NEBNext Immune Sequencing data analysis user guide

pRESTO tool introduction

The NEBNext Immune Sequencing data analysis workflow is based on the pRESTO tool suite, the Repertoire Sequencing TOolkit. pRESTO performs all stages of raw sequence processing prior to alignment against reference germline sequences. pRESTO is flexible and customizable and is composed of multiple modules. This tutorial is meant to be concise and allow you to understand and run an example workflow that is configured for the standard NEBNext Immune Sequencing workflow. For a more detailed tutorial please consider reading the pRESTO tutorial.

In an example of an Illumina MiSeq paired-end 2x300 cycle run, Figure 1 (adapted from pResto documentatioin figures) shows the read schematic. Each read was sequenced from one end of the target cDNA so that the two reads together cover the entire variable region of BCR and TCR. The V(D)J reading frame proceeds from the start of read 2 to the start of read 1. Read 1 is in the opposite orientation (reverse complement), contains a partial C-region, and is 300 nucleotides in length. Read 2 contains the 5'RACE template switch site with a 17 nucleotide UMI barcode preceding it and is 300 nucleotides in length.

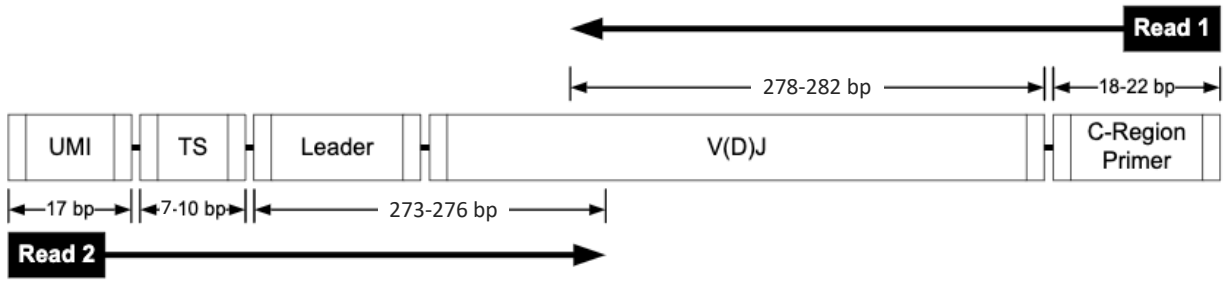


Figure 1. NEBNext Immune Sequencing library read schematic example of a run with Illumina MiSeq paired-end 2x300 reads.

pRESTO workflow overview

1. **Quality filter**: remove reads with mean Phred quality scores (Q) less than a user defined, sensitive value (set to 20%)
2. **Remove primers**: remove PCR primers and annotate reads with UMI (with maximum allowable error rate set to 20%)
3. **UMI consensus***: generate consensus sequences for each UMI barcode by multiple alignment
4. **Sequence assembly**: assemble paired-reads on the UMI consensus into a long sequence (by default, when the reads do not intercept each other, the full fragment can be inferred using IgBlast)
5. **Filters**: quality-filtersequences
6. **Final repertoire**: obtain the final repertoire output that can be used for VDJ alignment tools like IgBlast. Only unique sequences with more than *N* (set to 2) representative reads are used in downstream analysis.

*Troubleshooting consensus generation
Workflow parameters have been selected based on our experience with Illumina MiSeq instruments and V3 reagent kits. Since each experiment accumulates experimental errors differently, the parameters may require tuning based on the following factors.

- Filter reads if there is a high rate of mismatches between UMIs in a UMI consensus (**maxerror parameter**).
- Filter reads that do not share common primers within a UMI consensus group (**prcons parameter**).
- Align C-regions to a C-InternalRegions.fasta file, filtering for no more than 30% errors and limited to 100 nucleotides.
- The annotation specifying the number of raw reads used to build each sequence, is updated to the minimum of Fwd and Rev reads.
- Duplicate sequences, sharing the same constant region, are removed.
- Sequences with more than n (20) "N" are also removed.

For more UMI consensus troubleshooting details, please go to the pRESTO tool webpage for UMI consensus or the following reference.

Dysregulation of B Cell Repertoire Formation in Myasthenia Gravis Patients Revealed through Deep Sequencing.
Vander Heiden JA, et al.
J Immunol. 2017 198(4):1460-1473. doi:10.4049/jimmunol.1601415.

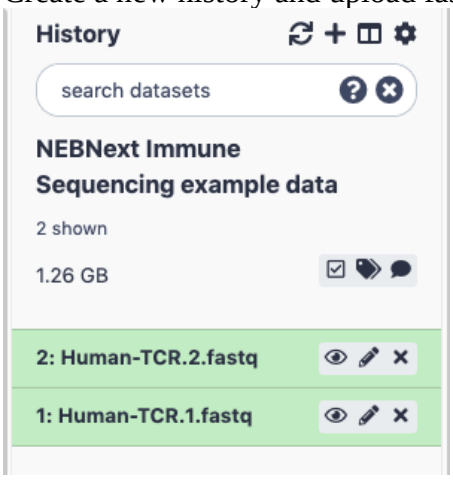
How to use the pRESTO workflow via Galaxy

NEB has implemented a pRESTO workflow on Galaxy for an easy start to users of immune sequencing data analysis. Galaxy is an open source, web-based platform for data intensive biomedical research. New users to Galaxy can follow the Galaxy Tours to start learning the platform. The pRESTO workflow published by NEB on Galaxy can be found here.

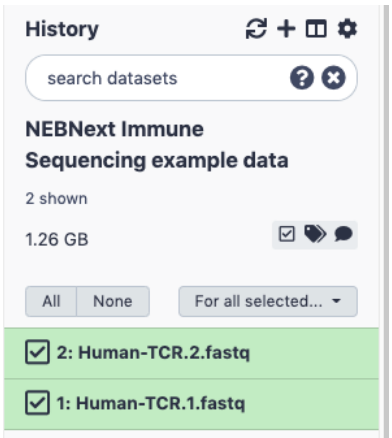
1. Import files before running pRESTO workflow

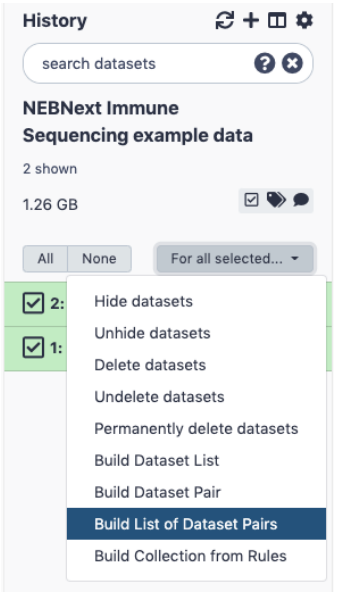
   Input files for running pRESTO and example dataset can be found at an example history published by NEB. Follow the steps below to import your own files.

   1.1 Create a new history and upload fastq files into the history.

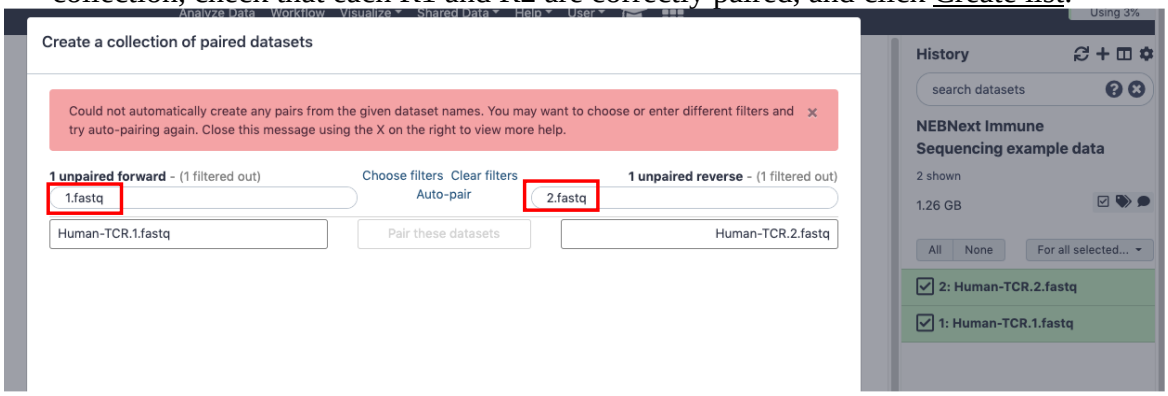


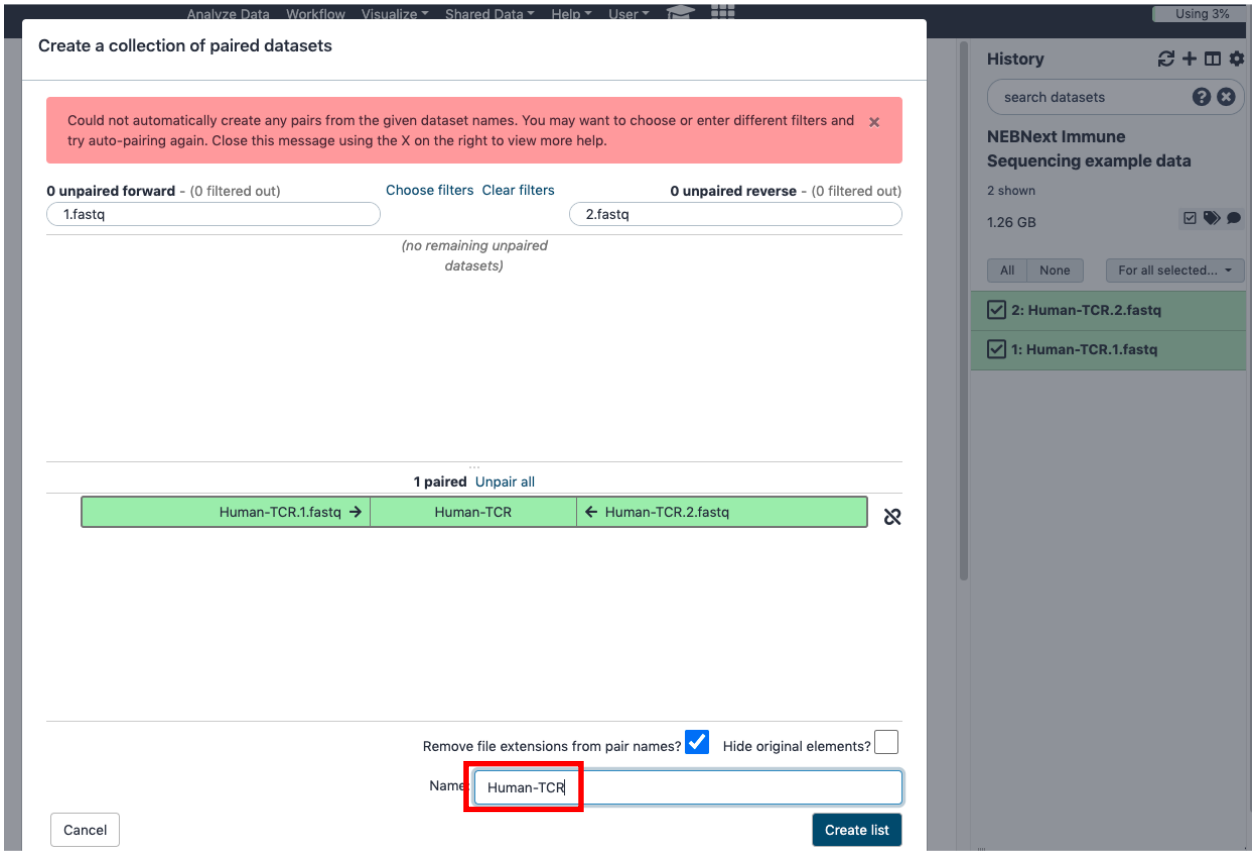
   1.2 Select all of the fastq files using the check button, click the "For all selected…" button and choose Build List of Dataset Pairs.

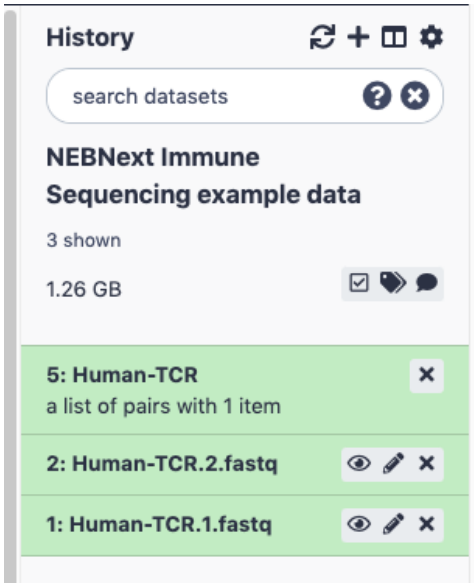




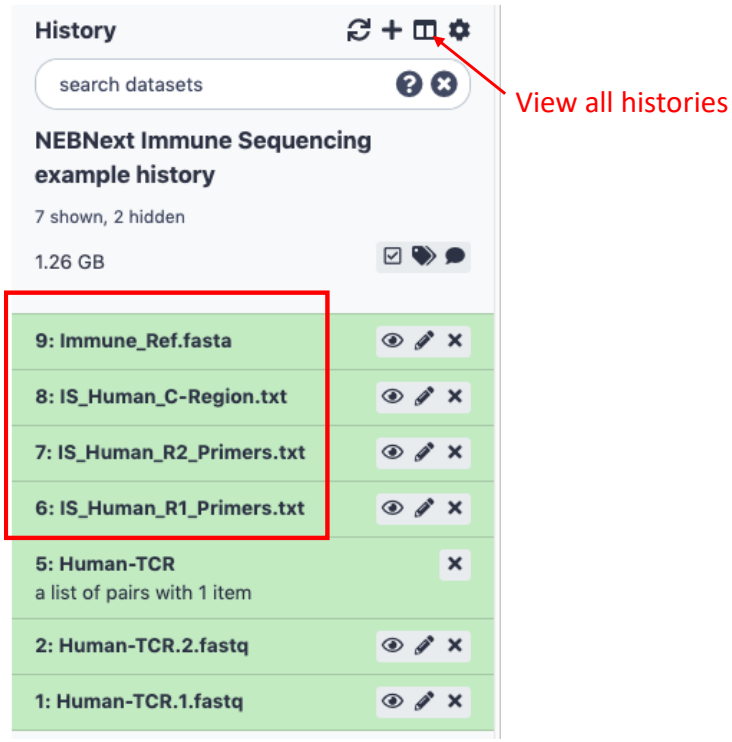
1.3 Enter the filters for forward and reverse fastq files, click Auto-pair, type the name of the collection, check that each R1 and R2 are correctly paired, and click Create list.

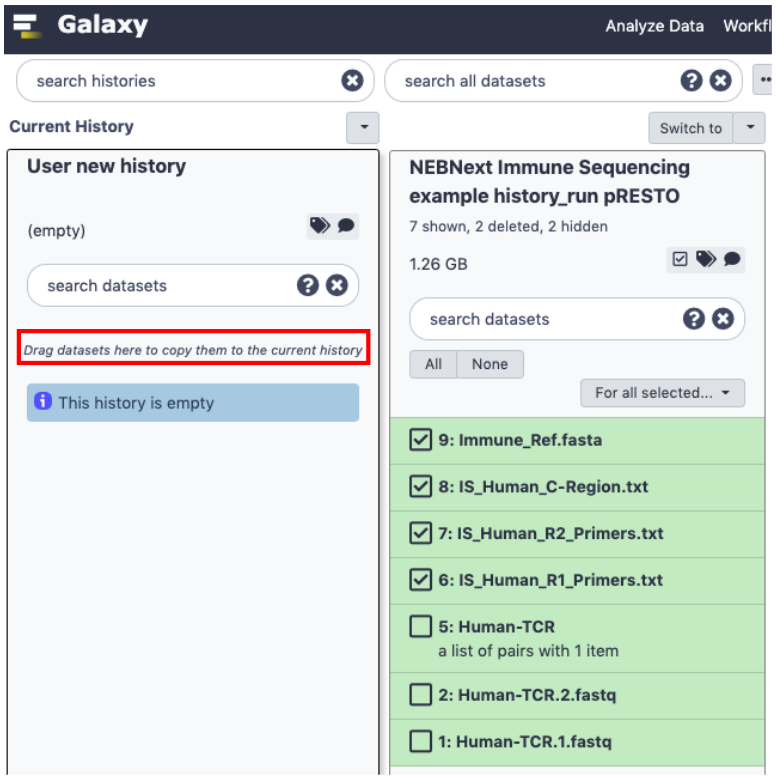




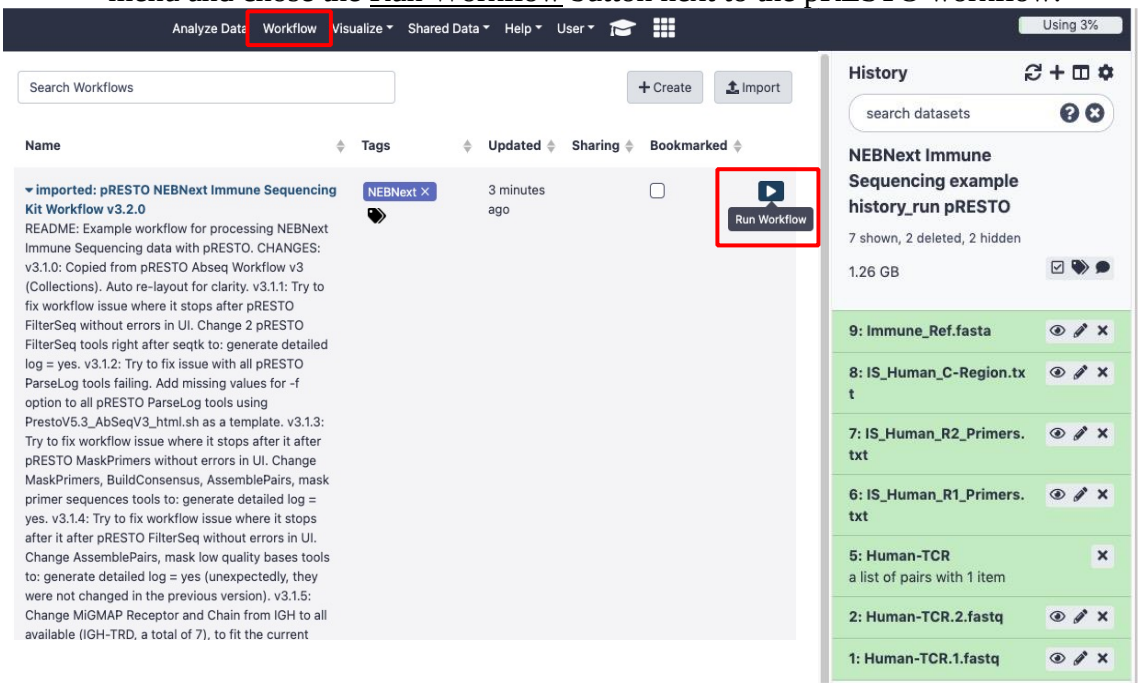


1.4 Import input files required to run pRESTO. The input files have been uploaded onto the example history as file number 6-9. The files can be copied to a new history or downloaded to local computer. To copy the files to a user history, click view all histories and click Switch to to make the destination history current, then select the files to be copied and drag them to the current history.





2. Run the pRESTO workflow
   2.1. Import the NEB pRESTO workflow to your Galaxy account.
   2.2. Activate the history containing your data and input files. Click Workflow on the top menu and chose the Run Workflow button next to the pRESTO workflow.

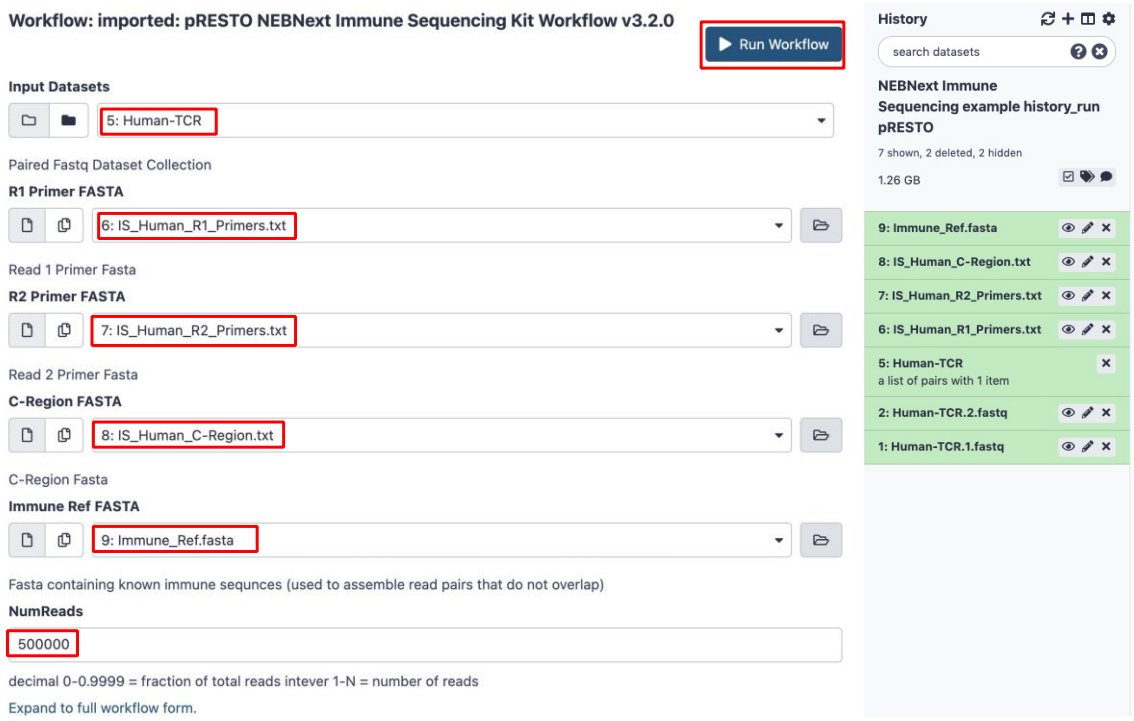


2.3. Run the pRESTO workflow

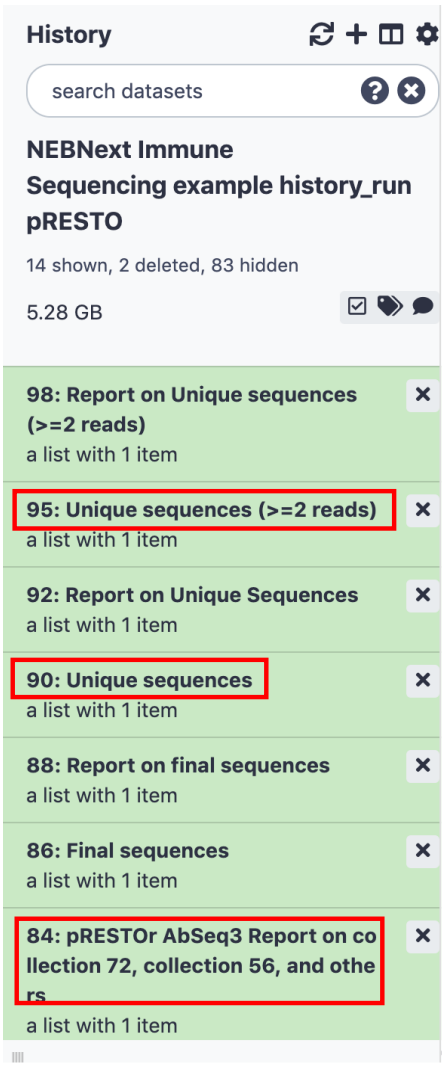
- Fill out Workflow Parameters and History Options

Num Pairs: number of paired end reads to include in the analysis. Consistent subsampling ensures that each library has the same power to detect transcripts but selecting a very large number will cause all available reads to be used if that is preferred. Example number: 500000.
Send results to a new history: select Yes or No
1: Input dataset collection: select the dataset collection in the history
2: R1 Primer FASTA: select input file IS_Human_R1_Primers in the history
3: R2 Primer FASTA: select input file IS_Human_R2_Primers in the history
4: C-Region FASTA: select input file IS_Human_C-Region in the history
5: Immune Ref FASTA: select input file Immune_Ref in the history

- Click the Run Workflow button.



3. QC report and output.
The sequencing reads QC metrics are visualized in the pRESTOr AbSeq3 Report.
For further reads alignment, the output files included in Unique sequences and Unique sequences (>=2 reads) can be used as input for V(D)J alignment tools, for example IgBlast.
An example history after running pRESTO is published here.


History
search datasets
NEBNext Immune Sequencing example history_run pRESTO
14 shown, 2 deleted, 83 hidden
5.28 GB
98: Report on Unique sequences (>=2 reads)
a list with 1 item
95: Unique sequences (>=2 reads)
a list with 1 item
92: Report on Unique Sequences
a list with 1 item
90: Unique sequences
a list with 1 item
88: Report on final sequences
a list with 1 item
86: Final sequences
a list with 1 item
84: pRESTOr AbSeq3 Report on collection 72, collection 56, and others
a list with 1 item