



Tutorial

Immune Repertoire Analysis using QIAseq Immune Repertoire panels

May 20, 2021

— Sample to Insight —

Immune Repertoire Analysis using QIAseq Immune Repertoire panels

This tutorial uses the capabilities of CLC Workbench and the Biomedical Genomics Analysis plugin to analyze sequences generated using the QIAseq Immune Repertoire RNA Library Kits.

In more detail, during this tutorial, we will:

- Make use of the Reference Data Manager to download reference data.
- Analyze QIAseq Immune Repertoire panel data, making use of the **Analyze QIAseq Panels** tool.

Prerequisites

For this tutorial, you must be working with *CLC Genomics Workbench 20.4* (or higher) with the Biomedical Genomics Analysis plugin 20.2 (or higher) installed. How to install plugins [in the CLC Genomics Workbench manual](#)

General tips

- Within wizard windows you can use the **Reset** button to change settings to their default values.
- You can access the in-built manual by clicking on **Help** buttons or by selecting the "Help" option under the "Help" menu.
- If you are connected to a *CLC Genomics Server* via your *CLC Genomics Workbench*, we recommend that you analyses on the *CLC Genomics Server* when you are offered this option.
- Information about running analyses in batch mode is available in the [the CLC Genomics Workbench manual](#)

Download and import data for this tutorial

For this tutorial we use data from a (not yet published) study of thymic B-cells role in shaping the T-cell receptor (TCR) repertoire in mice. The study uses transgenic mice, that have been genetically engineered to only express a single TCR β chain. Diversity in α : β T cells are therefore only generated by the α -chain.

Download the sample data

1. Download the sample data from our [website](#). The archive contains two samples from the experiment.
2. Unzip the data file to a location of you choice. Note that the files can also be automatically unzipped when imported into the CLC Workbench.

Import the sample reads

1. Start the CLC Workbench if it is not already running.
2. Import the sample reads, which are in CLC format, by going to:
File | Import (📁) | **Standard Import** (📁).
3. If you unzipped the data in the select the unzipped folders, otherwise select the zip-archive just downloaded. Leave the import type set to "Automatic", as shown in figure 1.

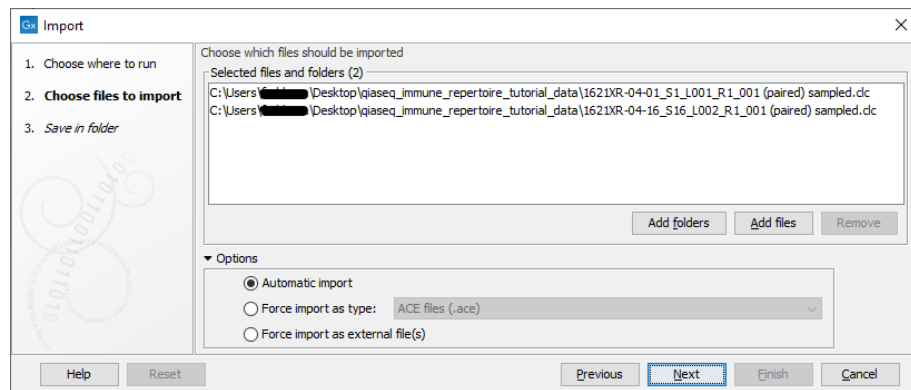


Figure 1: Standard import of folders containing the ".clc" files

4. Click on **Next** and select a folder to save to.
A new folder can be created by clicking on the "New Folder" option at the top of the window.

Download the reference data The "QIaseq Immune Repertoire Mouse" reference data sets are needed for the analyses we carry out in this tutorial.

1. Click on the **References** button in the top toolbar.
2. Select the **QIAGEN Sets** tab, and then select the **QIaseq Immune Repertoire Analysis** item under Reference Data Sets, on the left hand side.
3. Click on the **Download** button on the right, above the list of data elements (figure 2).
If the "Download" button is not enabled, and you see checkmarks by each data element listed, then you already have the reference data in this set.
4. Now select the **QIaseq Immune Repertoire Analysis Mouse** item under Reference Data sets, on the left hand side and download this data set as well.
5. When you are done, close the **Reference Data Manager**.

Run the Mouse Immune Repertoire workflow to analyze the samples

We now analyze the data from the mouse experiment using the Mouse Immune Repertoire workflow. We will launch the workflow in batch mode, such that the analysis will be run once for each sequence list supplied as input.

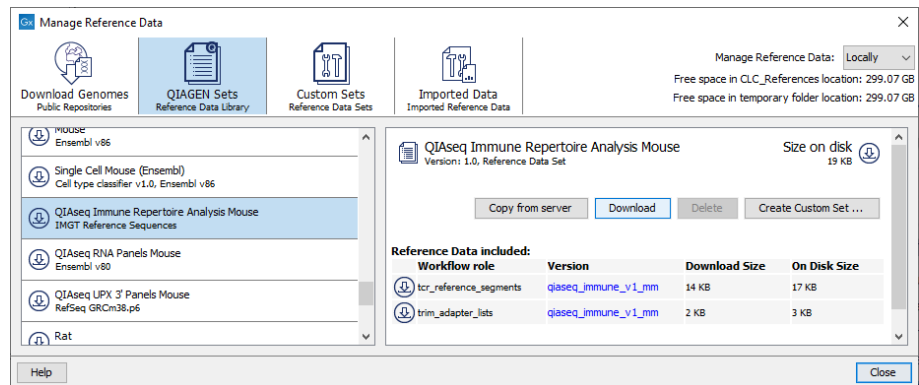




Figure 2: Reference data to be downloaded

For each sample, the workflow will produce QC related reports, an Immune Repertoire Analysis report, a Clonotype table and a Combined Report, containing a summary of information from the other reports.

Further information about this workflow can be found in the [Biomedical Genomics Analysis plugin manual](#).

1. Open the Analyze QIaseq Panels from the Toolbox:
Ready-to-Use Workflows | QIaseq Panel Analysis  | **Analyze QIaseq Panels** 
2. Open the **Immune** tab.
3. Click on **Run** next to the Mouse Immune Repertoire option (figure 3).

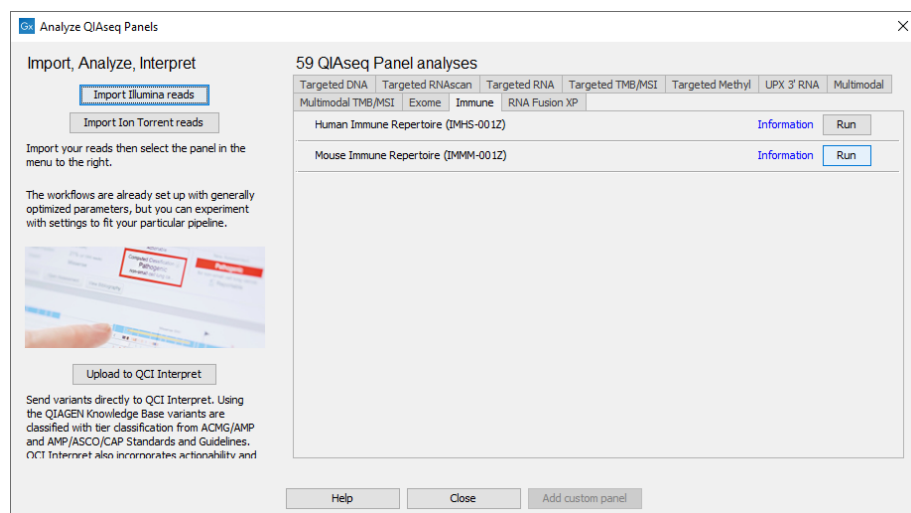


Figure 3: Run Mouse Immune Repertoire Analysis from Analyze QIaseq Panels

4. Select the sequences from the mouse experiment (figure 4).
5. Check the **Batch** option below the data selection area (figure 4).
6. Click on **Next**.
7. Specify a new folder to save the results to and click on **Finish**.

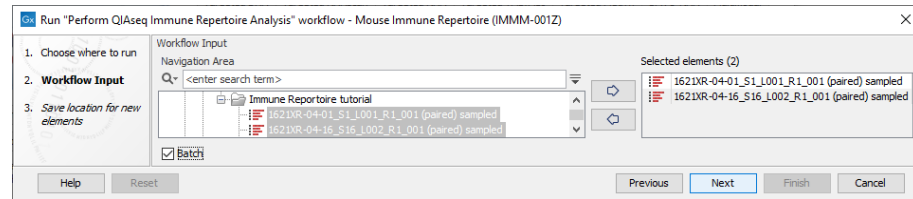


Figure 4: Reads to be analyzed in batch mode

Investigating the results of the Mouse Immune Repertoire workflow

The Perform QIAseq Immune Repertoire workflow will produce two top-level outputs:

- An immune repertoire report (📄) with a number of different summary statistics characterizing the immune repertoire.
- A clonotype table (📄) with a full list of identified clonotypes and their expressions.

The two outputs are described in detail below. The folder QC & Reports contains additional reports providing more detailed information for quality control.

Inspecting a single Immune Repertoire Analysis report

1. Click on a folder of results for one of the samples in the Navigation Area to expand its contents.
2. Open the Immune Repertoire Analysis report in that folder by double clicking on it.

In section 1 Summary, you will find a breakdown of the number of reads originating from each of the four chains, α , β , γ and δ . In the reports and tables the α , β , γ and δ -chains are denoted by TRA, TRB, TRG and TRD respectively to make it easier to search and filter for chains. For both mouse and human $\alpha:\beta$ T cells are much more common than $\gamma:\delta$ T cells in most tissues. It is therefore to be expected that the number of reads from the α and β chains is much larger than the number of reads from the δ and γ chains.

In section 3 Rarefaction, the so called rarefaction curve is shown for each of the four chains. The curve shows how many unique clonotypes would have been discovered for a given number of total sequenced reads originating from the chain. To begin with, when only a few clonotypes have been detected, there is a high chance that the next clonotype has not been seen before, so the curve rises steeply in the beginning. As more and more clonotypes are detected, it becomes more and more likely that the next clonotype has already been seen before, so the curve flattens. For a repertoire with high diversity, meaning that many different clones are present at modest frequencies, the curve will flatten slowly and asymptote at a high number of clonotypes. Whereas, for a repertoire with low diversity, meaning that there are a few dominating clones, the curve will flatten quickly and asymptote at a low number of clonotypes. This can be seen very clearly by contrasting the rarefaction curves for the α and β chains (figure 5). Since the β chain has been fixed diversity is low, even with 100,000 clonotypes sequenced only around 200 unique clonotypes have been identified. For the α chain around 3,500 unique clonotypes have been identified after clonotyping 10,000 reads.

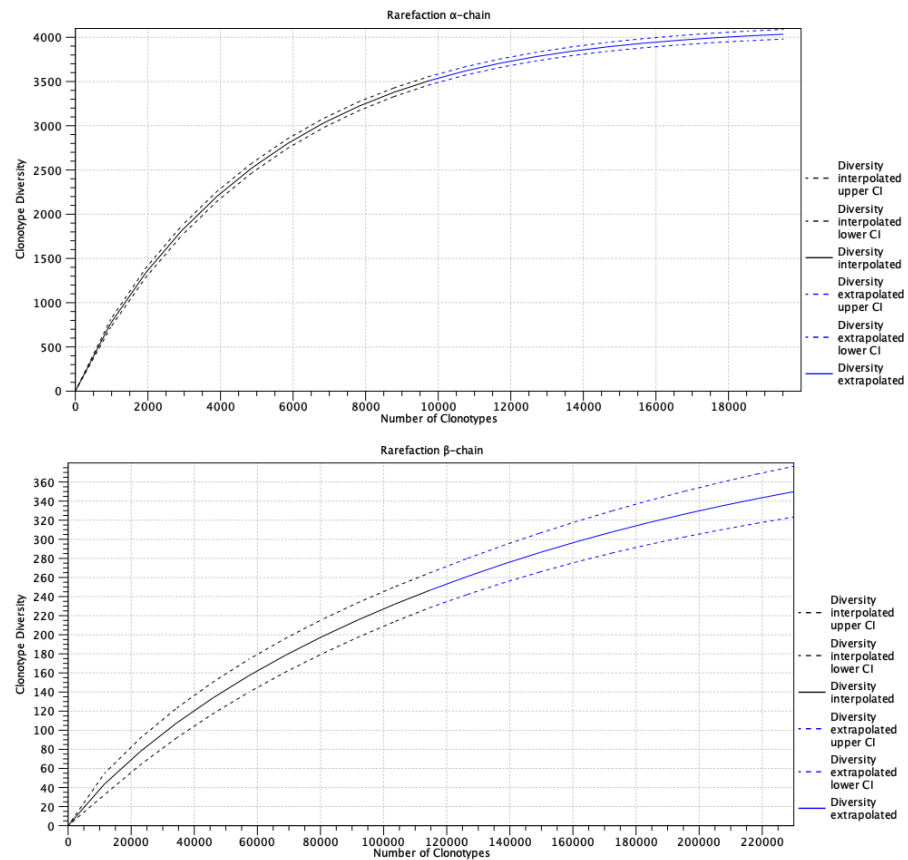


Figure 5: Rarefaction curves for the α and β chains

The diversity of a repertoire can be summarized using different metrics. In section 2 Diversity Indices, we can look at the Shannon-Wiener index, a high number means great diversity. Recapitulating the visual difference between the α and β rarefaction curve, we can see that the Shannon-Wiener index is close to zero for the β chain but more than 7 for the α chain.

Section 4 CDR3 length, shows histograms of the CDR3 lengths for each chain. We see characteristic peaks every 3nt, corresponding to in-frame TCR sequences (figure 6). It is common and no cause of concern that some CDR3 sequences are out-of-frame. In section 7 Productive summary, there are tables showing the percentage of clonotypes that are deemed un-productive, either because they have an out-of-frame CDR3 sequence or the CDR3 sequence introduces a premature stop codon.

Section 5 V and J usage, contains histograms grouped by chain and V and J, showing how often a gene segment is used. This will reveal if some gene segments are preferentially used.

Inspecting the clonotype table

1. Click on a folder of results for one of the samples in the Navigation Area to expand its contents.
2. Open the table in that folder with a name starting with "Clonotypes" by double clicking on it.

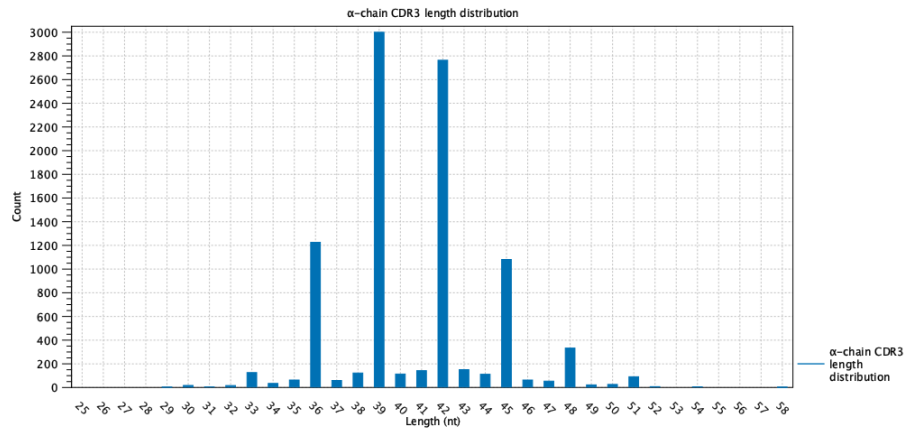


Figure 6: CDR3 length distribution for the α chain

The clonotype table contains all unique clonotypes identified along with their counts. A clonotype is characterized by its chain, V and J segments and its CDR3 sequence, which can be found in the first four columns of the table. The table also contains the amino acid translation of the CDR3 sequence. If the CDR3 sequence is out-of-frame this will be empty, if the CDR3 sequence contain a premature stop codon, it will be marked with an "*". The column Productive indicates whether the clonotype is productive or non-productive. Non-productive clonotypes have either out-of-frame CDR3 sequences or CDR3 sequences containing a premature stop codon.

By clicking on the "Count" table header twice we can sort the table by descending count, showing the most abundant clonotypes at the top. Again, since this repertoire is derived from a transgenic mouse, there is one clonotype from the β chain using V-19 and J-1-1 that is much more abundant than any other clonotype.

We can filter the table to only look at productive α chain clonotypes. The easiest way to filter to α chain clonotypes, is to right click on any cell in the Chain column contain "TRA", then navigate to "Table filters" and press "Chain = TRA" (Figure 7). We can do the same in the Productive column to filter out non-productive clonotypes.

Chain	V	J	CDR3 nucleotide sequence
TRB	V-19	J-1-1	TGTGCCAGCAGTATTCAGGG
TRB	V-19	J-2-7	TGTGCCAGCAGTATTCAGGG
TRA	V-19-1	J-1-1	TGTGCCAGCAGTATTCAGGG
TRA			GGGTAA
TRA			AGACTGG
TRA			GAATAG
TRA			CACAAA
TRA			CAATAC
TRA			GCAGGG
TRA			ACACAA
TRA			GTTATC
TRA		J-52	TGTGCAGCAAGTACTGGAGC
TRA		J-31	TGTGCTCTGAGTGATCAGGG
TRA		J-2	TGTGCAGCTAATACTGGAGG
TRA	V-19	J-58	TGCGCAGCAGGGCAAGGCA

Figure 7: Filter to show only α chain clonotypes

Comparing immune repertoires

To compare two or more immune repertoires we can use the Compare Immune Repertoires tool.

Run the Compare Immune Repertoires tool

1. To create a Combined Report, summarizing the QC results for each sample, go to: **Tools | QIAseq Panel Expert Tools** (🔧) | **QIAseq Immune Repertoire Expert Tools** (🔧) | **Compare Immune Repertoires** (👤)
2. Next select all the clonotype tables generated before. An easy way to select all the tables is by locating the folder with the results from the batch run. Right click the folder and press "Add folder contents (recursively)" (Figure 8).

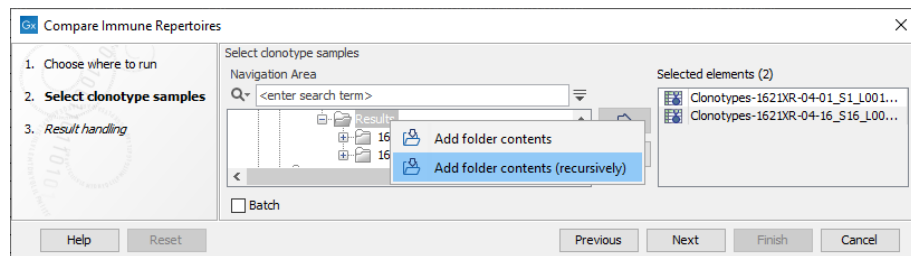


Figure 8: Select all clonotype tables

3. Click on **Next**.
4. In the output option check "Create similarity table". In the result handling chose "Save".
5. Click on **Next**.
6. Specify a new folder to save the results to and click on **Finish**.

The Compare Immune Repertoires report

1. Click on the folder of results for one of the samples in the Navigation Area to expand its contents.
2. Open the report in that folder by double clicking on it.

The first two sections contain tables where summary statistics and diversity indices can easily be compared. Such comparisons could also have been created using **Combine Reports** (📄). It is usually ill-advised to use "Observed diversity" directly to compare repertoire sizes, since this depends both on the number of reads clonotyped and the uniformity of the repertoire. *Extrapolated diversity (ChaoE)* is an estimate of the number of clonotypes we would find if we could completely saturate the repertoire [Chao, 1987]. *Interpolated to lowest sample diversity* is an estimate of the number of clonotypes we would have found in each sample if they all had the same number of reads clonotyped as the sample with the fewest. These estimates both have their advantages and drawbacks, but will make for a more "fair" comparison of the repertoire sizes, irrespective of the number of clonotyped reads.

The rarefaction curves in section 4 gives a complementary view of the repertoire diversity (Figure 9). A vertical red line has been added to the plot at the number of reads clonotyped from the α chain in the sample with the fewest clonotyped reads, here around 6,300. The intersections with the two rarefaction curves correspond to *Interpolated to lowest sample diversity*. The numbers the curves approach in the limit corresponds to *Extrapolated diversity (ChaoE)*.

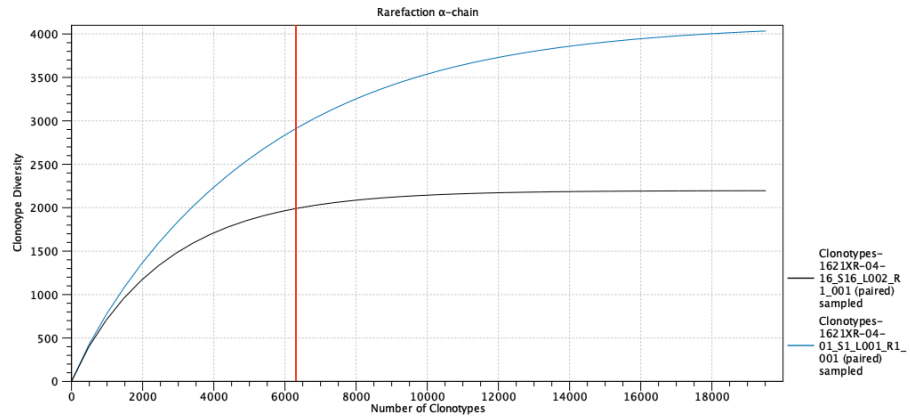


Figure 9: Rarefaction curves for the α chain. A vertical red line has been added to indicate the number of reads clonotyped in the sample with the fewest clonotyped reads.

Section 3 Scatter Plots, will contain scatter plots for each chain comparing the count of clonotypes in the two samples. This is useful for seeing if two repertoires have many common clonotypes (sometimes called "public") with similar expression or if the repertoires differ in their composition.

Section 6 V and J usage, can be used to compare the usage of different segments between repertoires (Figure 10). The plot will show if there are differential segment preference between the repertoires compared.

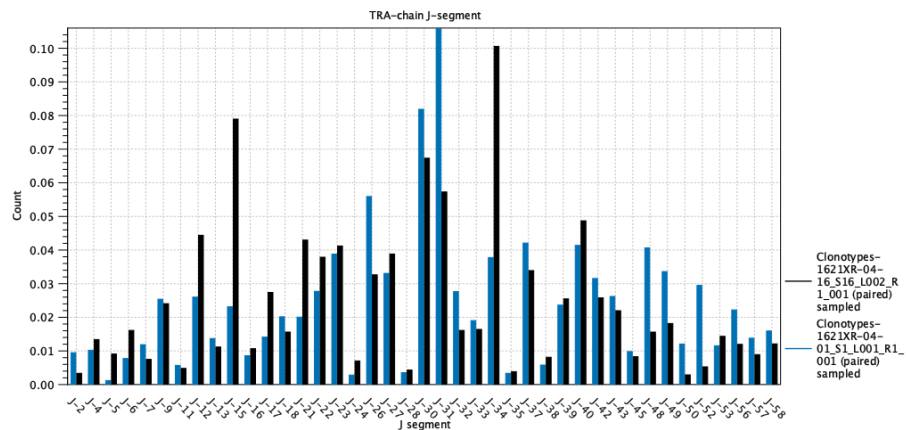


Figure 10: Comparison of α chain J-segment usage between the two samples.

Bibliography

[Chao, 1987] Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, pages 783–791.