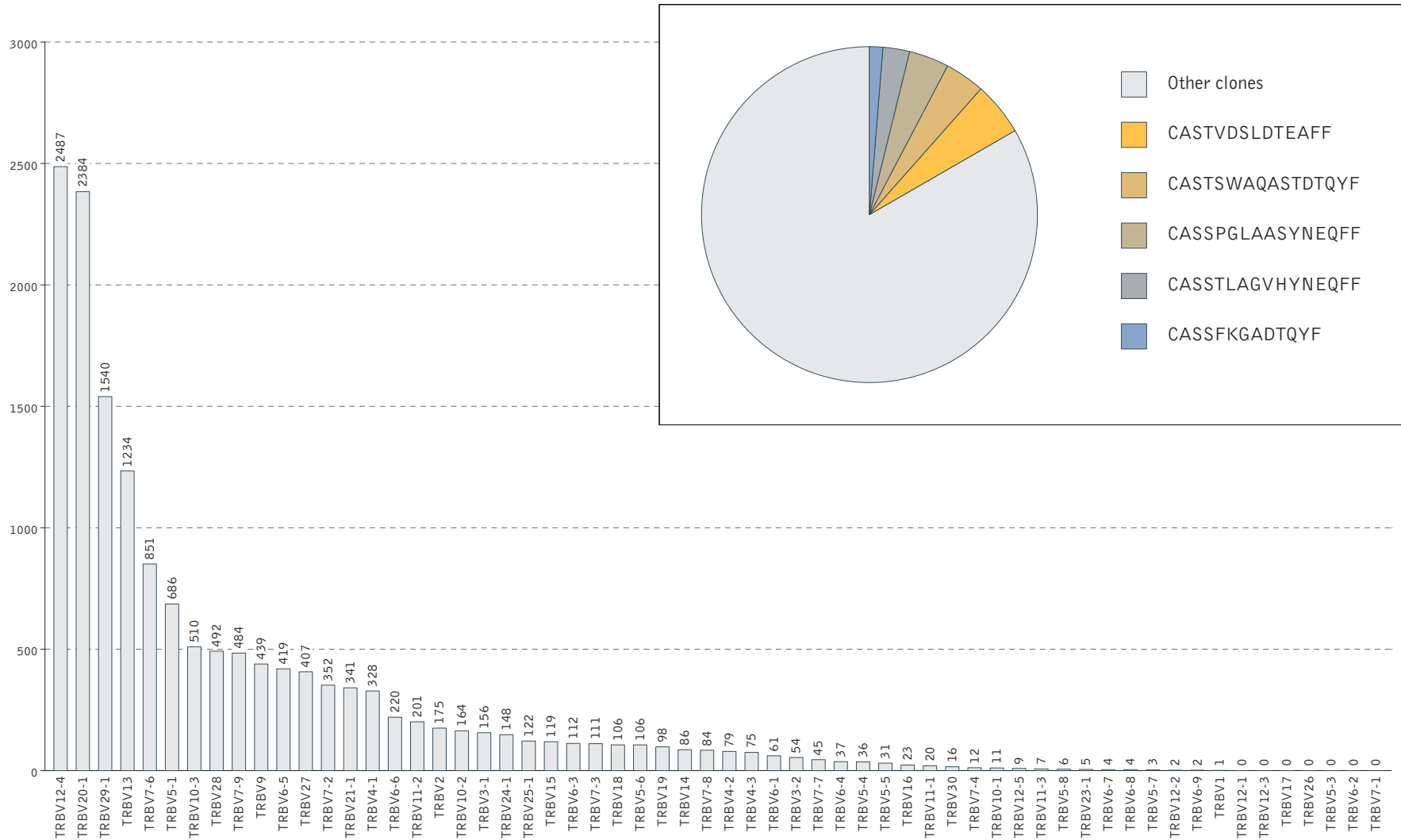
## b. Spectratypes



*In silico* CDR3 region length spectratypes for each of the V and J gene families 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. This simplifies analysis and allows for the comparison of data obtained through TCR profiling service with results from a classic spectratyping approach. CDR3 length is plotted along the x-axis. The number of sequences (i.e. abundance of a clone) is plotted along the y-axis. Spectratypes are available for all V and J genes families.

### c. Genes Abundance

The histogram shows abundance of TCR V beta genes. Gene names according to IMGT gene nomenclature are plotted along the x-axis. The number of sequence reads referred to the gene family is plotted along the y-axis. Pie-chart represents relative abundance of the major clones among other.



## IV. TRBase algorithm at a glance

BLAST databases containing V beta and J beta nucleotide reference sequences are obtained from the IMGT/ GENE-DB database. First, to extract the CDR3 sequence data, the 2<sup>nd</sup>-CYS<sup>1</sup> and J-PHE<sup>2</sup> positions in the raw sequence are determined. Next, 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 the V beta database.
2. Using the *blastn* program, the raw sequence is searched to identify matches in the J beta database.
3. In the case where matches are found in both the above searches, we take the best matches as V and J genes for the sequence.
4. The 2<sup>nd</sup>-CYS and J-PHE positions are determined using the resulting alignments.
5. Error tests are performed:
  - a) The first nucleotides of 2<sup>nd</sup>-CYS and J-PHE codons must be in the same reading frame.
  - b) No sequencing errors in CDR3 nucleotide sequence (e.g. "N" set by sequencer)
  - c) The CDR3 sequence must not have stop codons.

Such errors are shown in corresponding columns of output *\*.xls* report file and can be filtered out.

The program then clusterises 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 count (number of sequences). The set of clones is further analyzed and a set of graphical interpretations is generated.

---

<sup>1</sup>According to the [IMGT definition](#), CDR3 starts right after 2<sup>nd</sup>-cystein at position 104 of the amino acid sequence of the T-cell receptor chain

<sup>2</sup>Amino acid following the end of CDR3 according to the [IMGT definition](#)