InterClone Documentation

Release latest

Sep 27, 2022

HELP

The InterClone webserver and associated database provides easily accessible tools for storing, searching and clustering adaptive immune receptor repertoire (AIRR) sequence datasets. InterClone was designed to allow users to control the visibility of their own data. To this end data consists of "public" datasets, which are visible to all, and "private" datasets, which are visible only to the user who stored the data. To create a private dataset, you need to create an account, which is free to do. The source code of the backend pipeline is available on Gitlab.

CHAPTER

ONE

STORE A NEW DATASET

Storing a new dataset is easy as long as the files conform to the subset of AIRR standards required by InterClone. Please see the requirements and example datasets in order to understand what valid input looks like. If you do not know how to prepare AIRR-formatted files, please take a look at preparing AIRR files.

One or more AIRR-formatted files should be combined as a zip file (see here how to do this). To begin the process of storing a new dataset, please click the ”Store” menu item.

Give your dataset a name. This name can be anything, but we recommend following a convention, such as “<author>-<date>” so that it’s easy to browse later.

Next select the appropriate “Receptor Type” and “Chain Type”. Single-cell sequencing data should be stored as “paired”, while bulk data for a given chain should be uploaded separately. Please note that InterClone does not explicitly distinguish between species, but most of the data is human.

Next, provide tags for your data. Tags are used to filter dataset for use in search or cluster jobs. Typically, the species, or descriptions of the donors (healthy, etc.) should be provided.

When you are all ready, browse for the zip file, and click “Store Dataset”. Please be patient, as each sequence will be annotated (CDR regions defined, encoded and stored on our local filesystem.) Depending on the load on our server and the size of your data, this can take from minutes to hours.

CHAPTER

TWO

SEARCH DATASETS

The search tool allows you to find sequences whose CDRs match within specified identity thresholds. This can be helpful for locating receptors that bind to the same epitope as the query, although there are always tradeoffs between sensitivity (the fraction of true sequences that are found) and specificity (the fraction of found hits that are true). The default identity thresholds for each CDR are set to achieve a reasonable balance, but you should adjust as needed. Note, however, that reducing the coverage threshold below 90 may potentially yield matches with low significance.

Input consists of a full-length variable region amino acid sequence. The rest of the input fields are identical to those of the store tool. This is because your query will be stored and can be accessed at any time for reuse.

If your query is a TCR and you do not know the full length sequence, you can try to assemble it from the V and J gene names and the CDR3 sequence using our assembly tool.

Next, select the datasets that you want to search. In order to reduce load on our server, we restrict the searched data to be no more than 200 million sequences. Click "Search" and your search should start immediately. Please expect to wait a few minutes for a small to medium sized search (~100,000 sequences).

To follow a real world use case with inputs and results, please see the tutorial.

CHAPTER

THREE

CLUSTER DATASETS

Clustering involves selecting one or more datasets, then clicking “Cluster”. You will be directed to a waiting page while the job completes (which should take a few minutes or less). The results page should load automatically with a URL you can bookmark. The results consist of a table of clusters, sorted by decreasing size.You can download a summary of the clusters as a TSV or XLS file. You can also download an expanded table, which consists of the original AIRR file with additional columns containing the clusters.

To follow a real world use case with inputs and results, please see the tutorial.

CHAPTER

FOUR

PREPARING AIRR FILES

InterClone requires AIRR-formatted files, which are tab-delimited files with a number of specific column headers. In order to use InterClone, the following headers are required:

- sequence_id, containing a unique identifier for a sequence entry
- sequence_aa, containing the complete amino acid sequence
- v_call, containing the assigned V gene name for chain filtering
- clone_id, containing a unique identifier for each clone, used for paired result merging

Input files can be prepared from raw FASTA, 10X CellRanger, Illumina MIRA and MiXCR-formatted data. An import script is provided for each of these data sources. They can be found in the src/dataimport/ folder in the source code repository.

In the case of FASTA-formatted data, the user provides one or more input files containing full length amino acid sequences, along with the chain type of these sequences.

In the case of raw MiXCR TSV outputs, AIRR files are prepared as follows:

- Conjugate the full length amino acid sequence by merging the CDR and framework regions from 'aaSeqImputedFR1', 'aaSeqImputedCDR1','aaSeqImputedFR2', 'aaSeqImputedCDR2', 'aaSeqImputedFR3', 'aaSeqImputedCDR3' and 'aaSeqImputedFR4'. Discard sequences containing gaps or stop codons in these regions.
- The columns 'cloneId', 'allVHitsWithScore', 'allJHitsWithScore', 'aaSeqImputedCDR3', 'allCHitsWithScore', and 'cloneCount' from the MiXCR output file are renamed to 'clone_id', 'v_call', 'j_call', 'cdr3', 'c_call' and 'clone_count', respectively.

In the case of 10X CellRanger output, three files are required: 'airr_rearrangement.tsv', 'clonotypes.csv' and 'all_contig_annotations.csv'. To prepare AIRR formatted files, the following steps are taken:

- Examine the quality of each clone: Only clones that contain one paired heavy/light or alpha/beta chain are used. Clones containing multiple chains or single chains are discarded. The frequency of each clone taken from 'clonotypes.csv' and saved as 'clone_count' in the AIRR file.
- Select one of the contigs that share the same clonotype and use the columns 'cell_id', 'v_call', 'j_call', 'cdr3', 'c_call' and amino acid sequence for the AIRR file.

The following is a walk-through for the three use cases described in the paper:

CHAPTER

FIVE

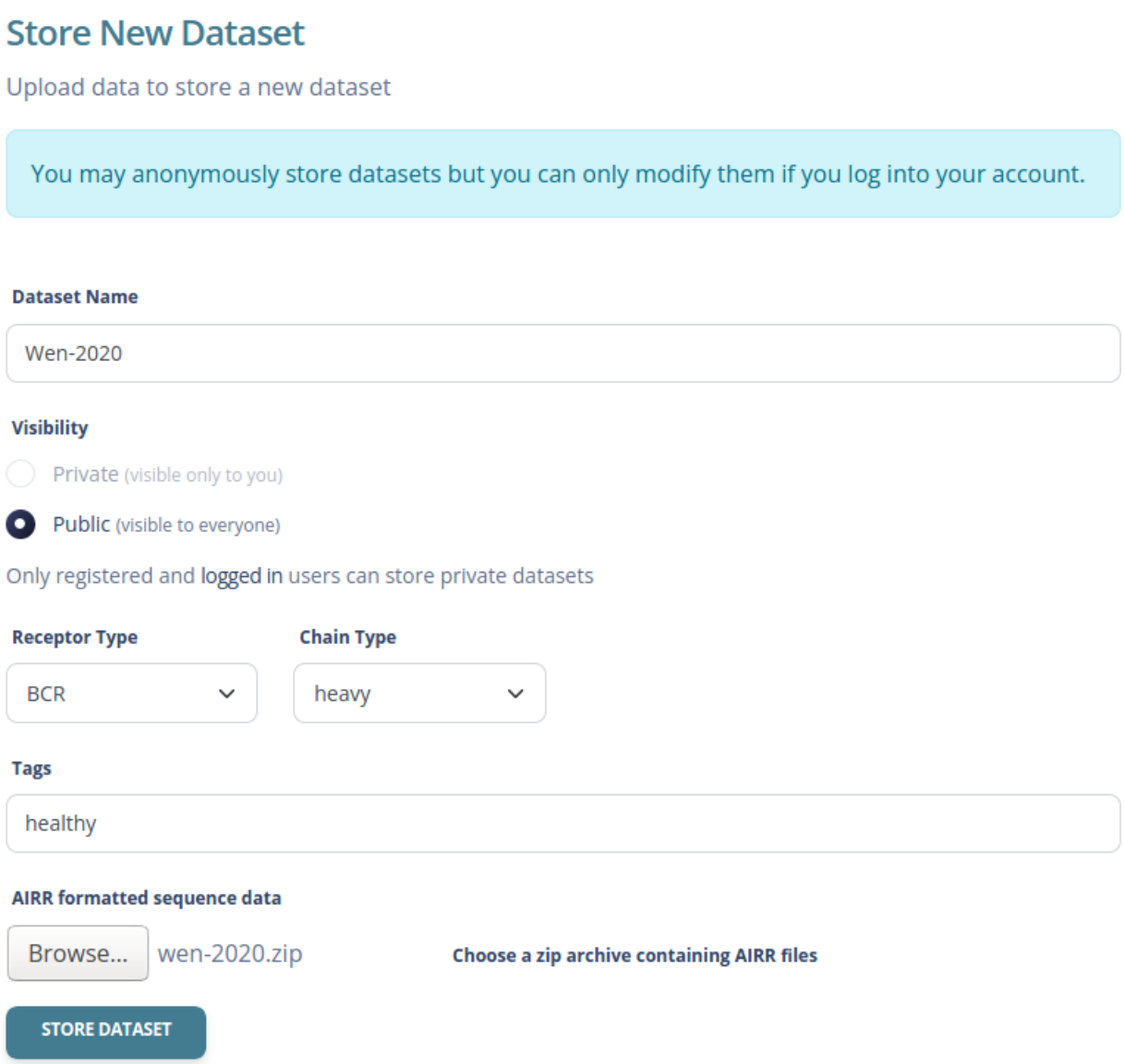
STORING A NEW DATASET

Searching healthy and COVID-19 data for infection enhancing antibodies

Clustering healthy and COVID-19 TCR data to investigate common CDR sequence motifs

All described functions are accessible without the need for a user account. However, additional features are available to those who register, like storing private datasets and reusing previous search queries. Registration is free and simple. Storing a new dataset To demonstrate storing a dataset in the InterClone database, we use one of the publicly available datasets, "Wen-2020", that was published by Wen, et al. Since the raw data needs to be processed into AIRR format in order to be usable by InterClone, we provide the `prepared dataset`. It consists of a zip archive containing four TSV files (one per donor) with full length amino acid sequences as well as chain identifiers.

On the InterClone web server, select the Store tool. Enter a name for the dataset (e.g. "Wen-2020") and choose the correct Receptor Type (in this case, "BCR") as well as Chain Type (in this case, "heavy"). It is recommended to add tags to make the dataset easier to find later. Since we have data from healthy donors here, we can enter "healthy" as a tag. Then, browse for the prepared zip archive and select it for upload. The filled out form should look like this:



After clicking on "Store dataset", you will be redirected to the Profile page which will show a summary of your dataset. Once the dataset has been stored in the database, the status will show as "PREPARED" and the dataset can be used for Searching and Clustering. Please check the number of successfully processed sequences and compare it with the size of the original input. A large disparity between the two counts means that a lot of your input data could not be processed properly. This can happen for a number of reasons, like unknown chain types or unusual donor species. You can contact us if you think your data is fine and should have been processed. Note that anonymous users can only store public datasets and are not able to delete these afterwards. Please consider creating a user account for advanced management of your data.

CHAPTER

SIX

SEARCHING BCR DATASETS

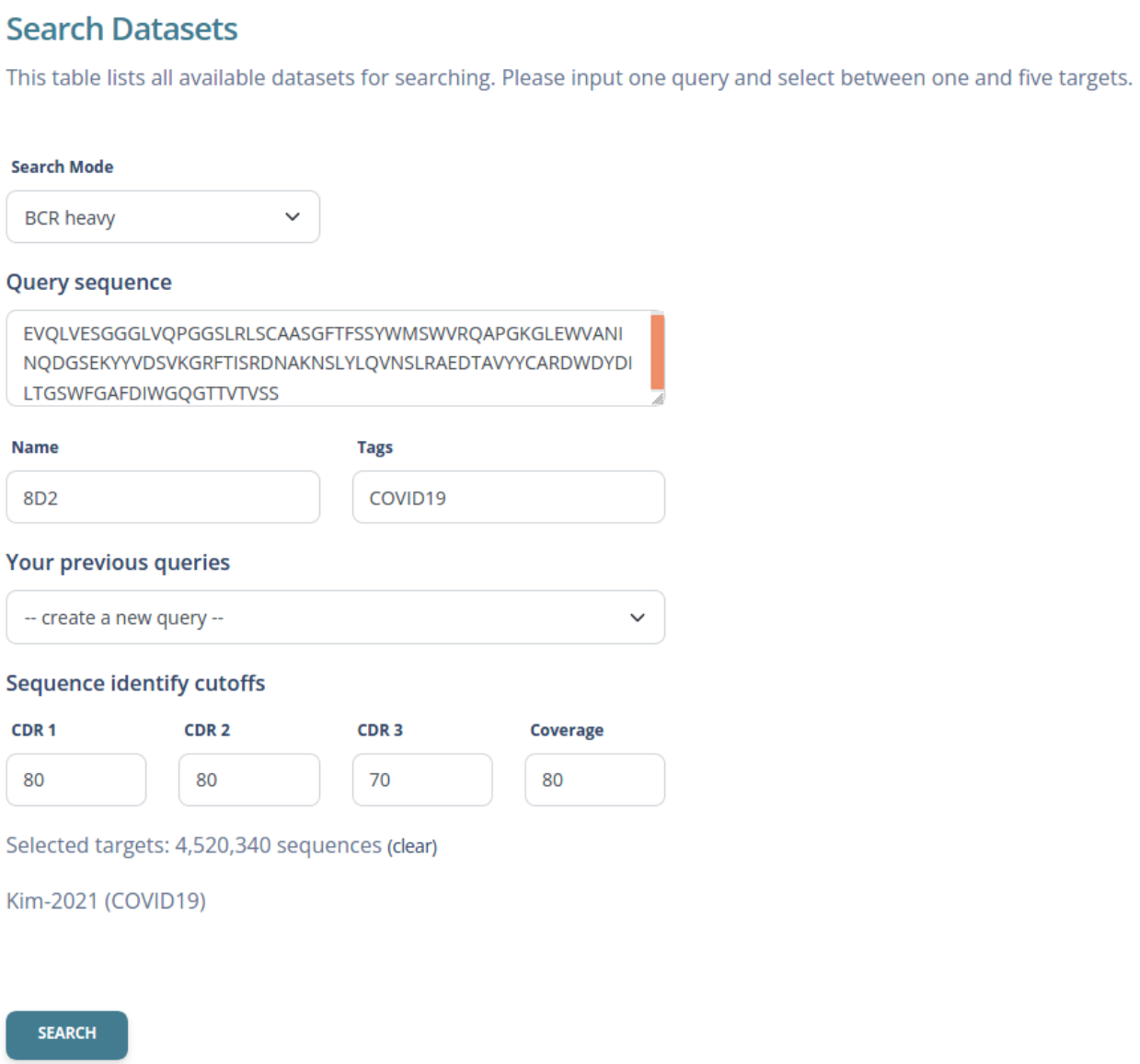
We will investigate the distribution of infection enhancing antibodies in both disease and healthy donors. For this, we will need the full length amino acid sequences of the enhancing antibodies, which can be obtained from Cov-AbDab. For example, the antibody 8D2 has the following sequence:

EVQLVESGGGLVQPGGSLRLSCAASGFTFSSYWMSWVRQAPGKGLEWVANIN-
QDGSEKYYVDSVKGRFTISRDNAKNSLYLQVNSLRAEDTAVYYCARDWDYDILTGSWFGAFDIWGQGTTVTVSS

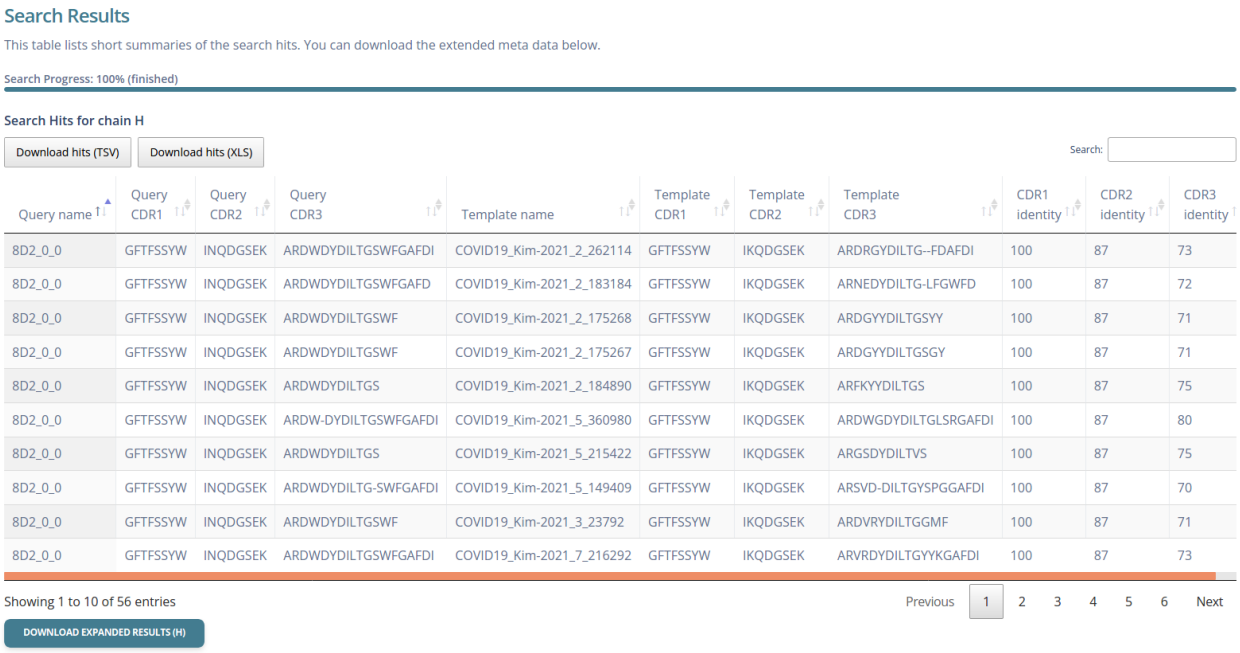
The results for the COVID19 search can be accessed here. The results for the healthy search can be accessed here.

6.1 Searching COVID19 data

On the InterClone web server, select the Search function and insert the above sequence into the "Query sequence" field. Enter a name for the query (e.g. "8D2") and, optionally, some tags (e.g. "COVID19"). Then choose the appropriate sequence identity cutoff values, i.e. 80/80/70 for CDRs 1, 2 and 3, respectively, and 80 for the coverage. Then, search for the "Kim-2021" dataset in the table of target datasets. You can filter the results by name, type and tags. Choose "Use as target" in the last column and the name of the dataset will appear in the "Selected targets" section, which will also indicate the total number of sequences that are about to be searched. The input form should look like this:



Click the "Search" button and wait for the result, which should appear after a few minutes. The progress will be indicated on the results page and will automatically reload so that once the search has finished, you should see this result:



Search Results

This table lists short summaries of the search hits. You can download the extended meta data below.

Search Progress: 100% (finished)

Search Hits for chain H

Download hits (TSV) Download hits (XLS) Search:

Query name	Query CDR1	Query CDR2	Query CDR3	Template name	Template CDR1	Template CDR2	Template CDR3	CDR1 identity	CDR2 identity	CDR3 identity
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGSWFGAFDI	COVID19_Kim-2021_2_262114	GFTFSSYW	IKQDGSEK	ARDRGYDILTG--FDAFDI	100	87	73
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGSWFGAFD	COVID19_Kim-2021_2_183184	GFTFSSYW	IKQDGSEK	ARNEDYDILTG-LFGWFD	100	87	72
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGSWF	COVID19_Kim-2021_2_175268	GFTFSSYW	IKQDGSEK	ARDGYYDILTGSYY	100	87	71
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGSWF	COVID19_Kim-2021_2_175267	GFTFSSYW	IKQDGSEK	ARDGYYDILTGSGY	100	87	71
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGS	COVID19_Kim-2021_2_184890	GFTFSSYW	IKQDGSEK	ARFKYYDILTGS	100	87	75
8D2_0_0	GFTFSSYW	INQDGSEK	ARDW-DYDILTGSWFGAFDI	COVID19_Kim-2021_5_360980	GFTFSSYW	IKQDGSEK	ARDWGDYDILTGLSRGAFDI	100	87	80
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGS	COVID19_Kim-2021_5_215422	GFTFSSYW	IKQDGSEK	ARGSDYDILTVS	100	87	75
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTG-SWFGAFDI	COVID19_Kim-2021_5_149409	GFTFSSYW	IKQDGSEK	ARSVD-DILTGYSPGGAFDI	100	87	70
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGSWF	COVID19_Kim-2021_3_23792	GFTFSSYW	IKQDGSEK	ARDVRYDILTGGMF	100	87	71
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGSWFGAFDI	COVID19_Kim-2021_7_216292	GFTFSSYW	IKQDGSEK	ARVRDYDILTGYYKGAFDI	100	87	73

Showing 1 to 10 of 56 entries Previous 1 2 3 4 5 6 Next

DOWNLOAD EXPANDED RESULTS (H)

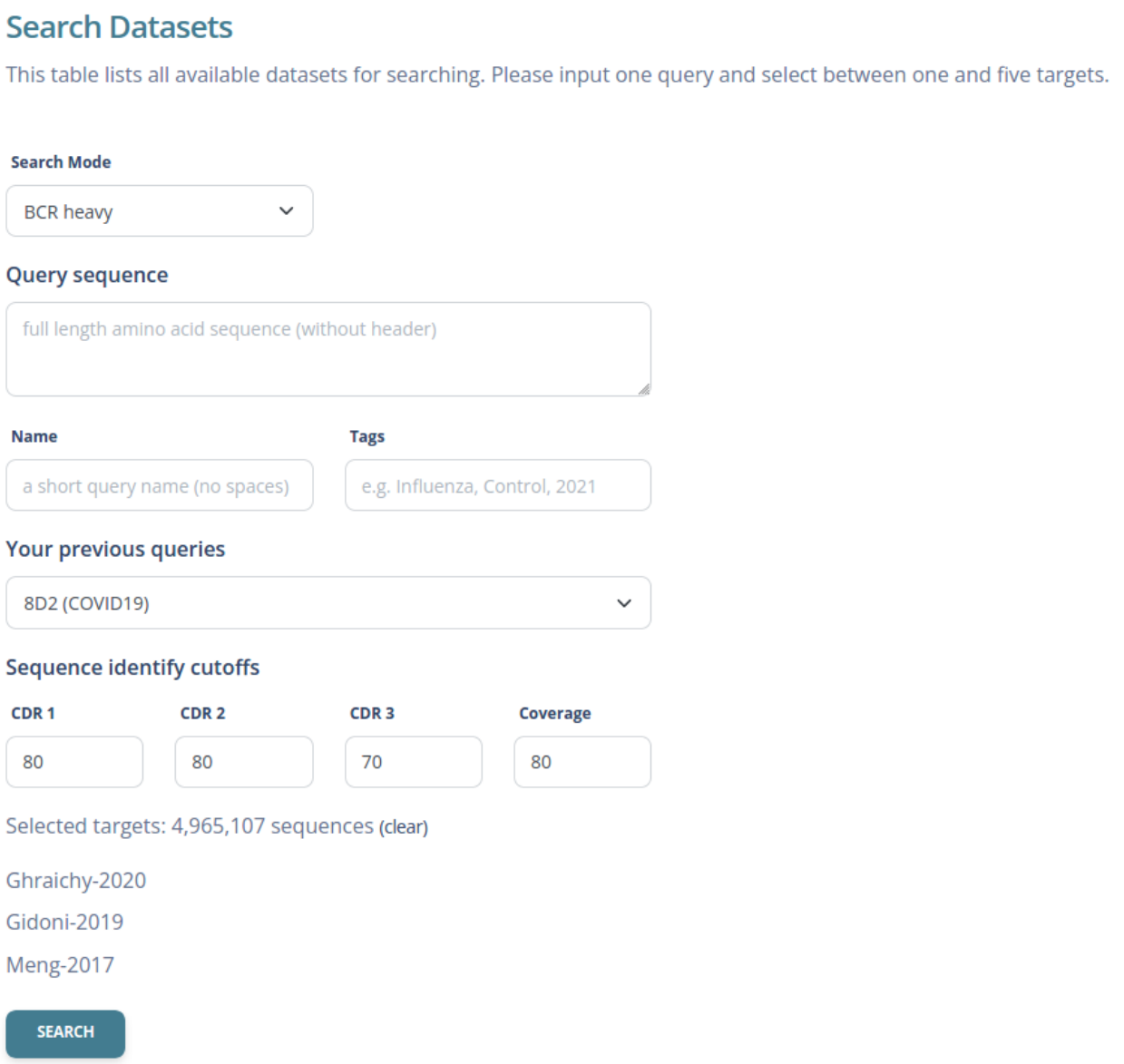
This table shows a summary of the search hits: the matched query and template sequences as well as their similarity scores, separately for each CDR. By clicking on “Download expanded results”, you can access additional metadata from the original inputs that might be useful for further analysis, e.g. “clone_count”.

6.2 Searching healthy data

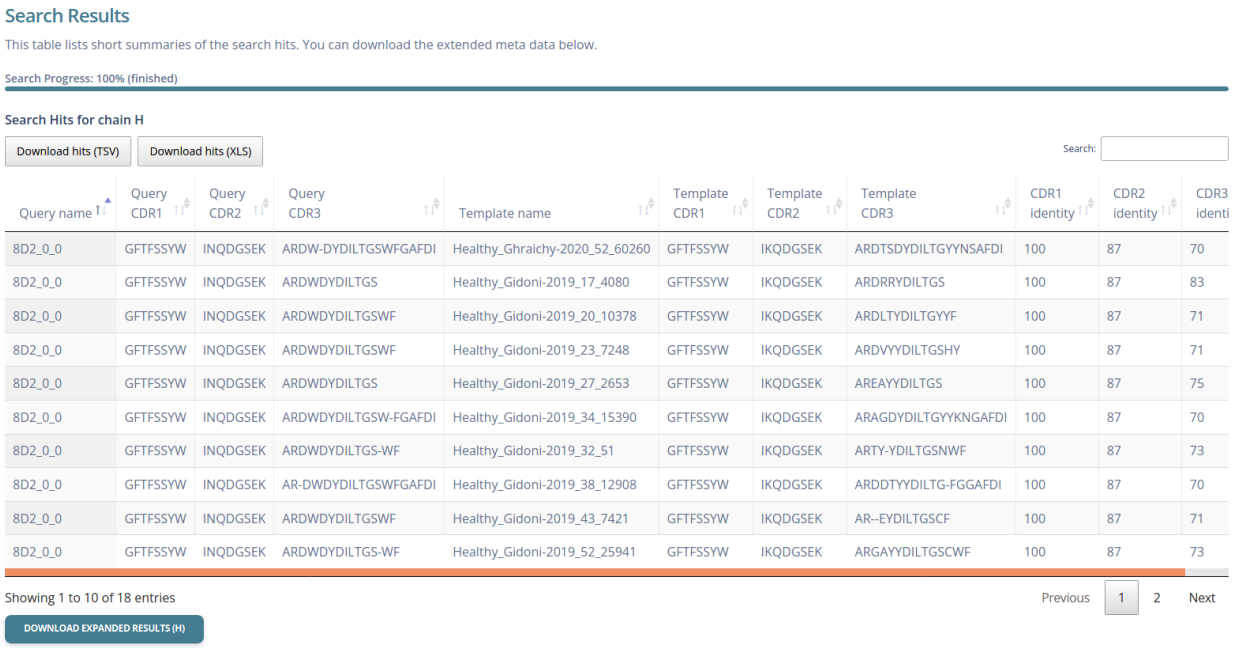
We will now repeat the above search for healthy data. If you are logged in, you can select the previously used query on the search page. Otherwise, you will have to re-enter the sequence. Again, select the appropriate thresholds (80, 80, 70 and 80). This time, filter the target datasets table for the following three datasets and select them as search targets:

- Gidoni-2019
- Ghraichy-2020
- Meng-2017

The input form should look like this, assuming you are reusing the previous query:



Press “Search” and wait for the results, which should look like this:



Search Results

This table lists short summaries of the search hits. You can download the extended meta data below.

Search Progress: 100% (finished)

Search Hits for chain H

Download hits (TSV) Download hits (XLS) Search:

Query name	Query CDR1	Query CDR2	Query CDR3	Template name	Template CDR1	Template CDR2	Template CDR3	CDR1 identity	CDR2 identity	CDR3 identi
8D2_0_0	GFTFSSYW	INQDGSEK	ARDW-DYDILTGSWFGAFDI	Healthy_Ghraichy-2020_52_60260	GFTFSSYW	IKQDGSEK	ARDTSDYDILTGYYNSAFDI	100	87	70
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGS	Healthy_Gidoni-2019_17_4080	GFTFSSYW	IKQDGSEK	ARDRRYDILTGS	100	87	83
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGSWF	Healthy_Gidoni-2019_20_10378	GFTFSSYW	IKQDGSEK	ARDLTYDILTGYYF	100	87	71
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGSWF	Healthy_Gidoni-2019_23_7248	GFTFSSYW	IKQDGSEK	ARDVYYDILTGSHY	100	87	71
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGS	Healthy_Gidoni-2019_27_2653	GFTFSSYW	IKQDGSEK	AREAYYDILTGS	100	87	75
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGSW-FGAFDI	Healthy_Gidoni-2019_34_15390	GFTFSSYW	IKQDGSEK	ARAGDYDILTGYYKNGAFDI	100	87	70
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGS-WF	Healthy_Gidoni-2019_32_51	GFTFSSYW	IKQDGSEK	ARTY-YDILTGSNWF	100	87	73
8D2_0_0	GFTFSSYW	INQDGSEK	AR-DWDYDILTGSWFGAFDI	Healthy_Gidoni-2019_38_12908	GFTFSSYW	IKQDGSEK	ARDDTYYDILTG-FGGAFDI	100	87	70
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGSWF	Healthy_Gidoni-2019_43_7421	GFTFSSYW	IKQDGSEK	AR--EYDILTGSCF	100	87	71
8D2_0_0	GFTFSSYW	INQDGSEK	ARDWDYDILTGS-WF	Healthy_Gidoni-2019_52_25941	GFTFSSYW	IKQDGSEK	ARGAYYDILTGSCWF	100	87	73

Showing 1 to 10 of 18 entries Previous 1 2 Next

DOWNLOAD EXPANDED RESULTS (H)

This time, there are fewer search hits than for the COVID19 data but as mentioned above, it's important to consider clonal expansion by checking the clone_count column in the extended results. The above steps can be repeated for all 11 known enhancing antibodies. The separate results can then be aggregated.

CHAPTER

SEVEN

CLUSTERING TCR DATASETS

Here, we want to analyze common patterns in TCR alpha sequences and specifically look for a recently discovered sequence motif in the CDRs that was published by Mudd, et al.:

SIFNT LYKAGEL CA[G/A/V]XNYGGSQGNLIF

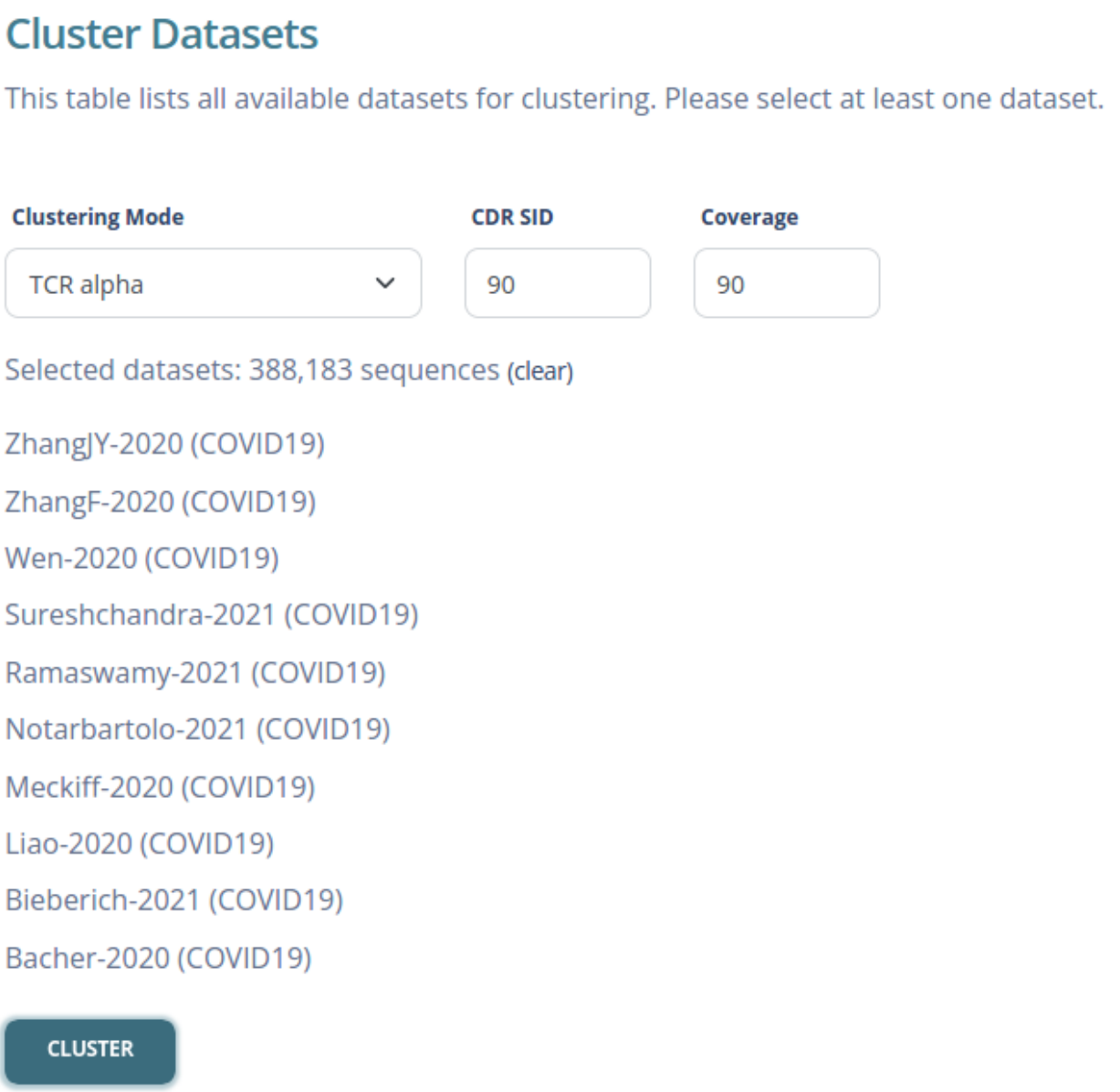
The final results of the COVID19 clustering can be accessed here. The healthy clustering results are accessible here.

7.1 Clustering COVID19 data

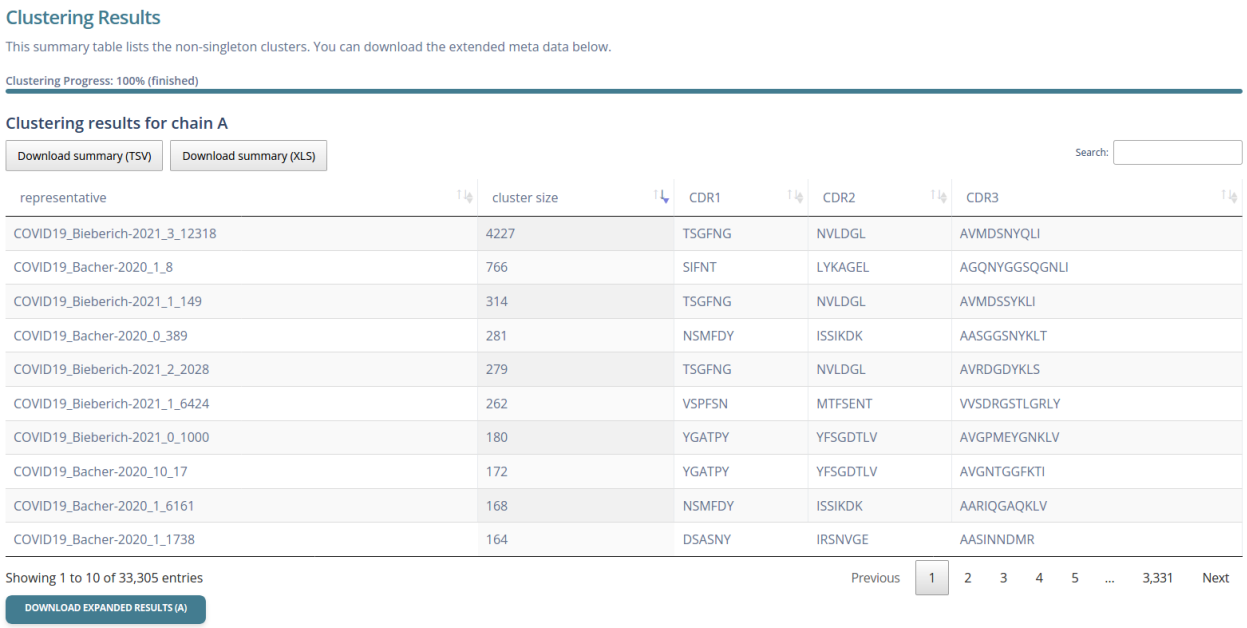
First, we will have a look at COVID19 datasets. Access the Cluster page and choose "TCR alpha" as Clustering Mode. The datasets table will update with the available datasets. Select the following via the checkbox in the last column:

- Bacher-2020
- Bieberich-2021
- Liao-2020
- Meckiff-2020
- Notarbartolo-2021
- Ramaswamy-2021
- Sureshchandra-2021
- Wen-2020
- ZhangF-2020
- ZhangJY-2020

Make sure to set appropriate values for sequence identity (90) and coverage (90). The input form should look like this:



Click "Cluster" and wait for the result to appear, this should only take a few minutes. The results should look like this:



Clustering Results

This summary table lists the non-singleton clusters. You can download the extended meta data below.

Clustering Progress: 100% (finished)

Clustering results for chain A

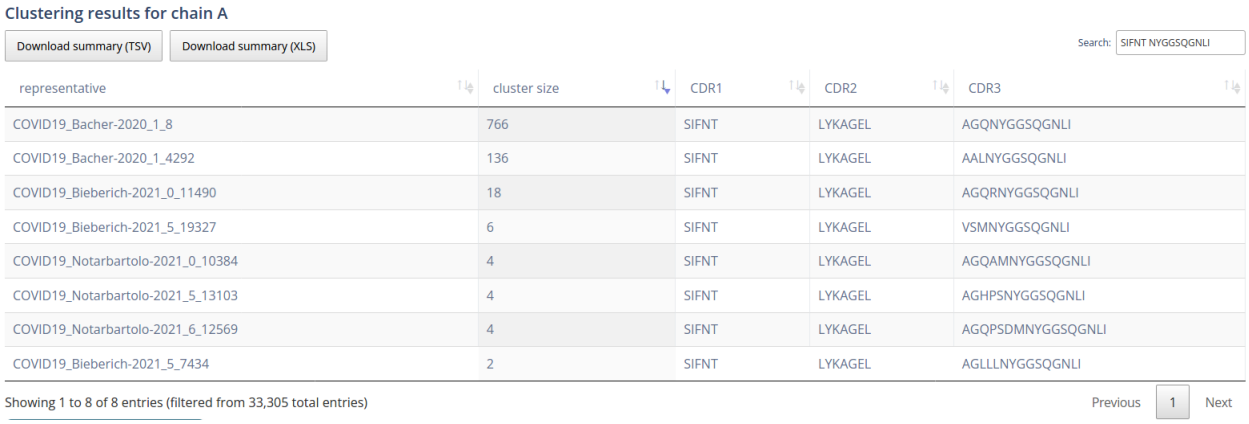
Download summary (TSV) Download summary (XLS) Search:

representative	cluster size	CDR1	CDR2	CDR3
COVID19_Bieberich-2021_3_12318	4227	TSGFNG	NVLDGL	AVMDSNYQLI
COVID19_Bacher-2020_1_8	766	SIFNT	LYKAGEL	AGQNYGGSQGNLI
COVID19_Bieberich-2021_1_149	314	TSGFNG	NVLDGL	AVMDSSYKLI
COVID19_Bacher-2020_0_389	281	NSMFDY	ISSIKDK	AASGGSNYKLT
COVID19_Bieberich-2021_2_2028	279	TSGFNG	NVLDGL	AVRDGDYKLS
COVID19_Bieberich-2021_1_6424	262	VSPFSN	MTFSENT	VVSDRGSTLGRLY
COVID19_Bieberich-2021_0_1000	180	YGATPY	YFSGDTLV	AVGPMEYGNKLV
COVID19_Bacher-2020_10_17	172	YGATPY	YFSGDTLV	AVGNTGGFKTI
COVID19_Bacher-2020_1_6161	168	NSMFDY	ISSIKDK	AARIQGAQKLV
COVID19_Bacher-2020_1_1738	164	DSASNY	IRSNVGE	AASINNDMR

Showing 1 to 10 of 33,305 entries Previous 1 2 3 4 5 ... 3,331 Next

DOWNLOAD EXPANDED RESULTS (A)

We can see that many clusters exhibit motifs from invariant TCRs (i.e. MAIT-like and iNKT cells), including the largest one. The second largest cluster however contains the above mentioned public Spike protein targeting motif. Just as with the Search function, additional metadata can be downloaded by clicking on the Download Expanded Results button. There are a few more clusters of interest, which we can find by filtering the table by the expected CDR sequences (in the upper right corner). Note that some of these don't conform to the motif definition because they are longer:



Clustering results for chain A

Download summary (TSV) Download summary (XLS) Search: SIFNT NYGGSQGNLI

representative	cluster size	CDR1	CDR2	CDR3
COVID19_Bacher-2020_1_8	766	SIFNT	LYKAGEL	AGQNYGGSQGNLI
COVID19_Bacher-2020_1_4292	136	SIFNT	LYKAGEL	AALNYGGSQGNLI
COVID19_Bieberich-2021_0_11490	18	SIFNT	LYKAGEL	AGQRNYGGSQGNLI
COVID19_Bieberich-2021_5_19327	6	SIFNT	LYKAGEL	VSMNYGGSQGNLI
COVID19_Notarbartolo-2021_0_10384	4	SIFNT	LYKAGEL	AGQAMNYGGSQGNLI
COVID19_Notarbartolo-2021_5_13103	4	SIFNT	LYKAGEL	AGHPSNYGGSQGNLI
COVID19_Notarbartolo-2021_6_12569	4	SIFNT	LYKAGEL	AGQPSDMNYGGSQGNLI
COVID19_Bieberich-2021_5_7434	2	SIFNT	LYKAGEL	AGLLLNYGGSQGNLI

Showing 1 to 8 of 8 entries (filtered from 33,305 total entries) Previous 1 Next

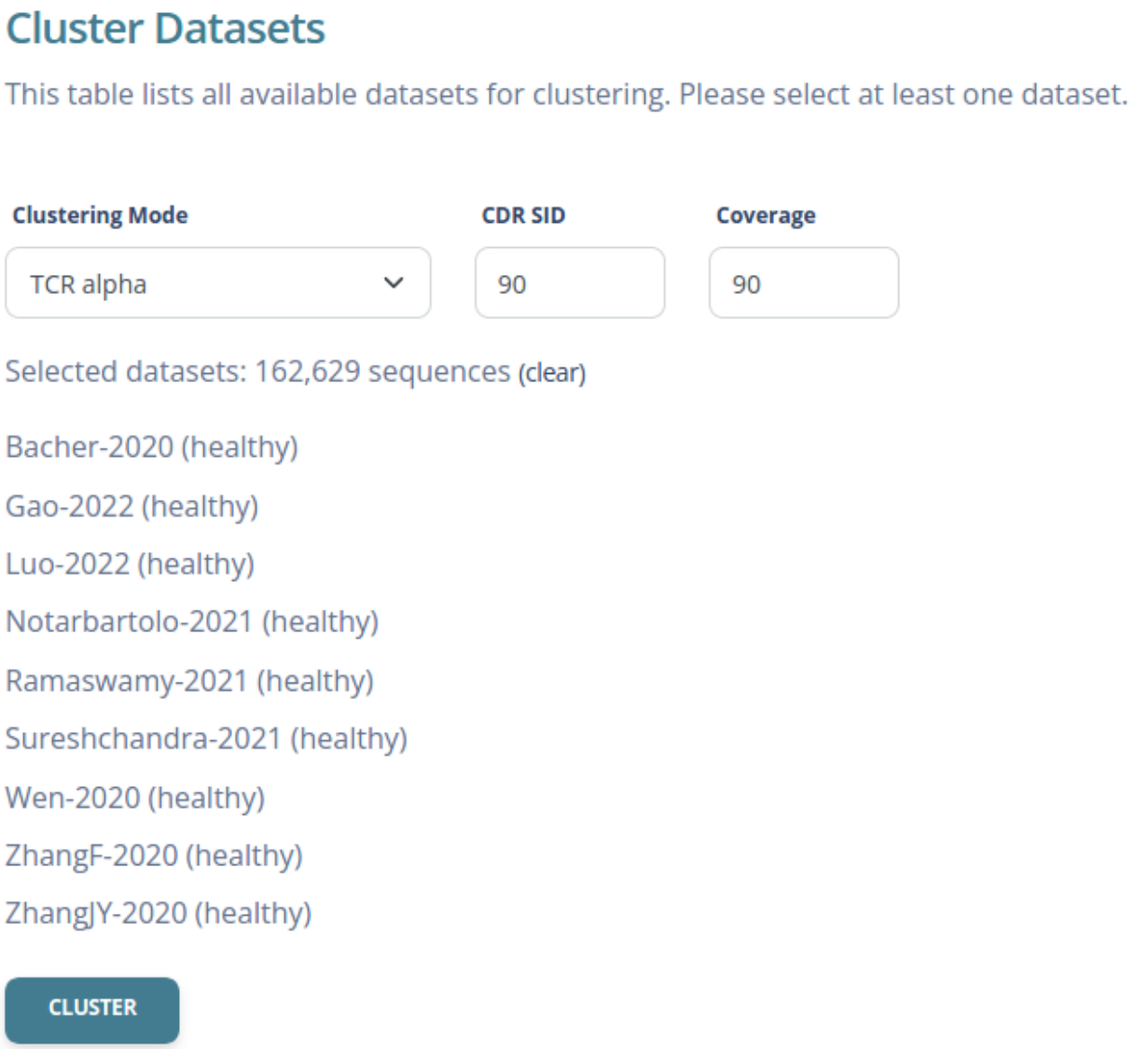
7.2 Clustering healthy data

For comparison, let's also have a look at healthy (pre-pandemic) data. Alternatively, select the following datasets on the Cluster page:

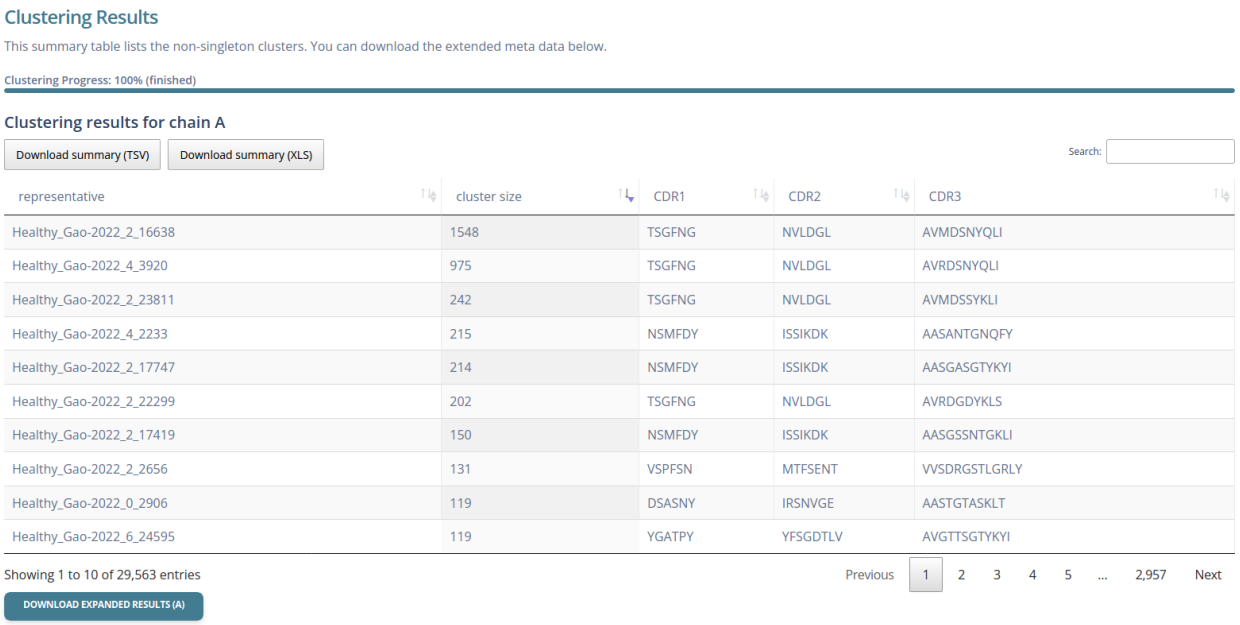
- Bacher-2020
- Gao-2022
- Luo-2022
- Notarbartolo-2021
- Ramaswamy-2021

- Sureshchandra-2021
- Wen-2020
- ZhangF-2020
- ZhangJY-2020

The input form should look like this:



After a few minutes, the following results should appear:



Clustering Results

This summary table lists the non-singleton clusters. You can download the extended meta data below.

Clustering Progress: 100% (finished)

Clustering results for chain A

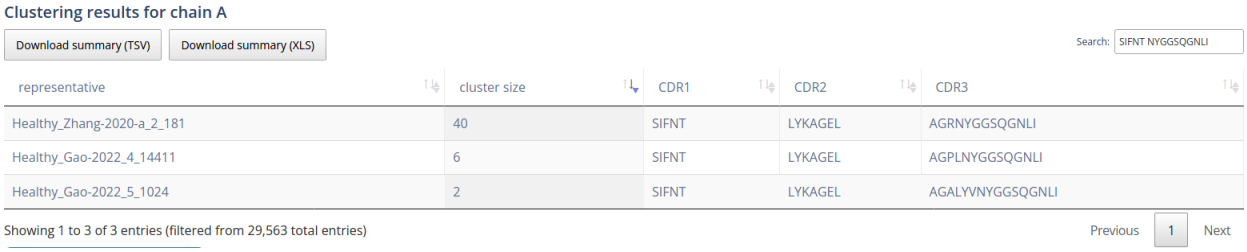
Download summary (TSV) Download summary (XLS) Search:

representative	cluster size	CDR1	CDR2	CDR3
Healthy_Gao-2022_2_16638	1548	TSGFNG	NVLDGL	AVMDSNYQLI
Healthy_Gao-2022_4_3920	975	TSGFNG	NVLDGL	AVRDSNYQLI
Healthy_Gao-2022_2_23811	242	TSGFNG	NVLDGL	AVMDSSYKLI
Healthy_Gao-2022_4_2233	215	NSMFDY	ISSIKDK	AASANTGNQFY
Healthy_Gao-2022_2_17747	214	NSMFDY	ISSIKDK	AASGASGTYKYI
Healthy_Gao-2022_2_22299	202	TSGFNG	NVLDGL	AVRDGDYKLS
Healthy_Gao-2022_2_17419	150	NSMFDY	ISSIKDK	AASGSSNTGKLI
Healthy_Gao-2022_2_2656	131	VSPFSN	MTFSENT	VVSDRGSTLGRLY
Healthy_Gao-2022_0_2906	119	DSASNY	IRSNVGE	AASTGTASKLT
Healthy_Gao-2022_6_24595	119	YGATPY	YFSGDTLV	AVGTTSGTYKYI

Showing 1 to 10 of 29,563 entries Previous 1 2 3 4 5 ... 2,957 Next

DOWNLOAD EXPANDED RESULTS (A)

Again, we see that most clusters have invariant receptors. This time, no major clusters exhibit the public Spike protein targeting motif. We can find some smaller ones, by filtering the table by the expected CDR sequences. As it turns out, only one cluster contains the correct sequence motif:



Clustering results for chain A

Download summary (TSV) Download summary (XLS) Search: SIFNT NYGGSQGNLI

representative	cluster size	CDR1	CDR2	CDR3
Healthy_Zhang-2020-a_2_181	40	SIFNT	LYKAGEL	AGRNYGGSQGNLI
Healthy_Gao-2022_4_14411	6	SIFNT	LYKAGEL	AGPLNYGGSQGNLI
Healthy_Gao-2022_5_1024	2	SIFNT	LYKAGEL	AGALYVNYGGSQGNLI

Showing 1 to 3 of 3 entries (filtered from 29,563 total entries) Previous 1 Next

All described functions are accessible without the need for a user account. However, additional features are available to those who register, like storing private datasets and reusing previous search queries. Registration is free and simple.

CHAPTER

EIGHT

REPORT AN ISSUE

This service is still under development. If you encounter any problems, please don't hesitate to contact us about them.