

Bioinformatic Note



T-Cell Receptor Sequencing

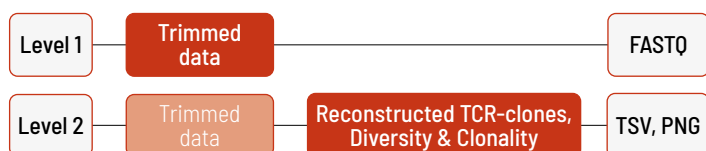
Located on the surface of T-cells, the T-cell receptor (TCR) is relevant for recognizing antigens presented by the major histocompatibility complex (MHC) molecules on antigen-presenting cells. Through somatic rearrangements, T-cells express a broad range of unique receptors. These highly diverse heterodimers are mostly composed of two subunits, the α and β chains, and a minor percentage of the γ and the δ chains. The TCR α and γ chains are generated by V/J recombination, which arises from random rearrangements of the variable (V) and joining (J) genes. The β and δ chains are generated by V/D/J recombination, which additionally includes the diversity (D) genes.

Thus, the individual TCR repertoire is shaped by V/D/J recombination. This recombination results in a highly diverse complementary-determining region 3 (CDR3). This region is an attractive target to assess the overall TCR repertoire diversity, given that it is thought to be unique to each TCR- β variant.

Investigation of the TCR repertoire can

- ✗ provide insights into functions of T-cells in immune response, e.g., immunosuppression.
- ✗ enable monitoring drug therapies, such as immunotherapies in cancer and the related change in T-cell status.
- ✗ improve personalized medicine by tumor-infiltrating T-cell analysis.

Different levels of bioinformatic data analysis are available:



With increasing bioinformatics levels, more data is delivered. Thus, Level 2 includes the reconstructed TCR-clones, and diversity and clonality analyses, as well as the trimmed data from Level 1.

Level 1

If you wish to analyze your data, we recommend Level 1, where trimmed reads in FASTQ format are delivered. At this level, the sequencing data are demultiplexed and trimmed. This level is provided for every project, regardless of additional purchased bioinformatic analyses.

The project report that is generated at the end of every project provides information for every sample about the laboratory protocol, including data about quality control of the starting material, library preparation, sequencing parameters, and the Q30 value of the sequencing. Additionally, the number of sequenced fragments and bases is reported, and the sequence length, quality of the reads and the GC content are illustrated in bar plots for all samples.

Level 2

For the TCR analyses, the FASTQ files are downsampled to 2 million read pairs to ensure comparability between samples. Reconstruction of T cell receptor sequences is performed. A first set of functional clones is generated, and a second set of clones is discarded as the clones are considered non-functional due to alternative reading frames or premature stop codons. The non-functional clones are not used in further analysis of the reconstructed TCR repertoires. Unexpectedly long CDR3 β sequences (>29 amino acids) are removed from the functional clone set because they are considered assembly artifacts.

A table in the project report summarizes the number of reads that are used as input that can successfully be merged and that are incorporated into the reconstructed clone set.

For every sample, a filtered results list is provided in TSV format that contains all TCR clones with ≤ 29 amino acids. This file is very large and contains 24 different columns. Amongst others, it includes for every identified TCR clone the nucleotide and amino acid sequences, the names of the V and J gene segments, the number of present duplicates (= reads supporting this clone), start and end positions of the different segments, and quality scores. This TSV file is in accordance with the Adaptive Immune Receptor Repertoire (AIRR Community) standard. For further information on the AIRR standards, we refer you to the AIRR standards documentation (<https://docs.airr-community.org/en/stable/>).

Reconstructed TCR-Clones

Further analyses of the TCR repertoire are performed. In the first step, the TCR repertoire is analyzed. The TCR repertoire is defined as the set of T-cell clones in a sample. The clonotypes of a TCR repertoire are the set of unique TCR sequences in a sample. Thus, a clonotype is defined by a unique nucleotide and amino acid sequence and is annotated with a frequency. The number of TCR clonotypes in every sample is calculated and stored in a TSV file (see table 1). These numbers are also visualized in a bar plot as shown in figure 1. For your convenience, both are also included in the project report. Every project receives a unique S-number (SXXXX), and every sample a unique identifier. In this example, the S-number is S1163.

In addition to the clone count frequencies, we calculate and visualize the distribution of the CDR3 lengths (see figure 2). This figure is also included in the project report.

Table 1 | Number of TCR Clonotypes after CDR3 β length filtering.

Sample	Number of clones
S1163Nr310	304200
S1163Nr311	295617
S1163Nr312	299060
S1163Nr313	290437
S1163Nr314	285382
S1163Nr315	288463
S1163Nr316	309389
S1163Nr317	304967
S1163Nr318	307266

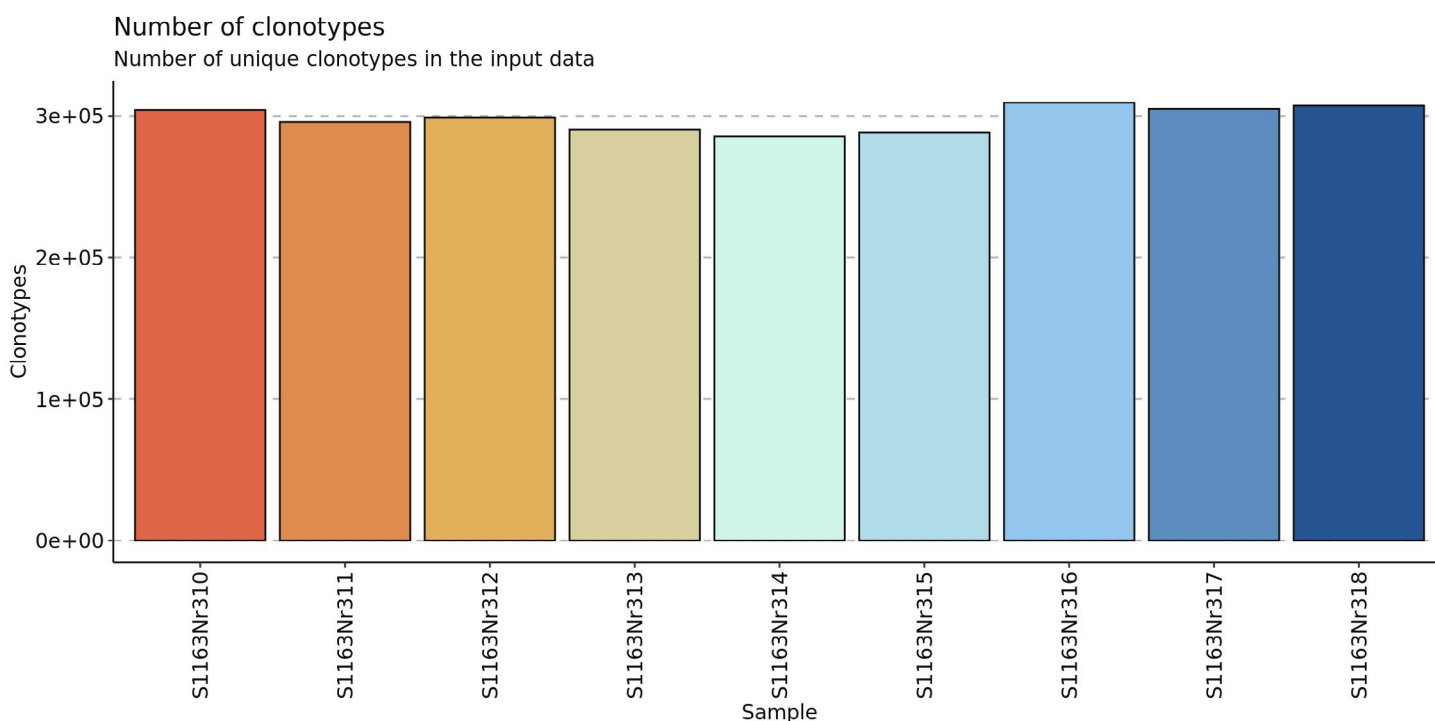


Figure 1 | Number of different clonotypes (CDR3 β sequences) in the sample.

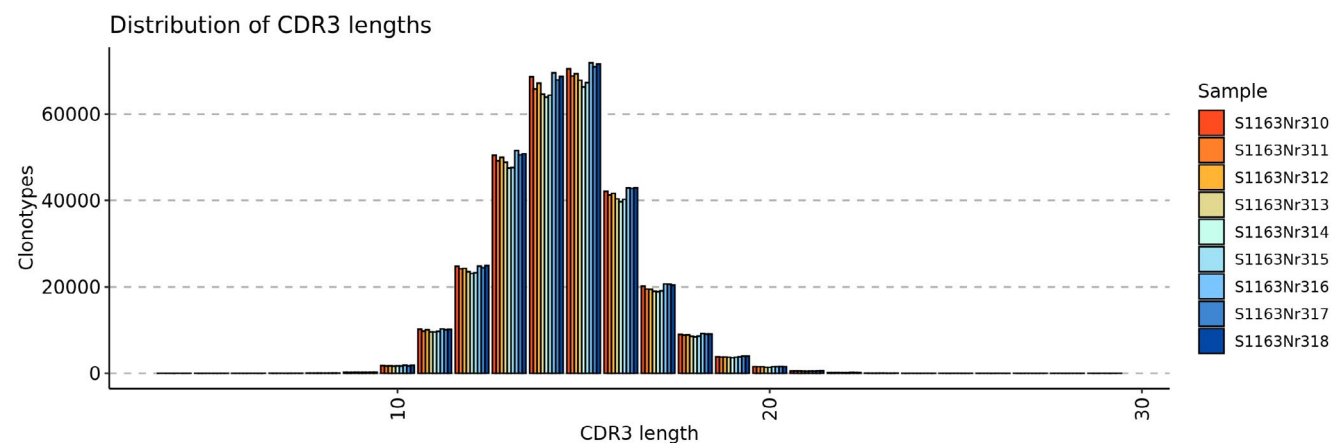


Figure 2 | Sequence length distribution of reconstructed CDR3 β regions (in AA).

We create an additional TSV file for every sample with the clones and their frequencies for your convenience. This table is a standardized subset of the filtered results table. An excerpt of one additional TSV file is shown in table 2. For the sake of clarity, we excluded the columns without values in this excerpt. Additionally, we shortened the nucleotide sequences.

In the delivered TSV file, all columns and the whole sequences are of course, reported. As seen in table 2, this TSV file lists the absolute abundance of the clones, their proportion, the nucleotide and amino acid sequences, and the names of the V and the J segments.

Table 2 | Excerpt of a clones tsv file that is additionally provided for every sample.

Clones	Proportion	CDR3.nt	CDR3.aa	V.name	J.name	Sequence
64134	0.0350	TGTGCCAGCA...	CASSFSTCSANYGYTF	TRBV12-4*01	TRBJ1-2*01	GATGCTGGAG...
8704	0.0048	TGTGCCAGCA...	CASSSRSNPEQYF	TRBV7-2*01	TRBJ2-7*01	GGAGCTGGAG...
8348	0.0046	TGTGCCAGTA...	CASSRHGQDTQYF	TRBV19*01	TRBJ2-3*01	GATGGTGGAA...
7285	0.0040	TGTGCTAGTG...	CASGSKRDRGQETQYF	TRBV12-5*01	TRBJ2-5*01	GATGCTAGAG...
5870	0.0032	TGCGCCAGCA...	CASSQVQGGNQPQHF	TRBV4-1*01	TRBJ1-5*01	GACTACTGAAG...

Diversity & Clonality

Besides the basic analysis of the TCR repertoire, we also analyze the diversity and clonality of the TCRs. Diversity and clonality are measures to describe the composition and frequency distribution of a TCR repertoire. We used the Shannon Entropy as a measure of diversity of a TCR repertoire. Simpson Clonality is a single measure for the frequency distribution of the clones within the TCR repertoire. The third measure is the repertoire clonality index.

These measures are calculated for every sample. The whole list is delivered as a TSV file. Additionally, the information can be found in the project report. An example of such a TSV file is shown in table 3: The diversity, calculated with the Shannon entropy, the Simpson clonality, and a repertoire clonality index are calculated and delivered.

In addition to a single measure for clonality, a graphical representation of the frequency distribution can be used to assess differences in the clonality between samples. Besides the graphical representation shown in figure 3, we also provide the data as a TSV file (see table 4). The definition of the clonotype groups (rare, small, medium, large, hyperexpanded) is given in the head of each column. For each sample, the relative abundance of these groups is given.

The last provided TSV file contains the public repertoire. An excerpt is shown in table 5. In this file, all TCR clones from all samples are listed, and their frequency in each sample is indicated. The CDR3 amino acid sequence and the name of the V segment are given, together with the number of samples, in which the clone occurs.

Additional Analyses

Upon request, we can provide further TCR analyses, such as clonotype tracking, or repertoire changes across different time points.

Table 3 | Diversity (Shannon Entropy), clonality, and repertoire clonality index of the TCR repertoires.

Sample	Diversity (Shannon Entropy)	Simpson Clonality	Repertoire Clonality Index
S1163Nr310	14.7	0.17	0.65
S1163Nr311	14.62	0.17	0.65
S1163Nr312	14.66	0.17	0.65
S1163Nr313	16.36	0.04	0.61
S1163Nr314	16.42	0.04	0.61
S1163Nr315	16.45	0.04	0.61
S1163Nr316	16.89	0.01	0.6
S1163Nr317	16.87	0.01	0.6
S1163Nr318	16.89	0.01	0.6

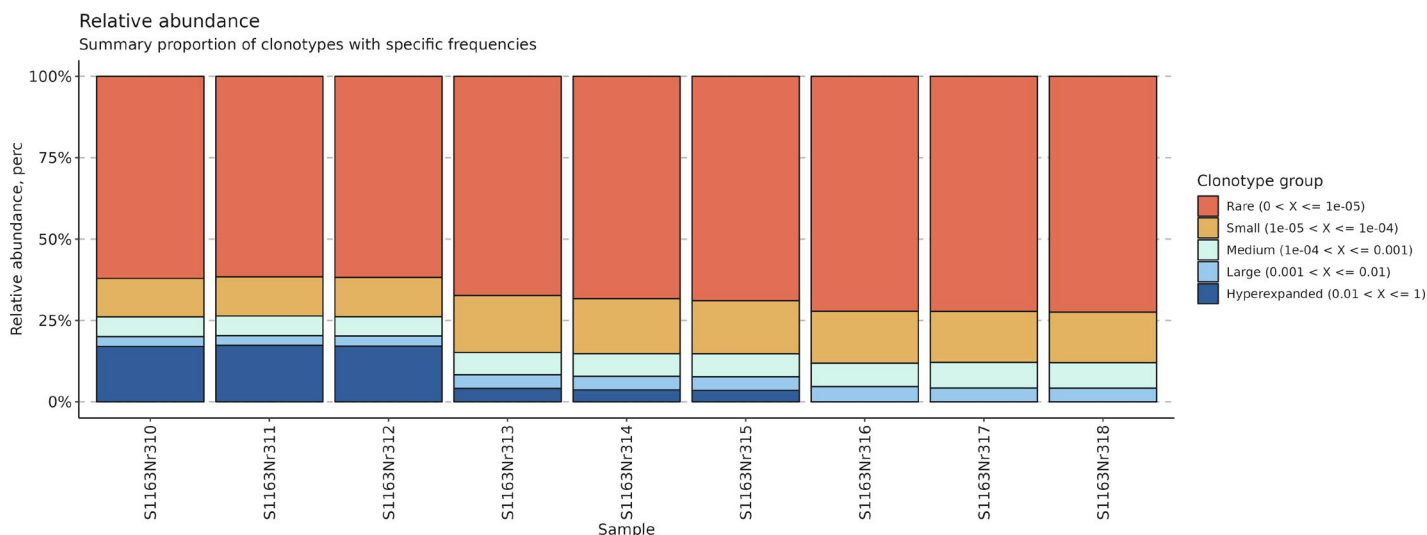


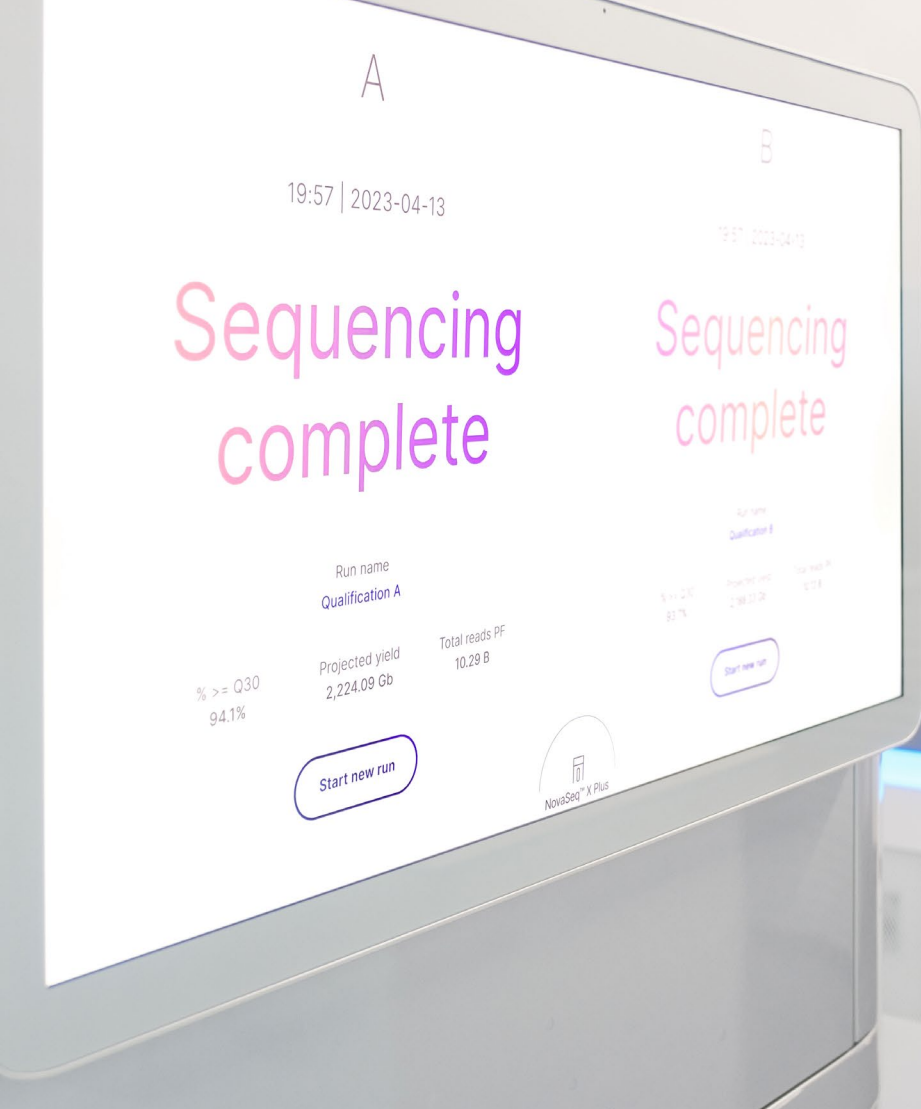
Figure 3 | Relative abundance: The clones within the T cell receptor repertoire were grouped by their frequency (see Figure legend for the definition of the clonotype groups). For each sample the relative abundances of rare, small, medium, large, and hyperexpanded groups are given in percent.

Table 4 | Relative abundance in TSV format.

Sample	Rare ($0 < X \leq 1e-05$)	Small ($1e-05 < X \leq 1e-04$)	Medium ($1e-04 < X \leq 0.001$)	Large ($0.001 < X \leq 0.01$)	Hyperexpanded ($0.01 < X \leq 1$)
S1163Nr310	0.62095	0.11854	0.06063	0.03003	0.16986
S1163Nr311	0.61598	0.12043	0.06077	0.02935	0.17347
S1163Nr312	0.61773	0.12143	0.05903	0.03067	0.17114
S1163Nr313	0.67332	0.17559	0.06818	0.04164	0.04128
S1163Nr314	0.68351	0.16919	0.06880	0.04194	0.03656
S1163Nr315	0.68943	0.16342	0.07011	0.04199	0.03505
S1163Nr316	0.72213	0.15934	0.07207	0.04646	0
S1163Nr317	0.72263	0.15638	0.07870	0.04229	0
S1163Nr318	0.72469	0.15523	0.07718	0.04290	0

Table 5 | Public repertoire.

	CDR3.aa	V.name	Sam- ples	S1163 Nr310	S1163 Nr311	S1163 Nr312	S1163 Nr313	S1163 Nr314	S1163 Nr315	S1163 Nr316	S1163 Nr317	S1163 Nr318
1	CAAERGGHNNQFF	TRBV19*01	9	4,89E+08	6,51E+08	5,42E+07	1,04E+09	9,84E+08	1,20E+09	5,51E+08	1,42E+09	1,42E+09
2	CAALGGNTGELFF	TRBV10-2*01	9	7,06E+08	1,41E+09	1,63E+08	9,83E+08	3,28E+08	2,19E+08	6,06E+08	1,25E+09	9,80E+08
3	CAASGGTDTQYF	TRBV18*01	9	1,41E+09	5,42E+08	6,51E+08	2,19E+08	1,15E+09	1,20E+09	8,26E+08	2,73E+08	9,25E+08
4	CACLLGTGLEYGTYF	TRBV30*01	9	1,41E+09	3,25E+09	2,66E+08	1,20E+09	3,23E+09	1,09E+09	1,93E+09	2,02E+08	7,08E+08
5	CACLYGGAGLNEQFF	TRBV30*01	9	4,34E+08	7,59E+08	7,05E+08	7,10E+08	4,92E+08	7,10E+08	1,16E+09	8,73E+07	1,14E+09



About Us

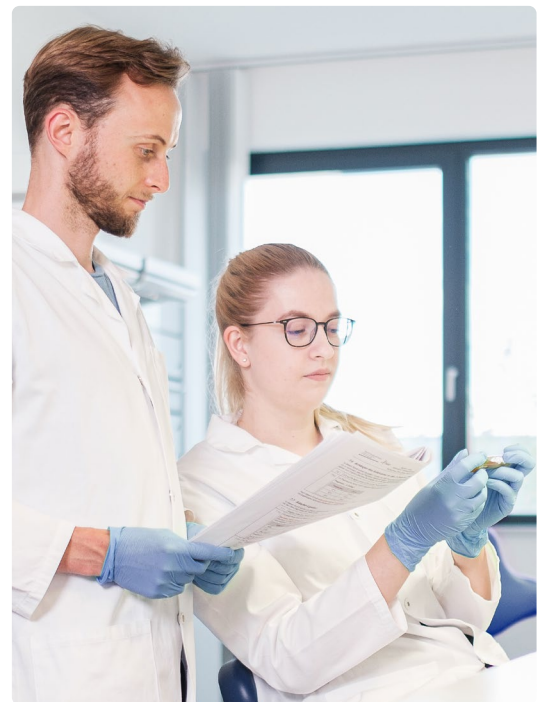
CeGaT was founded in 2009 in Tübingen, Germany. Our scientists are specialized in next-generation sequencing (NGS) for genetic diagnostics, and we also provide a variety of sequencing services for research purposes and pharma solutions. Our sequencing service portfolio is complemented by analyses suited for microbiome, immunology, and translational oncology studies.

Our dedicated project management team of scientists and bioinformaticians works closely with you to develop the best strategy to realize your project. Depending on its scope, we select the most suitable library preparation and conditions on our sequencing platforms.

We would be pleased to provide you with our excellent service.
Contact us today to start planning your next project.



For more details please visit
www.cegat.com/t-cell-receptor-sequencing



CeGaT GmbH
Research & Pharma Solutions
Paul-Ehrlich-Str. 23
72076 Tübingen
Germany

Phone: +49 7071 56544-333
Fax: +49 7071 56544-56
Email: rps@cegat.com



CLIA CERTIFIED ID: 99D2130225



Accredited by DAKKS according to
DIN EN ISO/IEC 17025