VisTCR

Visual T Cell Receptor Sequencing Data Analysis Software

User Manual

Biomedical Analysis Center
31 Gaotanyan Street
Shapingba District, Chongqing, 400038
People's Republic of China

Telephone
+86-23-68752190

Website: http://core.tmmu.edu.cn/VisTCR/index.html

E-mail: wanying516@foxmail.com

Table of Contents

1. Introduction

VisTCR is an open source software that provides an interactive visualization of high-throughput TCR sequencing data, while also incorporating a friendly graphical user interface and a flexible workflow for data analysis. The software is a client-based HTML program written in ROR (Ruby on Rails) and Data-Driven Documents Javascript (D3.js)[1]. The major features of the software include:

1. **Independent modules for data management and analysis**. VisTCR consists of two modules; one to store data, and another to analyze data, denoted the "Data Storage Module" and "Data Analysis Module," respectively. This modular architecture allows the user to be versatile in their methods of data organization and re-organization, while allowing freedom to design various analysis strategies for a given data set.
2. **Freedom in grouping samples for individual analysis**. VisTCR includes an "Experiment Design File," which allows the user to deconstruct complex experimental designs into a combination of multiple variables. This design gives the user complete freedom in sample grouping and re-grouping for individual data analysis.
3. **Integration of multiple cutting-edge analysis algorithms**. VisTCR is capable of performing full TCR repertoire analysis, including calculations of distribution, diversity, similarity and clonotype tracking. All of these data analysis methods are organized in a hierarchical way and can cover multiple types of analyses, ranging from analyzing single samples, performing pairwise comparisons, and assessing the statistical merit of multiple samples.
4. **User-friendly interactive interface and data visualization**. Results from each individual's analysis can be visualized according to the user's preferences and can be generated and downloaded as tables, charts or graphs during multiple parts of the analysis workflow.

- ## Basic workflow

The workflow of the VisTCR software includes three steps:

1. **Uploading TCR sequencing data files in the Data Storage Module.** Raw FASTQ sequencing data files can be stored and organized in the 'Experiments' tab in this module.
2. **Creating an analysis project in the Data Analysis Module.** An Experiment Design File is created, allowing the user to select multiple variables used in the analysis of their project.
3. **Performing analysis and obtaining results in the Data Analysis Module.** Multiple TCR repertoire analysis methods are organized in a hierarchical way to cover several modes of analyses ranging from single sample features, pairwise comparisons and multi-sample statistical assessments.

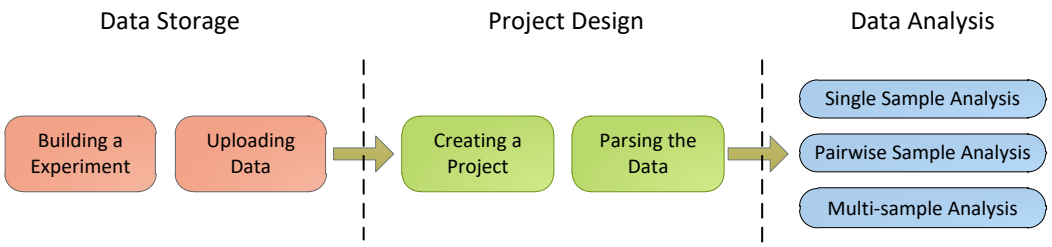


Figure 1.1 Basic workflow of the VisTCR software

2. Getting Started

2.1 Install dependencies

Before running VisTCR, you will need to install

- The Ruby language (version 2.0)
- Rails (version 3.2, See the article Installing Rails for detailed instructions and advice)
- R (R packages Rserve,Biostrings,seqinr,ShortRead,stringdist,gplots,ggplot2,vegan should be installed)
- Java (Version 1.8 or higher)
- python
- mysql

2.2 Getting VisTCR

You can download the code with the command

```
$git clone git://github.com/qingshanni/VisTCR.git
```

The source code is managed with Git. You'll need Git on your machine (install it from http://git-scm.com/).

To use mysql database, you'll need to modify the file **database.yml** to include your mysql password.

```
development:
        adapter: mysql2
        database: tcr1
        pool: 5
        host: localhost
        password: *
        encoding: utf8
        timeout: 5000
production:
        adapter: mysql2
        database: tcr1
        pool: 5
        host: localhost
        password: *
        encoding: utf8
timeout: 5000
```

Replace * by your mysql password.

To load required packages and codes when Rserve starts, RServe config file should be created. Create a file called Rserv.conf under /etc directory using vi or other text editor with the following contents:

```
workdir ***/tools/R
remote enable
fileio enable
interactive yes
port 6311
maxinbuf 262144
encoding utf8
control enable
source init_rserve.R
eval xx=1
```

Replace *** by the full path of vistcr directionary, such as /home/vistcr.

2.3 Run VisTCR

Start Rserve

```
$R CMD Rserve
```

Be sure you are in the vistcr directionary, and run the following commond.

Install the required gems on your computer:

```
$bundle install
```

Prepare the database and add the default user to the database by running the commands:

```
$ rake db:create
$ rake db:migrate
$ rake db:seed
```

Start the server

```
$rails s
```

Open a browser window and navigate to http://localhost:3000. You should see the log in page. You can sign in using:

- Email: user@example.com
- Password: changeme

2.4 Navigating the VisTCR Homepage

The VisTCR homepage consists of two tabs used to navigate between the Data Storage Module (Figure 2.5) and the Data Analysis Module (Figure 2.6).

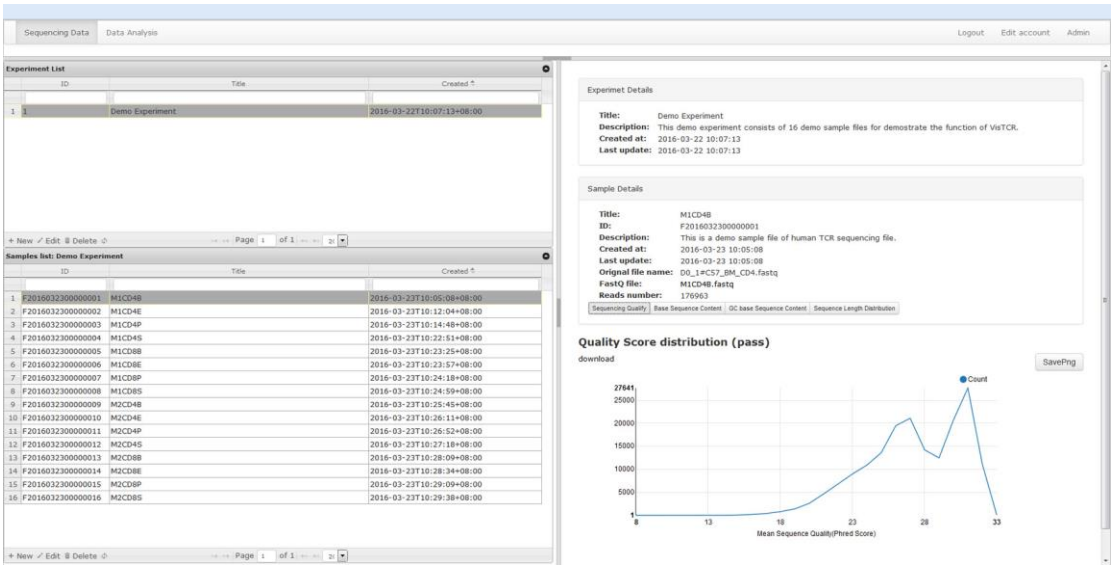


Figure 2.5 Screenshot of the Data Storage Module on the VisTCR homepage

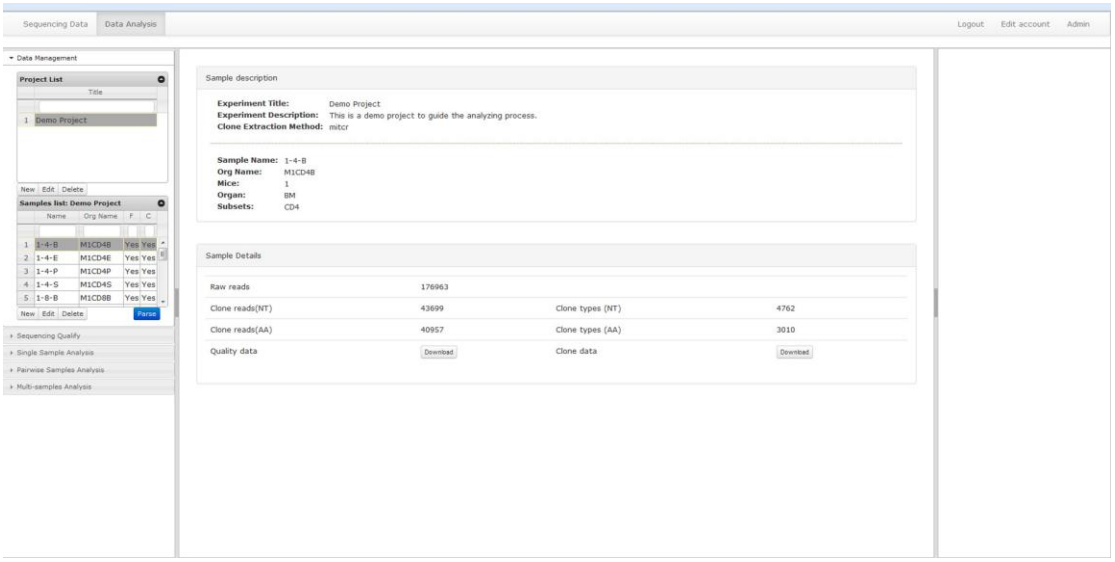


Figure 2.6 Screenshot of the Data Analysis Module on the VisTCR homepage

3. Data Storage Module

3.1 Overview

The Data Storage Module manages the raw TCR sequencing data (Figure 3.1). In the CONTROL PANEL, users can create and manage their experiments, assign sample IDs to TCR sequencing data files that have been uploaded, and check the quality of raw sequencing data. The INFORMATION LIST shows the details of the experiment and provides quality control checks on raw TCR sequencing data.

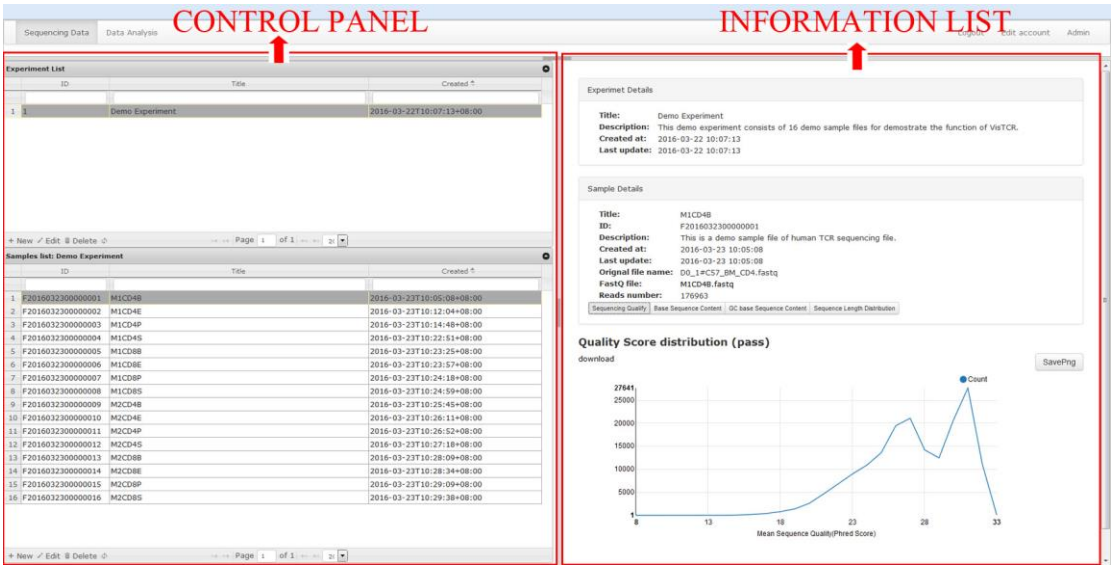


Figure 3.1 Screenshot of the Data Storage Module that displays the CONTROL PANEL and INFORMATION LIST sections.

3.2 Creating an Experiment

Users will organize their samples in the "Experiment List." Click the 'New' or 'Edit' buttons underneath the Experiment List to create, revise or remove an experiment (Figure 3.2). Multiple experiments can be uploaded in this section, allowing users to organize experiments however they like.

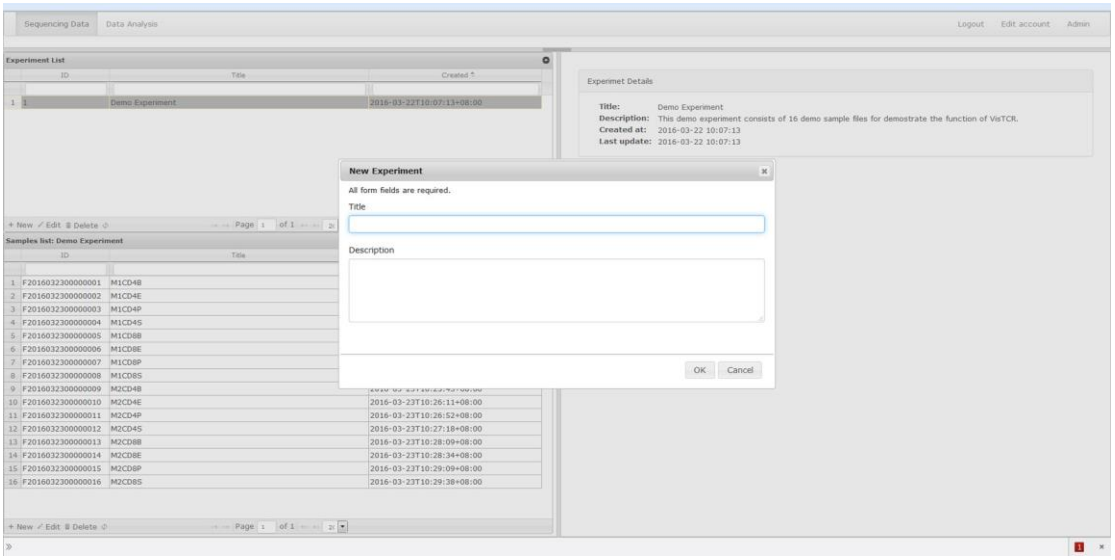


Figure 3.2 Screenshot the dialogue box used to create a new experiment in the Data Storage Module

3.3 Uploading TCR sequencing data

To upload TCR sequencing data files (in FASTQ format), click the “New” button underneath the “Sample List” table (Figure 3.3). Uploaded sample files are automatically stored in the selected experiment and will be assigned a unique sample ID. Each sample file needs to be uploaded individually.

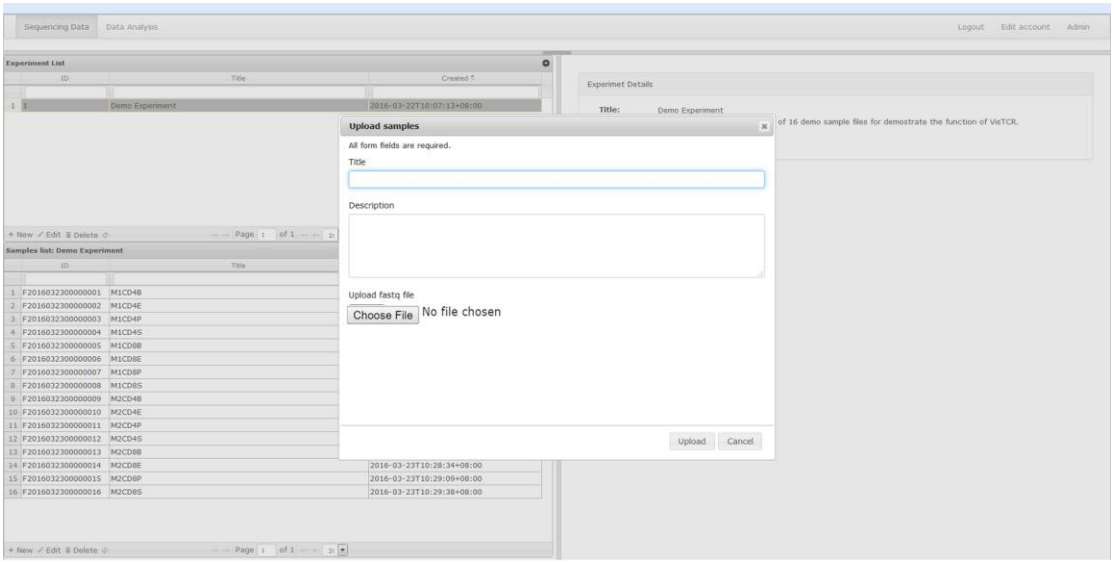


Figure 3.3 Screenshot of the dialogue box to upload samples in the Data Storage Module

3.4 Checking the quality of TCR sequencing data

The quality score distribution window provides quality control checks on the TCR sequencing data by using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The QC report of the selected TCR sequencing samples is shown in the INFORMATION LIST window (Figure 3.4). There are four QC analysis methods in VisTCR: Quality Score Distribution using the Phred score[2] (Figure 3.5), Base Sequence Content (Figure 3.6), GC Base Sequence Content (Figure 3.7), and Sequence Length Distribution (Figure 3.8).

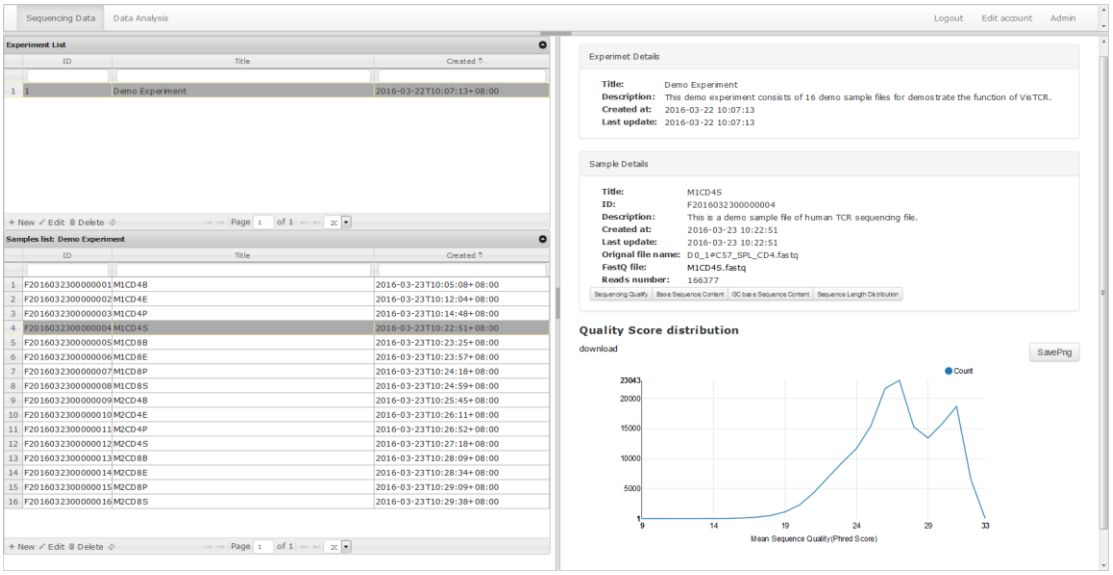


Figure 3.4 Screenshot of the QC report graph in the INFORMATION LIST window

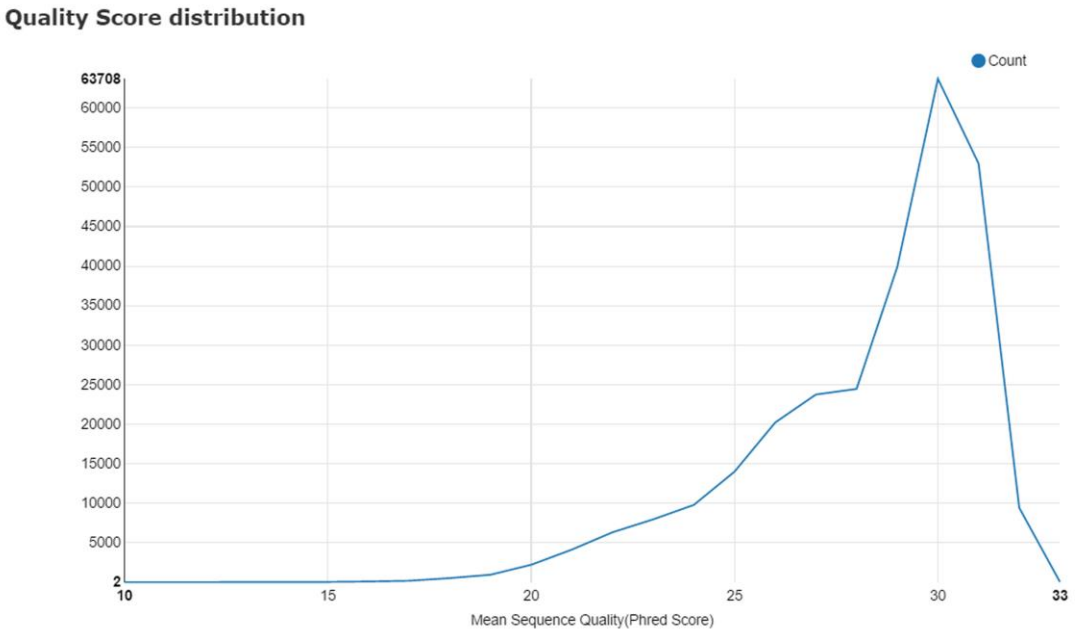


Figure 3.5 The Phred Quality Score distribution graph indicates if a subset of sequences has universally low sequencing quality values.

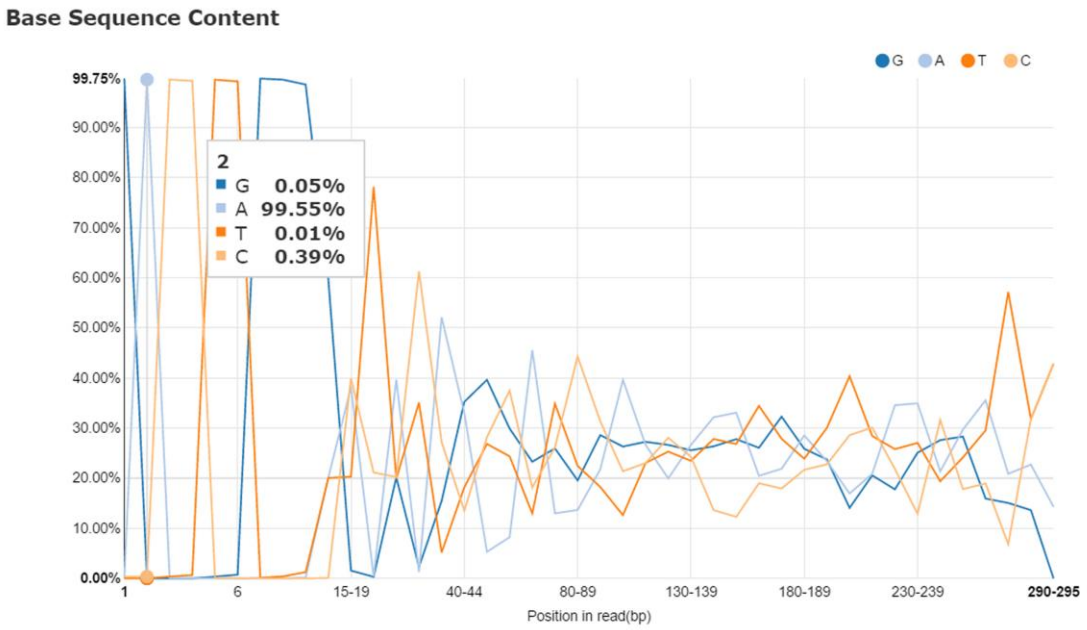


Figure 3.6 The Base Sequence Content graph shows the distribution of each base position in a selected TCR sequencing data file.

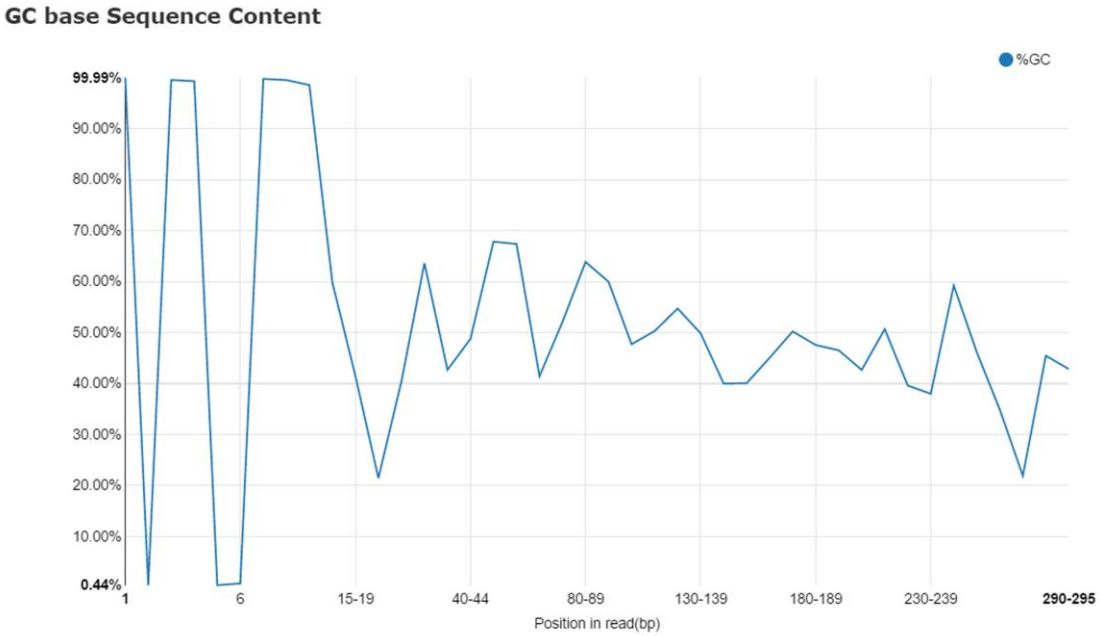


Figure 3.7 The GC Base Sequence Content graph shows the GC content across the length of each sequence in a selected TCR sequencing data file.

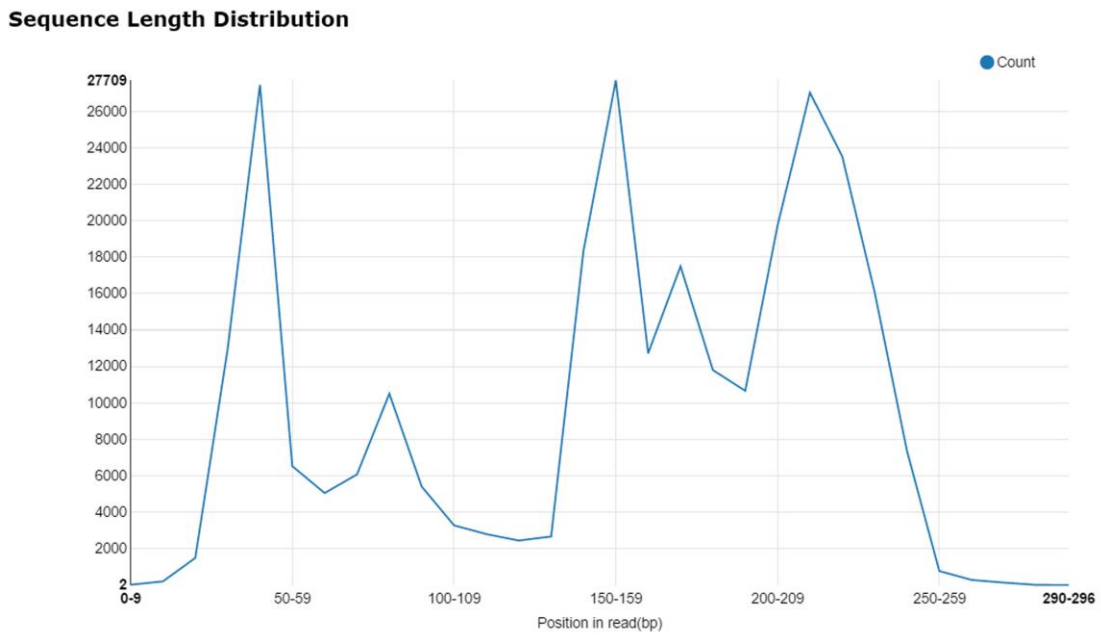


Figure 3.8 The Sequence Length Distribution graph indicates the distribution of fragment sizes in a selected TCR sequencing data file.

4. Data Analysis Module

4.1 Overview

In the Data Analysis Module, users can assign their TCR repertoire data samples to individual projects. A separate experiment design file is created per experiment, which defines the specific experimental conditions and any dependent variables or factors in a particular TCR repertoire data file. A unique sample ID is assigned to each individual sample and the data is combined in the experiment design file. This allows the user to select one of several analysis methods such as MiTCR[3], MiXCR[4], or Decombinator,[5] to parse through the raw TCR sequencing data. Once the clonotypes are extracted from each data file, the user can analyze their data using a variety of cutting-edge analysis methods which include features of single samples analysis, pair-wise sample comparisons, and multi-sample statistical analyses.

The workspace in the Data Analysis Module consists of three elements: a CONTROL PANEL, DISPLAY WINDOW, and INFORMATION LIST (Figure 4.1). Subsections of the CONTROL PANEL include: Data Management, Sequencing Quality, Single Sample Analysis, Pairwise Sample Analysis, and Multi-Sample Analysis. The DISPLAY WINDOW shows the results of the TCR repertoire analysis by using the JavaScript library D3.js, and allows users to interact with their data. The INFORMATION LIST provides details of the enumerated data results, and allows users to download their complete data in a CSV format file.

Figure 4.1 Screenshot of the workspace of Data Analysis Module consisting of the CONTROL PANEL, DISPLAY WINDOW and INFORMATION LIST

4.2 Creating a project for analysis

The control panel includes a list of several options that are visualized by a set of collapsible panels. In the Data Management subsection, users can assign their uploaded data files to individual projects and view sample details, including total raw reads and specific clone reads (both AA and NT). Users can download this data as a CSV file in this section (Figure 4.2).

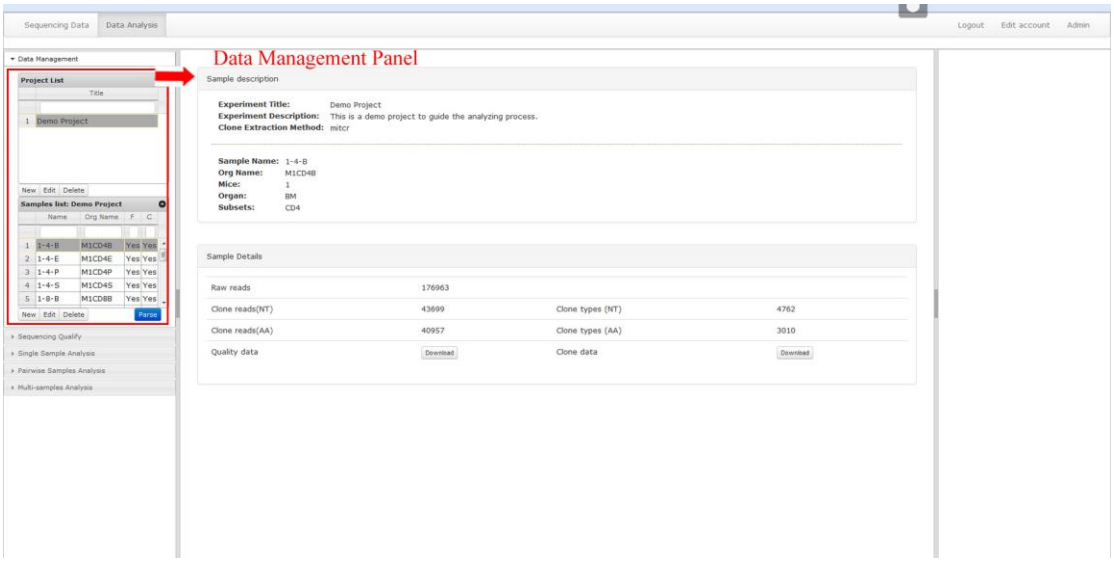
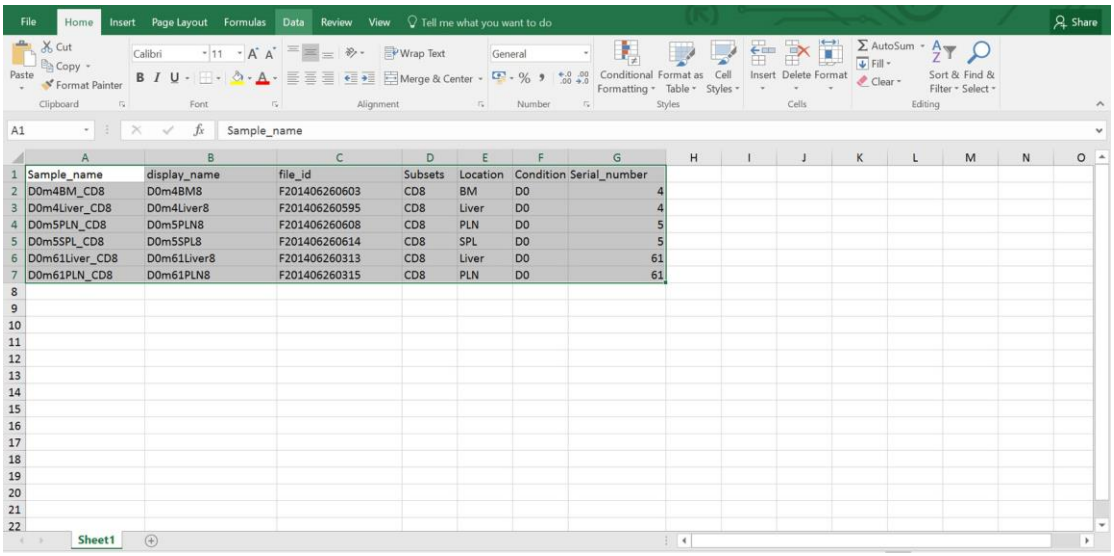


Figure 4.2 Screenshot of the Data Management subsection in the Data Analysis module

4.2.1 Creating an experiment design file

Before the clonotypes of each experimental file can be extracted for complete TCR analysis, an Experiment Design File must be created for each independent project. This is a CSV formatted file used to define the specific experimental conditions and any dependent variables or factors that can be used in TCR repertoire data analysis. Users can create this experiment design file in Microsoft® Office Excel (Figure 4.3.1) or a text editing software such as UltraEdit™ (Figure 4.3.2), as long as each file is saved in the CSV format.

The experiment design file must be created to meet the following specifications in order for the VisTCR software to recognize the items: the first three column names should be "sample_name," "display_name," and "fild_id." Please note that these three title names cannot be changed. In the "sample_name" column, users can describe basic information about the sample, using underscores (_) instead of spaces. The "display_name" column allows users to specify the sample name which will show up in future analysis in the Data Analysis Module. Finally, the "file_id" is the unique ID for each sample that is automatically created by the Data Storage Module once samples are uploaded. It is important to make sure that the file ID exactly matches the ID that shows up in the Data Storage Module once samples are uploaded. The subsequent column names can be defined according to each user's individual experimental conditions or other dependent variables or factors that can be used as parameters in the statistical analysis. For example, in the demo.csv file (Figure 4.3.1), the customized labels are "Subsets," "Location," "Condition," and "Serial number." In total, the VisTCR software can support a maximum of ten different column identifications.



	A	B	C	D	E	F	G
1	Sample_name	display_name	file_id	Subsets	Location	Condition	Serial_number
2	D0m4BM_CD8	D0m4BM8	F201406260603	CD8	BM	D0	4
3	D0m4Liver_CD8	D0m4Liver8	F201406260595	CD8	Liver	D0	4
4	D0m5PLN_CD8	D0m5PLN8	F201406260608	CD8	PLN	D0	5
5	D0m5SPL_CD8	D0m5SPL8	F201406260614	CD8	SPL	D0	5
6	D0m61Liver_CD8	D0m61Liver8	F201406260313	CD8	Liver	D0	61
7	D0m61PLN_CD8	D0m61PLN8	F201406260315	CD8	PLN	D0	61

Figure 4.3.1 Screenshot of the demo experiment design file in Microsoft® Excel.

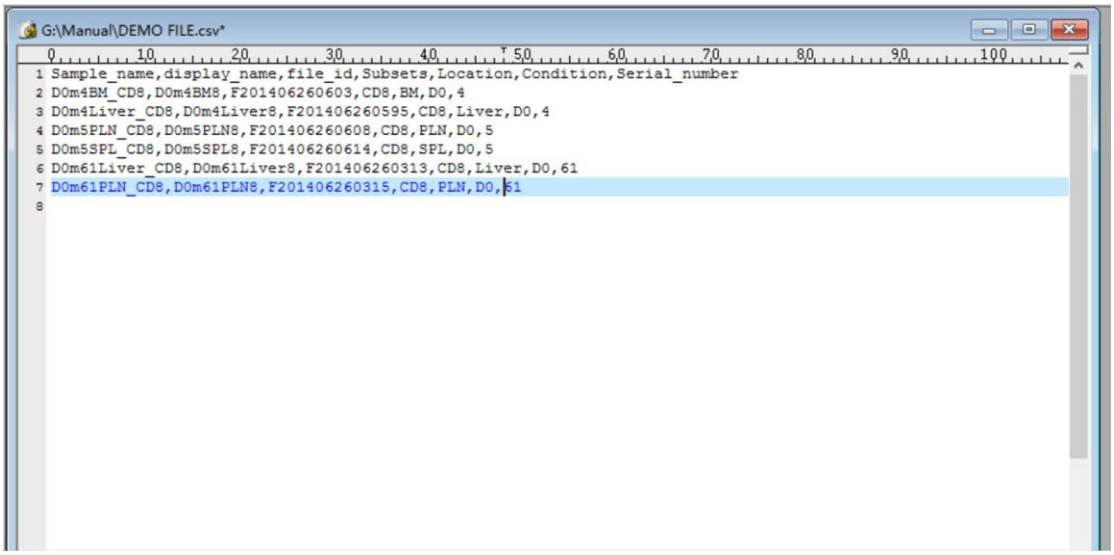


Figure 4.3.2 Screenshot of the demo experiment design file in UltraEdit™

4.2.2 Uploading the experiment design file

To upload the experiment design file, users must first create a new project. To do this, click the "New" button under the table that says "Project List" in the DATA MANAGEMENT panel. Input the appropriate title and project description in the dialogue box and click "Next" (Figure 4.4). In the following window, users can choose the TCR data analysis method of their choice ((MiTCR[3], MiXCR[4],or Decombinator[5],) and set additional parameters depending on their experiment (Figure 4.5). After clicking "Next," users can click the "Choose File" button to upload their Experiment Design File (Figure 4.6), remembering that the file MUST be saved in a CSV format for it to be read by the VisTCR software. After the project is created, users can click on their newly uploaded project in the Project List table to check individual project details (Figure 4.7)

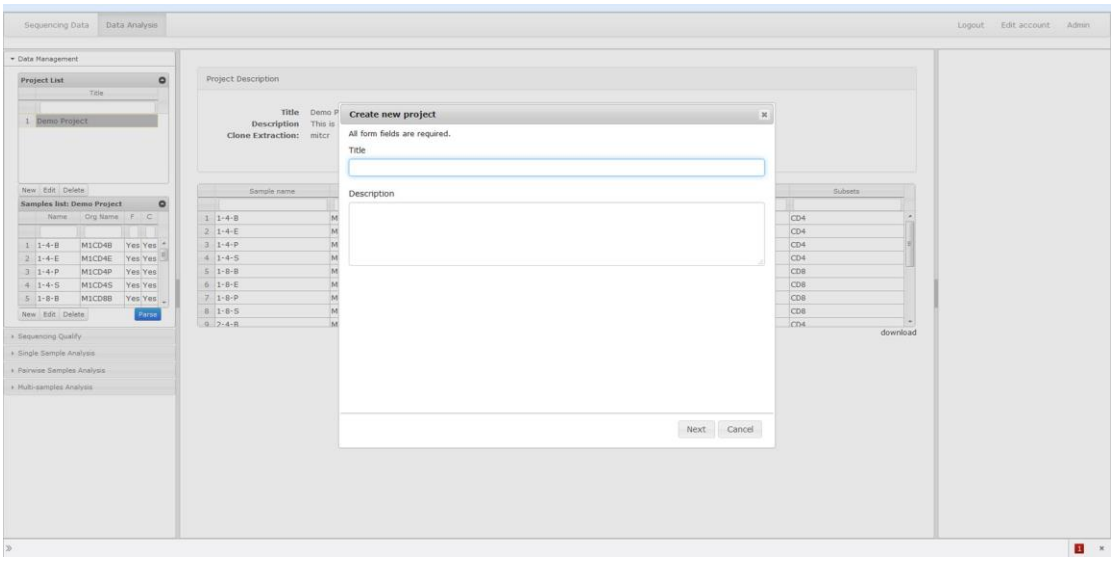


Figure 4.4 Screenshot of the first dialogue box allowing users to create a new project file.

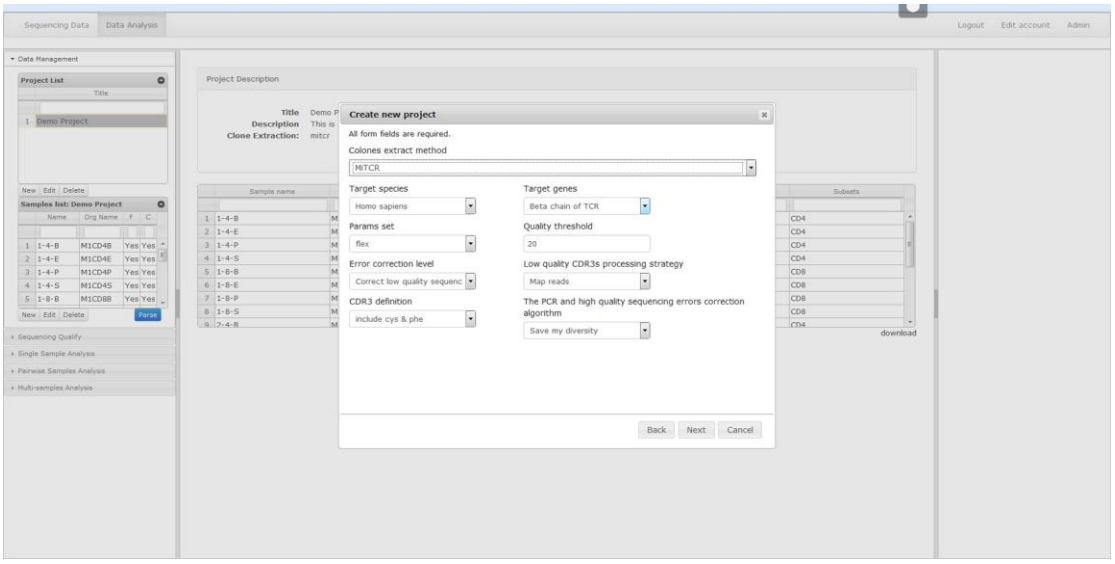


Figure 4.5 Screenshot of the second dialogue box allowing users to select appropriate methods and parameters for analysis.

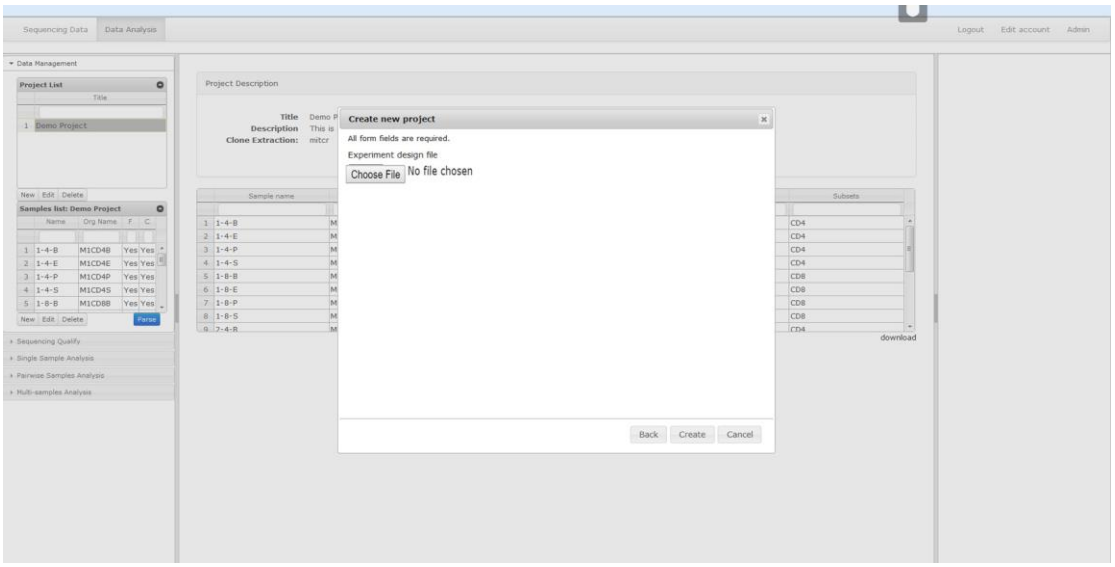


Figure 4.6 Screenshot of the third dialogue box for users to upload their Experiment Design File.

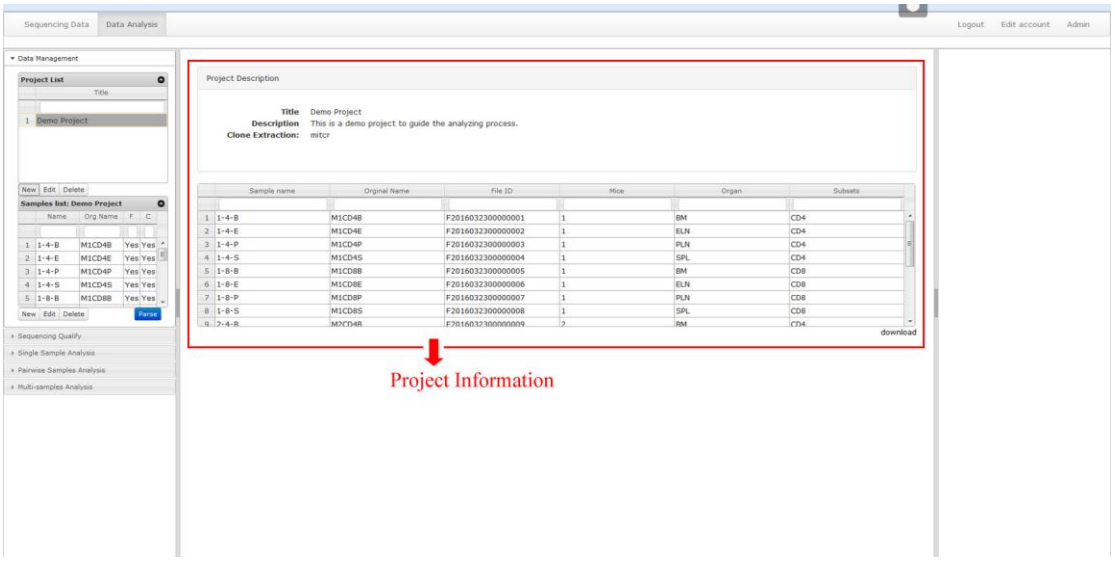


Figure 4.7 Screenshot of the “Project Description” window

4.2.3 Parsing TCR sequencing data

To parse the TCR sequencing data, an individual project needs to be selected. Users will select the project that they want analyzed from the “Project List” table, and click the blue “Parse” button under the sample list table (Figure 4.8). VisTCR will search the individual file ID of each sample from the Data Storage Module, the available raw sequencing data files will be parsed, and the TRBV, TRBJ and CDR3 regions will be recognized. Once the unique file ID has been found in the Data Storage Module, the status in the “F” column will switch from “No” to “Yes.” Once the samples have been successfully parsed, the status in the “C” column will also switch to “Yes.” If the sample ID cannot be found or the sample is not parsed successfully, the status will show up as “No” (Figure 4.8).

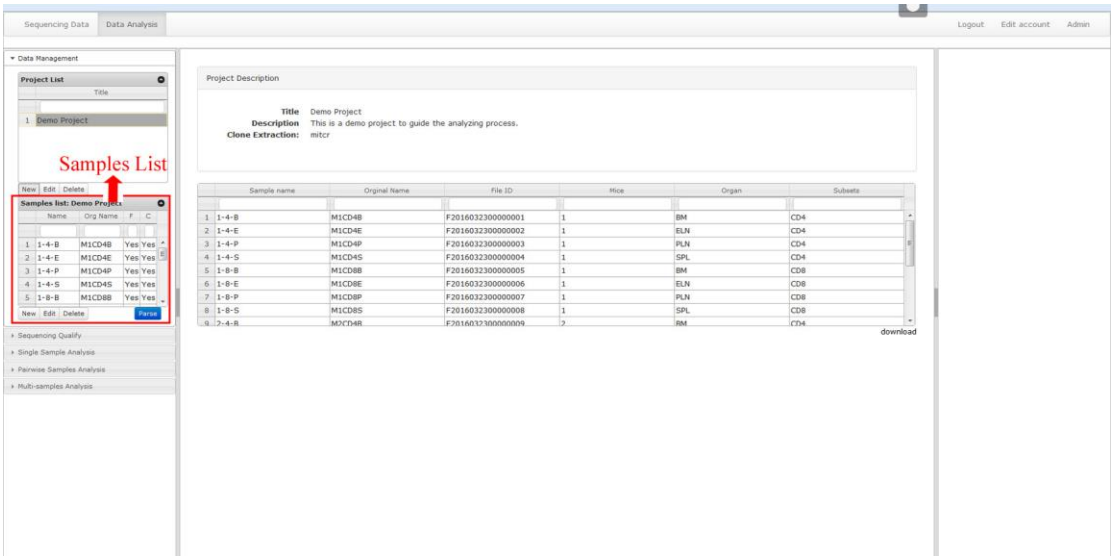


Figure 4.8 Screenshot of the sample list panel under the Data Management tab.

4.2.4 Editing the project

Individual sample parameters defined in the experiment design file can be revised. To revise samples, users can select a sample in the sample list and click the ‘Edit’ button to open the dialogue box (Figure 4.9). A new sample can be added into the experiment by clicking the ‘New’ button in the sample list (Figure 4.10).

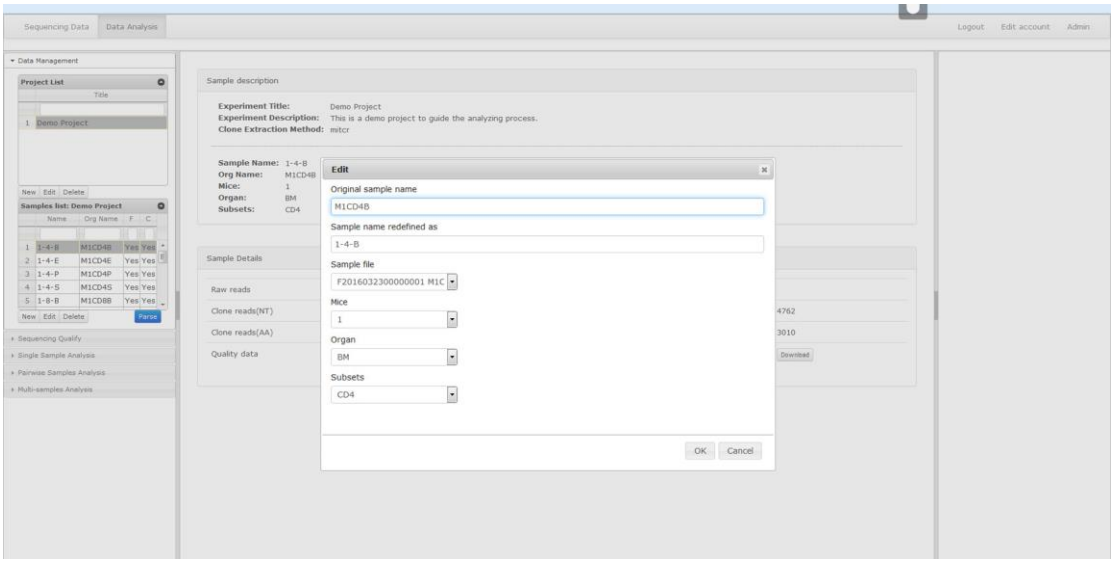


Figure 4.9 Screenshot of the dialogue box for editing sample parameters.

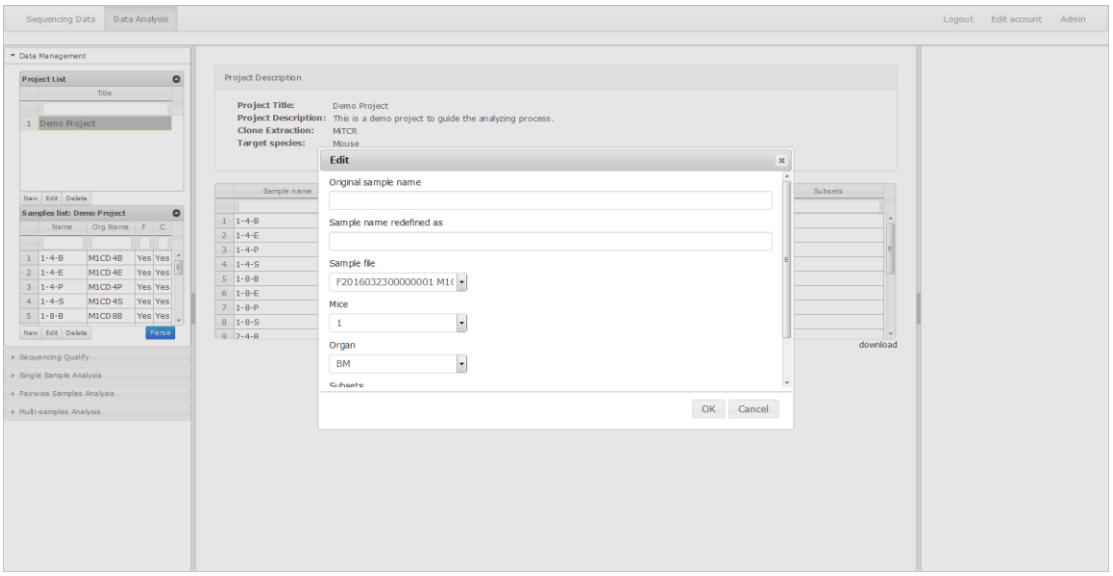


Figure 4.10 Screenshot of the dialogue box for adding a sample

4.2.5 Checking the quality of TCR sequencing data

The “Sequencing Quality” section of Data Analysis Module allows users to check the quality of their TCR sequencing data. For additional information on this, refer to Section 3.4.

4.3 Performing single sample analysis

In the single sample analysis section, the features of each individual sample are shown, including their TRBV and/or TRBJ usage, CDR3 spectratype, and their clonotype distribution. To navigate all the features of the single sample analysis, users can select a sample in the sample list table in the “Data Management” panel, and click the tab indicating “Single Sample Analysis.” Specific details of the data are shown in the INFORMATION LIST, and can be downloaded (Figure 4.11).

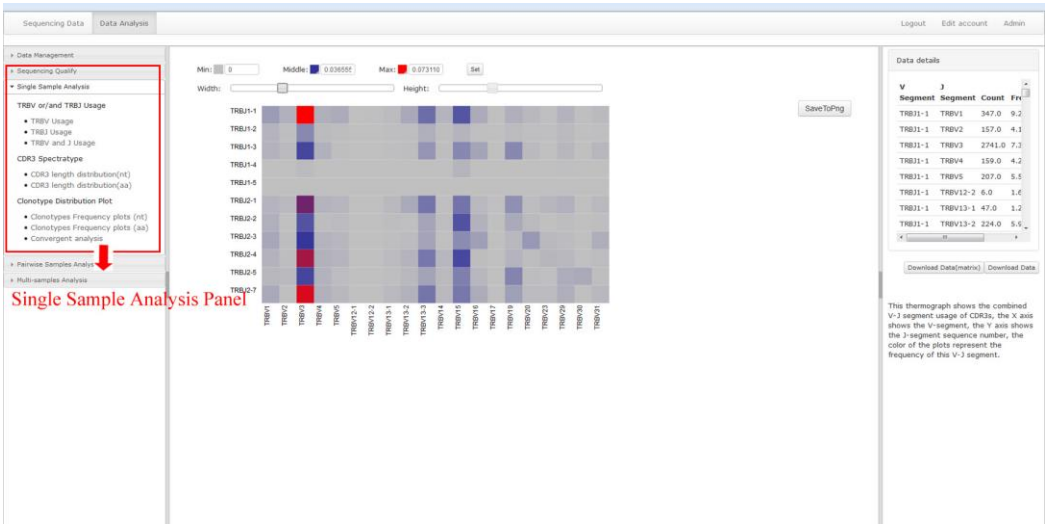


Figure 4.11 Screenshot of the single sample analysis panel which displays different methods for analysis.

- **TRBV and/or TRBJ Usage**

The TRBV and TRBJ usage of each sample is shown as a histogram in the DISPLAY WINDOW (Figure 4.12). In the INFORMATION LIST, users can find specific details of the TRBV and TRBJ usage. TRBV and TRBJ usage can also be shown as a heat map (Figure 4.13). Users can select individual boxes in the heat map to see specific details of the corresponding rows and columns. The size and colors of the heat map can be adjusted to the user’s preference (Figure 4.14).

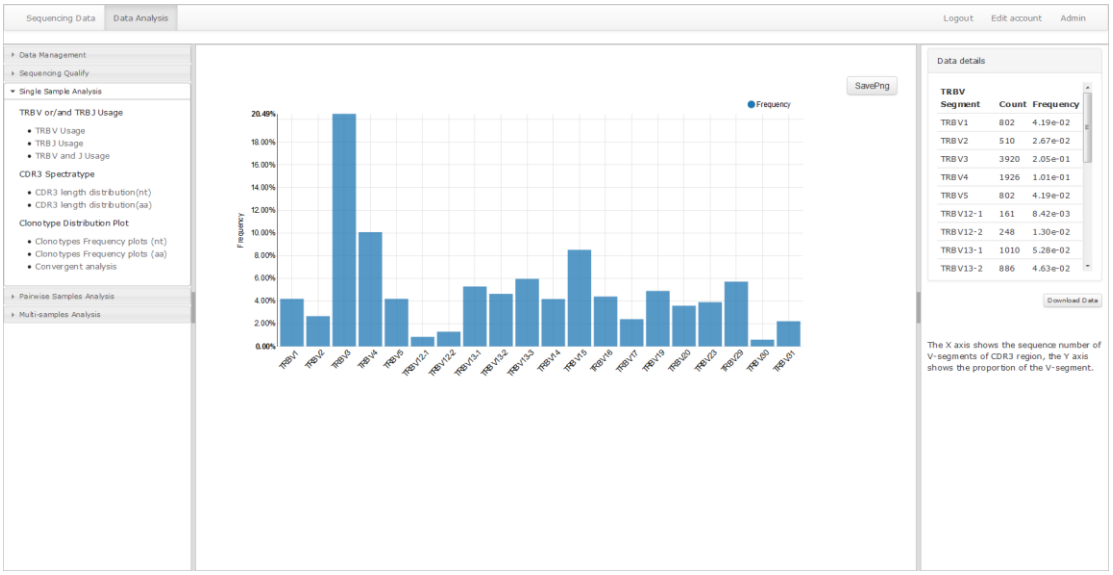


Figure 4.12 Screenshot showing a histogram of the TRBV and TRBJ usage.

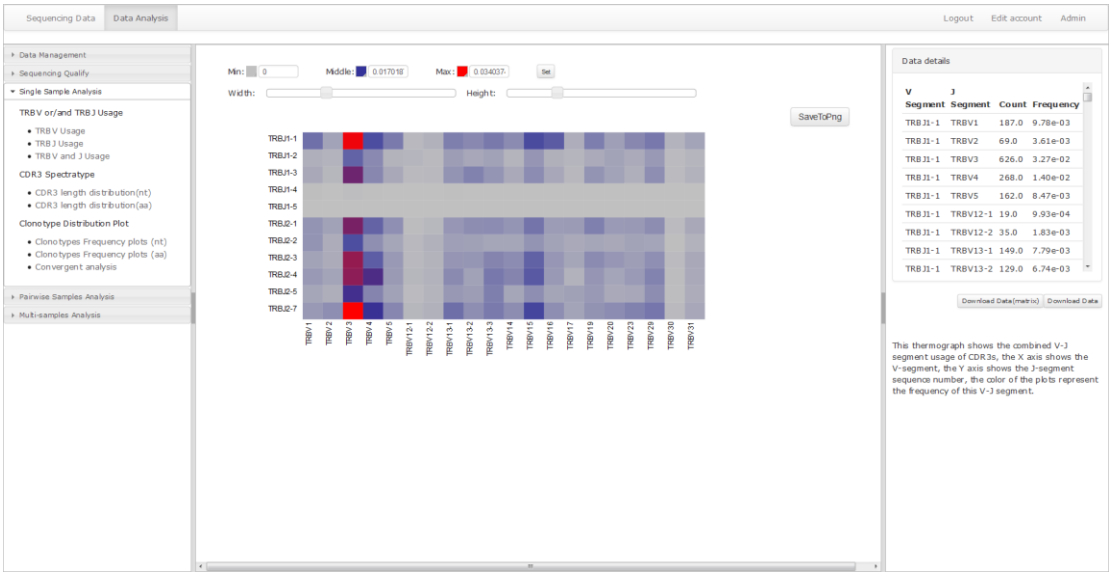


Figure 4.13 Screenshot showing a heat map of the TRBV and TRBJ usage.

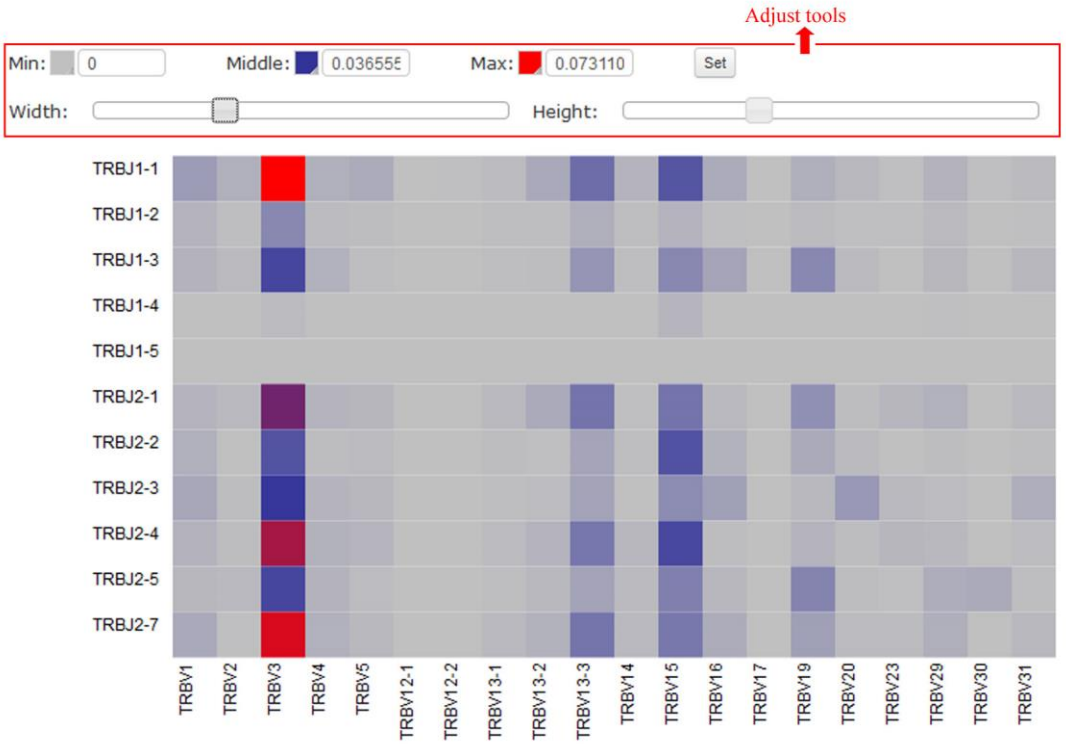


Figure 4.14 Tools to adjust the TRBV and TRVJ heat map. Users can change the size and color schemes of their heat map to their preference.

- **CDR3 Spectratype**

The CDR3 length distribution of either the nucleotide or amino acid sequences can be visualized by either stacked or grouped bars (Figure 4.15). To switch between these two visual outputs, click the “Grouped” or “Stacked” buttons located at the top of the graph (Figure 4.16).

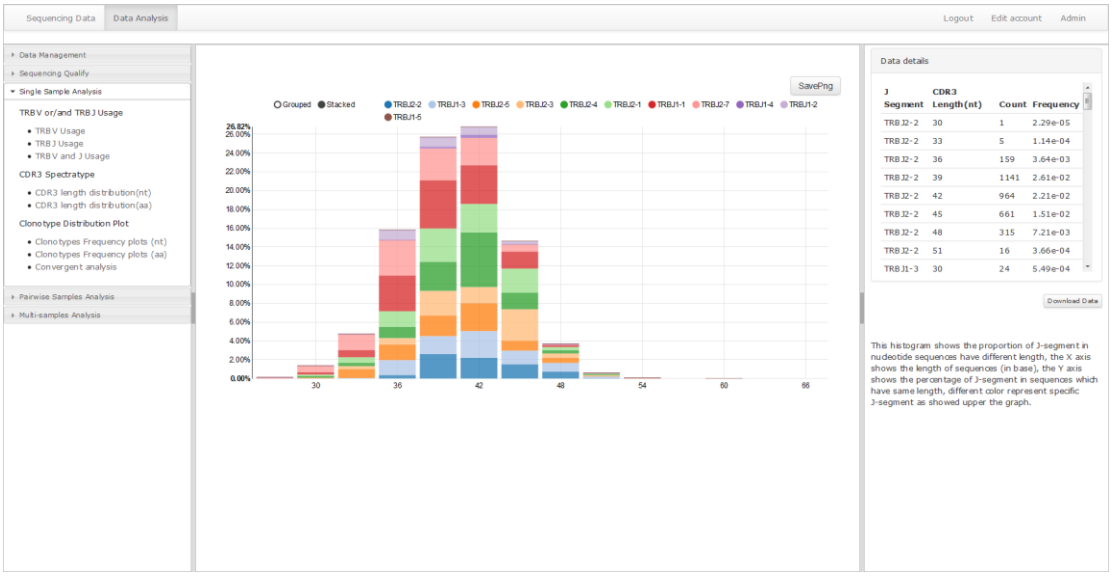


Figure 4.15 Screenshot showing the CDR3 length distribution by stacked bars

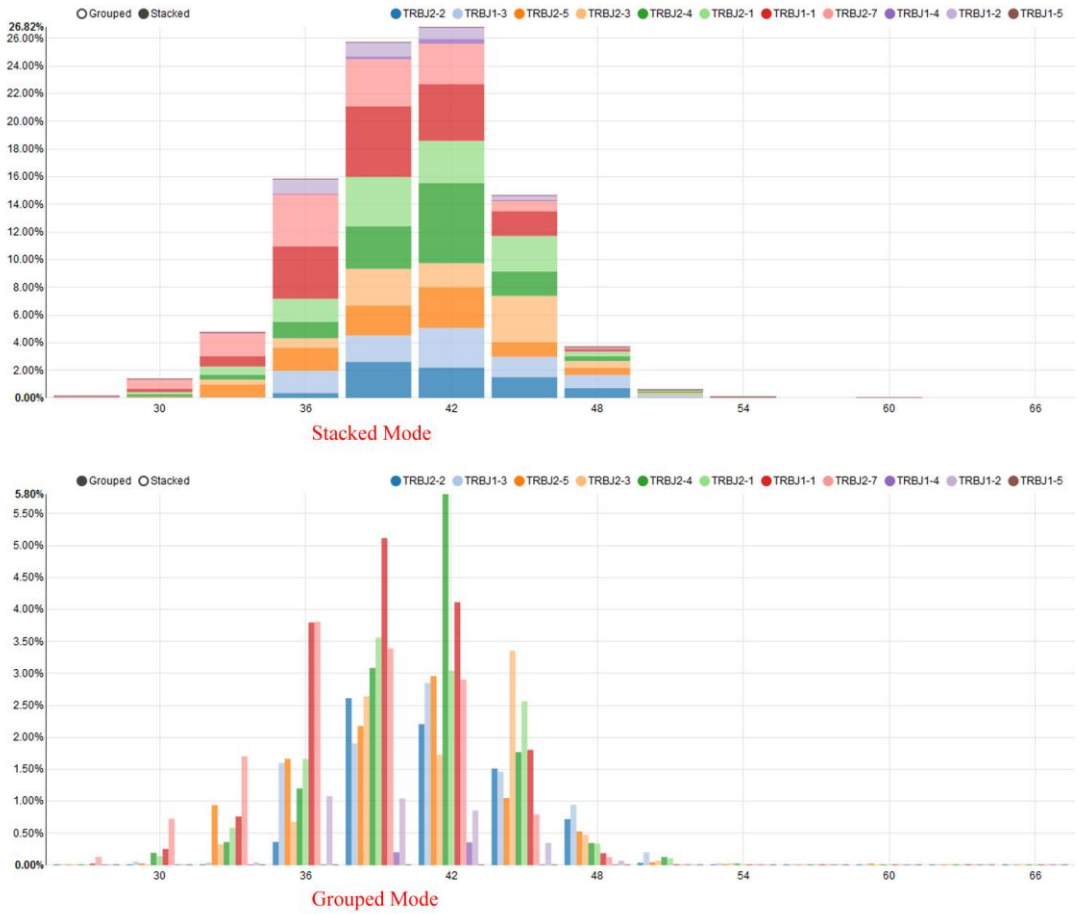


Figure 4.16 Stacked and grouped modes of the CDR3 spectratype distribution.

- **Clonotype Distribution Plot**

The Clonotype Distribution Plot shows the frequency of each specific nucleotide or amino acid sequence (Figure 4.17) in the TCR repertoire of the selected sample. The convergent analysis (Figure 4.18), shows a variety of dots, whose area represents the number of unique CDR3 nucleotide sequences that have been translated into the same CDR3 amino acid sequence. The degenerated nucleotide sequences are listed in the INFORMATION LIST, and can be downloaded (Table 2).

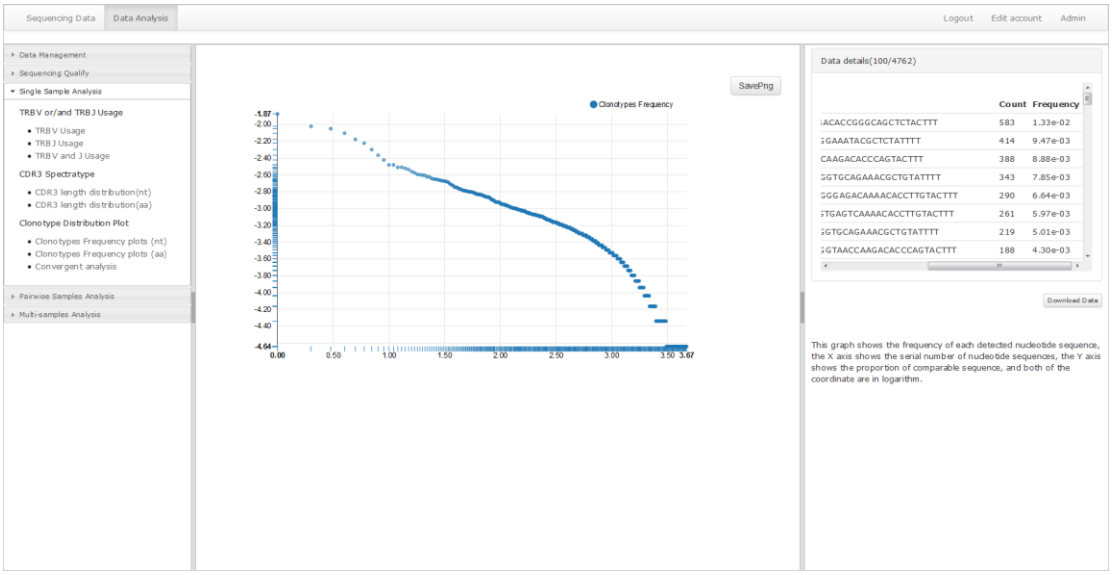


Figure 4.17 Screenshot showing the Clonotype Frequency.

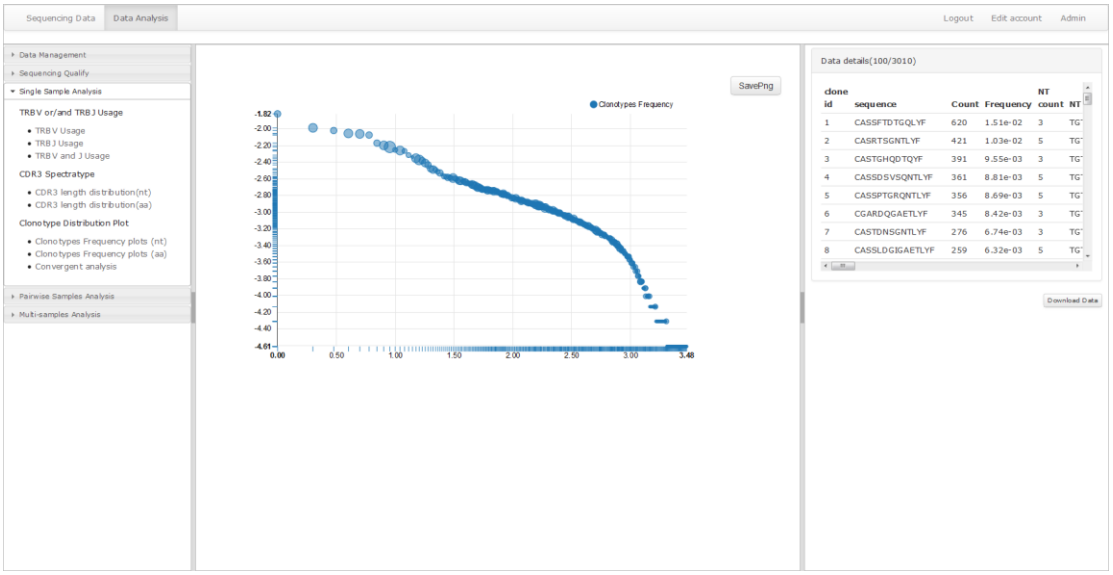


Figure 4.18 Screenshot showing the Convergent Clonotype Frequency Plot.

4.4 Pairwise Sample Analysis

In the Pairwise Samples Analysis section, the shared clonotypes between a selected pair of datasets are analyzed. To compare a pair of samples, they first need to be selected. To do this, click “Select samples” in the Pairwise Samples Analysis section (Figure 4.19). One data set for sample comparison can comprise a single sample or a combined group of samples. The results of this pairwise sample analysis include both un-overlapping and overlapping clonotype distribution and convergence analyses.

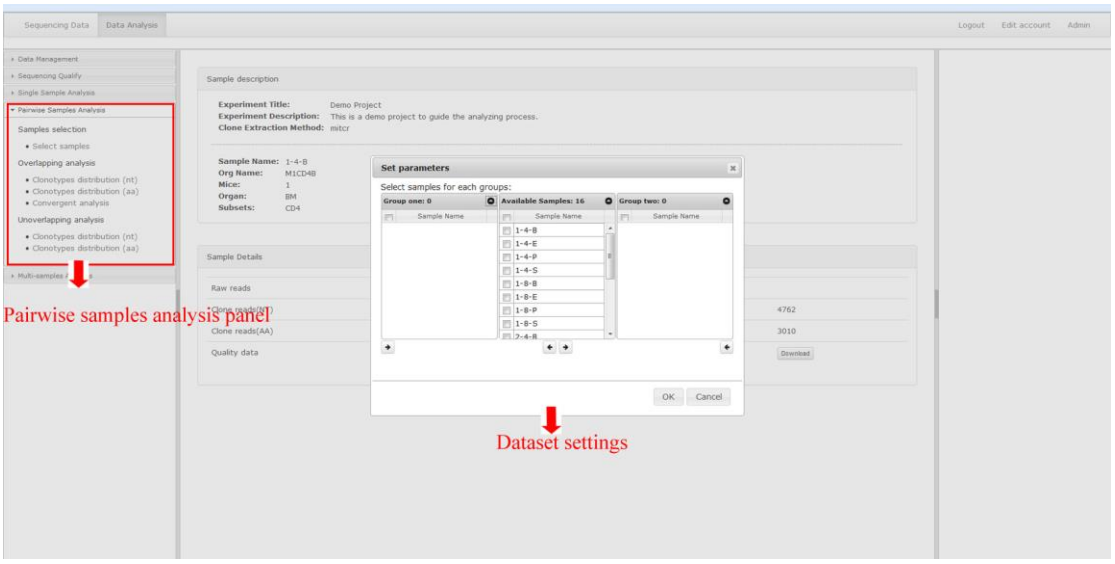


Figure 4.19 Screenshot of the Pairwise Samples Analysis panel and the dialogue box for selecting samples for comparison.

- **Overlapping clonotype distribution**

The overlapping analysis section calculates the frequency of nucleotide or amino acid (nt/aa) sequences that overlap in the two selected sets of sequencing data (Figure 4.20). The resultant plots show the frequencies that the clones overlap in the first and second groups, shown in the x and y axes, respectively.

Figure 4.20 Screenshot showing the overlapping clonotype distribution plot. The axes indicate the frequency of the shared clones in the paired data sets. The X axis indicates the frequency in Group 1 while the Y axis indicates the frequency in Group 2.

- **Convergent analysis**

The convergent analysis is used to evaluate the degeneracy of the shared T cell clonotypes (Figure 4.21). The area of the dots represents the number of unique CDR3 nucleotide sequences

that are translated into same CDR3 amino acid sequence.

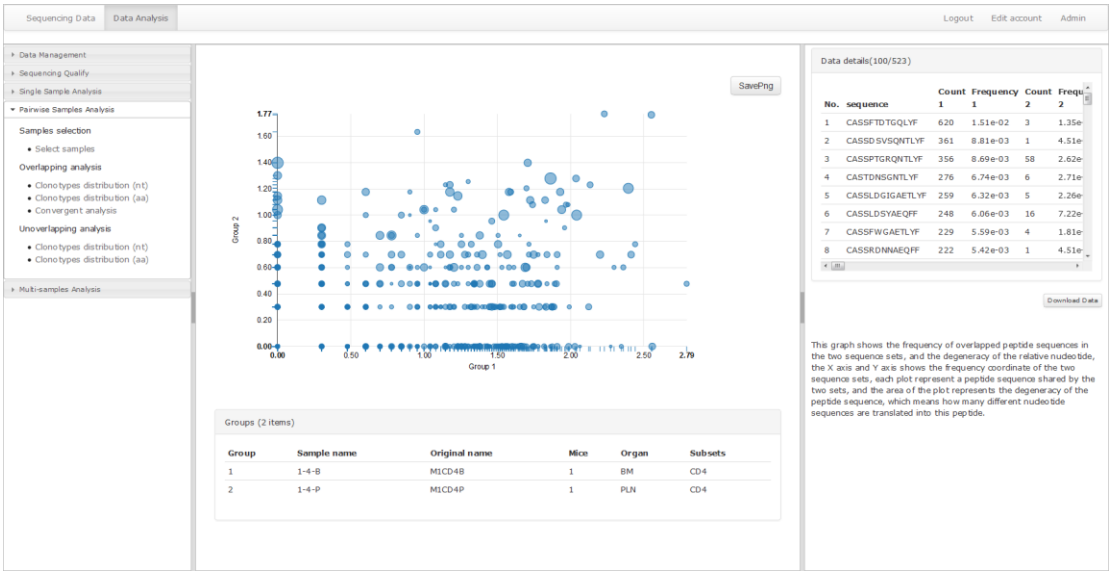


Figure 4.21 Screenshot showing the convergent analysis of two sample sets.

- **Un-Overlapping clonotype distribution**

The un-overlapping clonotype analysis displays the nt/aa distribution frequency of clonotypes that are not shared between the two data sets (Figure 4.22).

Figure 4.22 Screenshot showing the frequency of un-overlapping clonotypes between two selected samples.

4.5 Multi-sample Analysis

The Multi-sample Analysis tab uses the pre-defined experimental conditions, or the dependent variables and factors specified in the experiment design files to provide statistical analysis for multiple samples in the data set. The different statistical analyses that can be

compared are: descriptions of TCR clonotypes, similarity between grouped data sets, and analysis of the biodiversity of grouped data sets (Figure 4.23). The resultant charts and information are shown in the DISPLAY WINDOW, and specific details about the data are shown in the INFORMATION LIST.

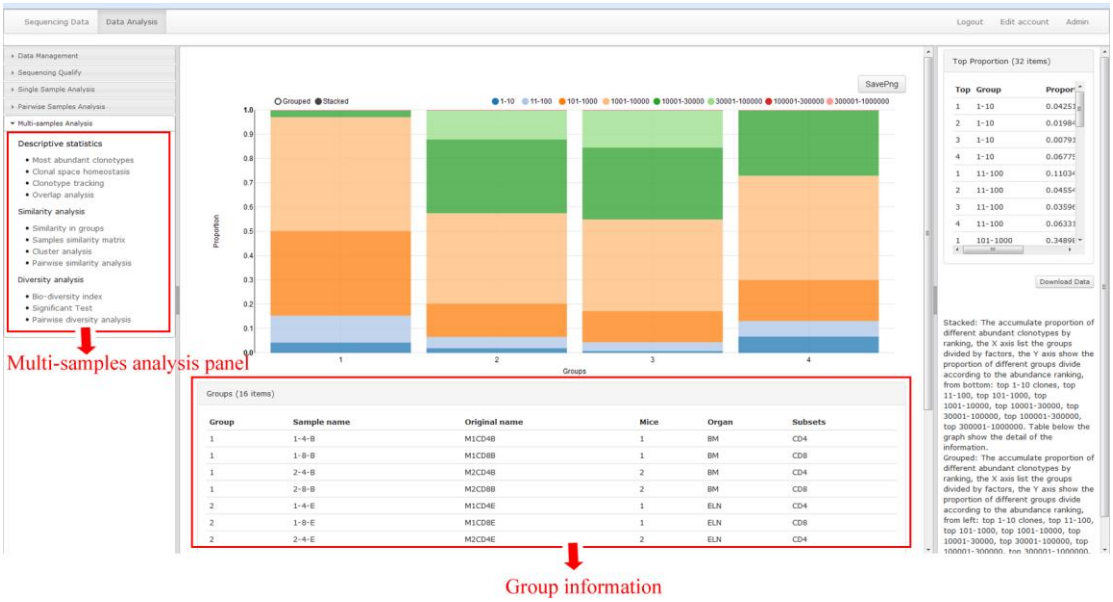


Figure 4.23 Screenshot showing the workspace in the Multi-sample Analysis tab.

4.5.1 Descriptive Statistics

The “Descriptive Statistics” tab allows users to look at the most abundant clonotypes, clonal space homeostasis, clonotype tracking, and overlapped analyses.

- **Most abundant clonotypes**

The “most abundant clonotype” option gives proportions of the most abundant clonotypes’ sum of reads to the overall number of reads in a combined samples’ TCR sequencing dataset. To open the dialogue box, click on the “most abundant clonotypes” button and group the samples according to the select data types and factors which have been pre-designated in the uploaded experiment design file (Figure 4.24). The resultant chart shows the most abundant clonotypes, and can be viewed in either stacked or grouped charts (Figure 4.25 and 4.26).

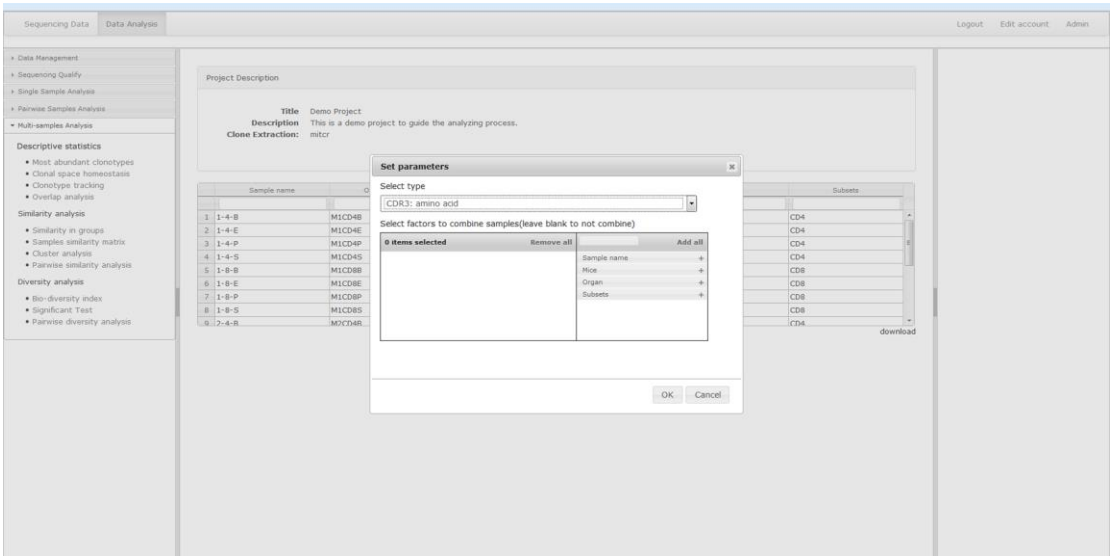


Figure 4.24 Screenshot of the dialogue box for setting sample parameters

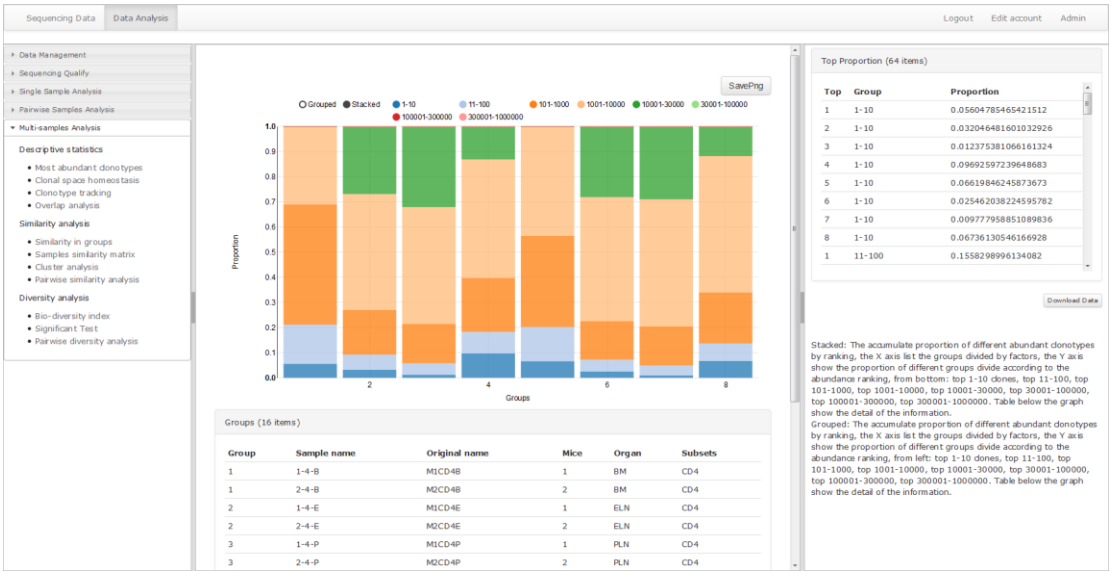


Figure 4.25 Screenshot showing the stacked chart of the most abundant clonotypes.

Figure 4.26 The most abundant clonotype analyses can be shown in either stacked (top), or grouped (bottom) charts.

- **Clonal space homeostasis**

The "Clonal space homeostasis" tab shows the proportional distribution of clones according to their sequence frequency in grouped samples. The resulting analysis shows whether the clonotypes have expanded or contracted, which is called space homeostasis. To open the dialogue box, click the "Clonal space homeostasis" tab and select the data type and factors used to group the samples, which have been pre-designated in the experiment design file. The clonal space homeostasis chart can be viewed in either a stacked or grouped chart (Figure 4.27).

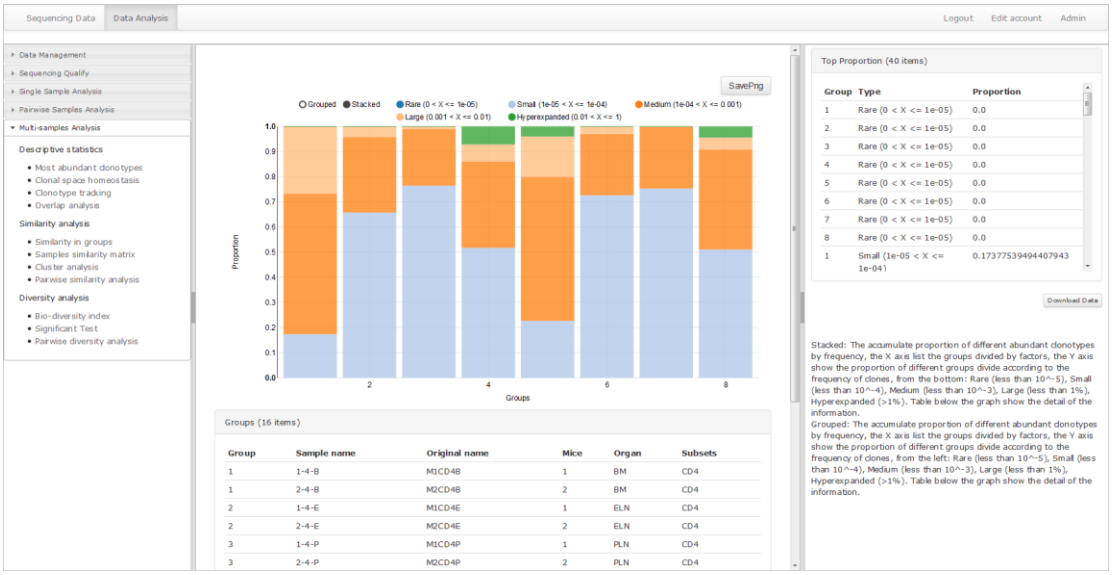


Figure 4.27 Screenshot showing the stacked distribution of expanded or contracted clones- the clonal space homeostasis.

- **Clonotype tracking**

The "Clonotype tracking" analysis shows the frequencies that specific clonotypes arise in different datasets. To open the dialogue box, click the "Clonotype tracking" button and select the desired factors to group the samples, which are pre-designated in the uploaded experiment design file. The clonotype tracking analysis will generate both a line and curved chart. In the line chart, each line represents a shared clonotype, while each point on the line represents the count of that shared clonotype in each group (Figure 4.28). The curved chart shows the number of clonotypes found in each sharing category, with the Y-axis showing the counts of the shared clonotypes, and the X-axis showing how many groups share that clonotype (Figure 4.29).

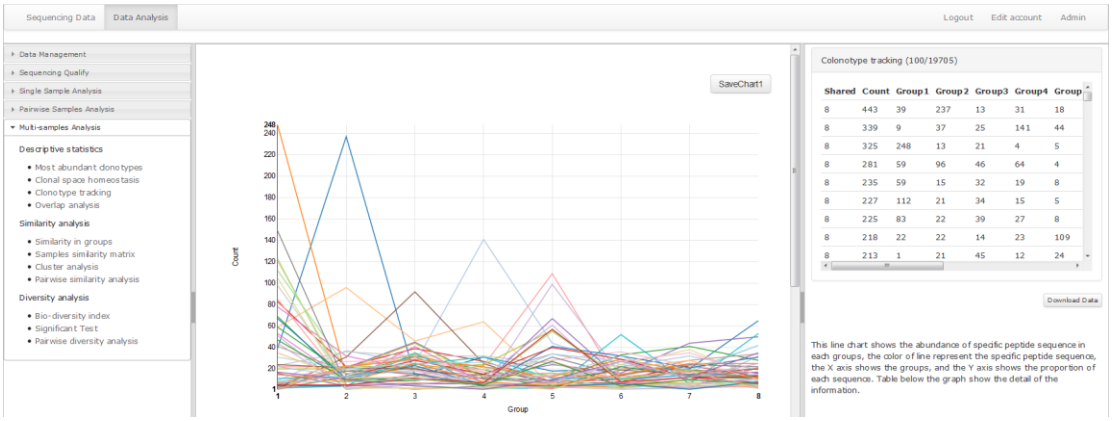


Figure 4.28 Screenshot showing the line chart for clonotype tracking.

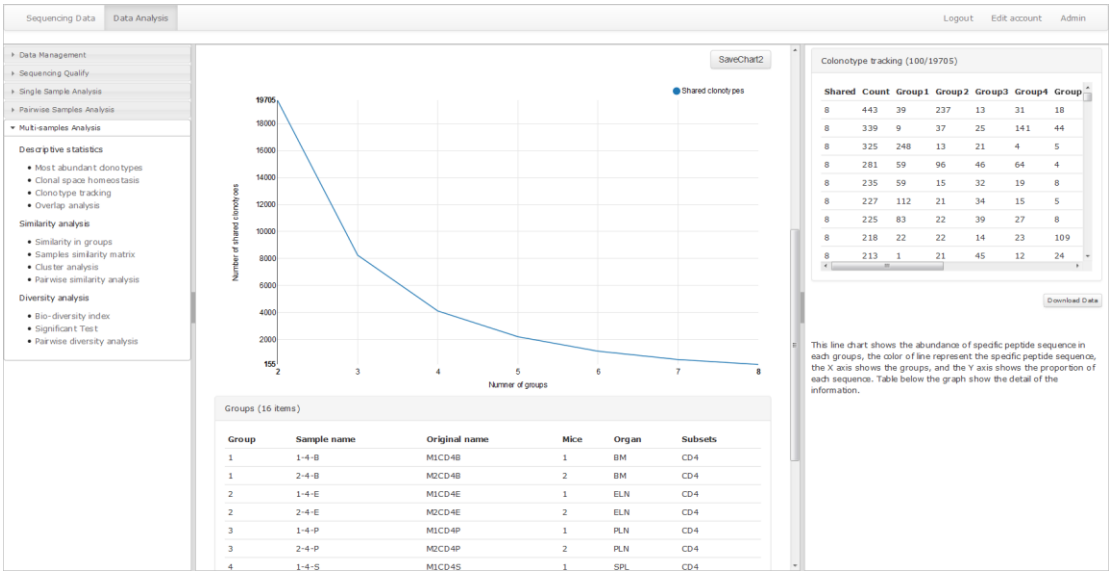


Figure 4.29 Screenshot showing the number of clonotypes found in each sharing category

- **Overlap analysis**

The overlap analysis provides a heat-map matrix of the shared number of clonotypes between the grouped data sets (Figure 4.30). To open the dialogue box, click the “Overlap analysis” button to group the samples and to select individual factors and methods for displaying the data. Users can select individual cells in the heat map to look at more specific details of the data in that corresponding row and column. The color and size of the heat map can be adjusted by changing the scales at the top of the graph.

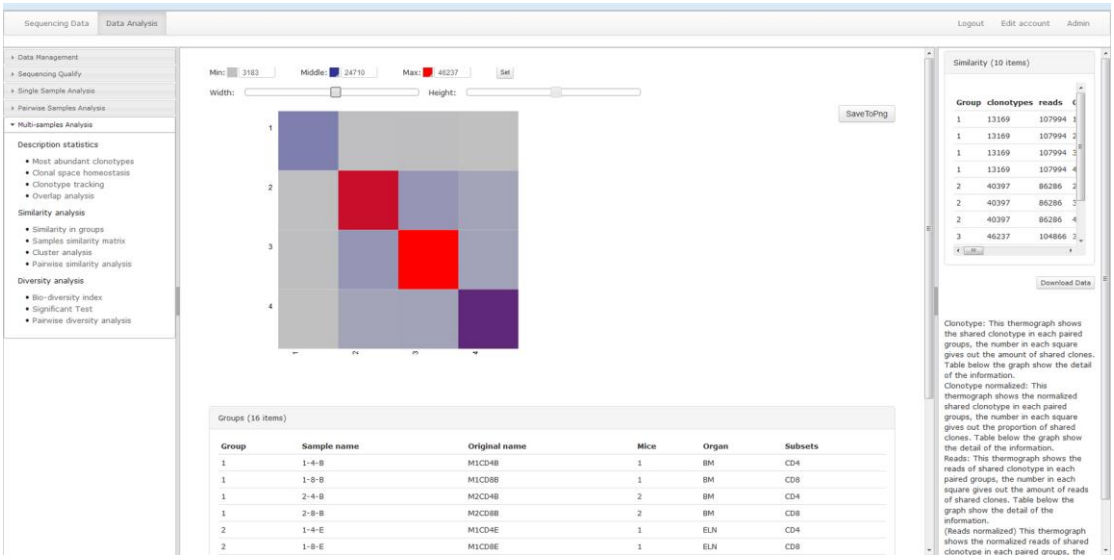


Figure 4.30 Screenshot showing the heat map of overlapping clonotypes.

4.5.2 Similarity analysis

- **Similarity in groups**

The “Similarity in groups” tab shows the average value of similarity between pairwise

samples in one data set. To open the dialogue box, click the “Similarity in groups” button and select the type of data, method for analyzing similarity, and factors for grouping (Figure 4.31). The resultant data will be displayed in a histogram form (Figure 4.32). For more information and details about the methods for determining similarity, refer to Section 5.1.

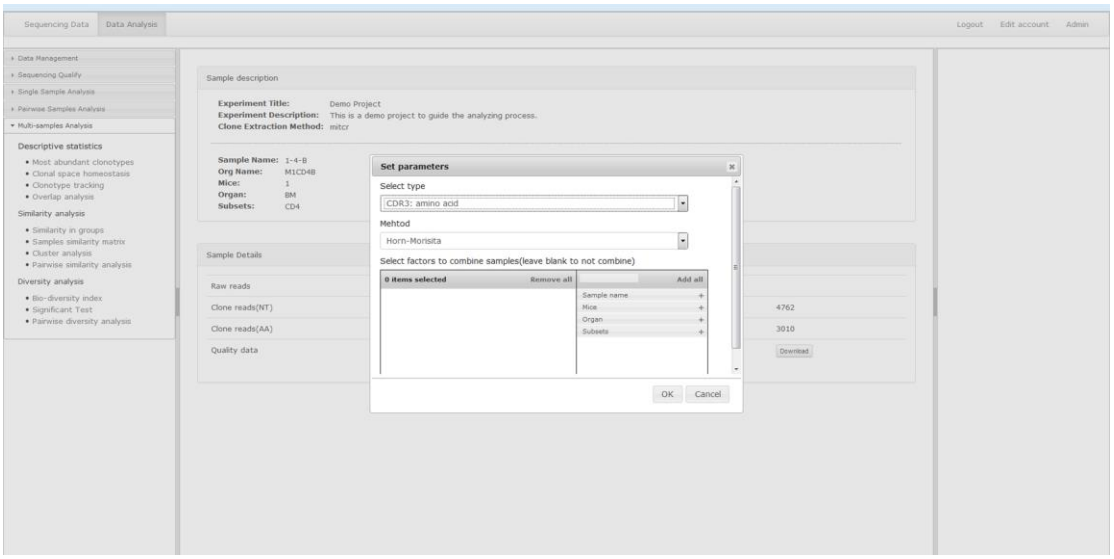


Figure 4.31 Screenshot of the dialogue box comparing the similarity of groups.

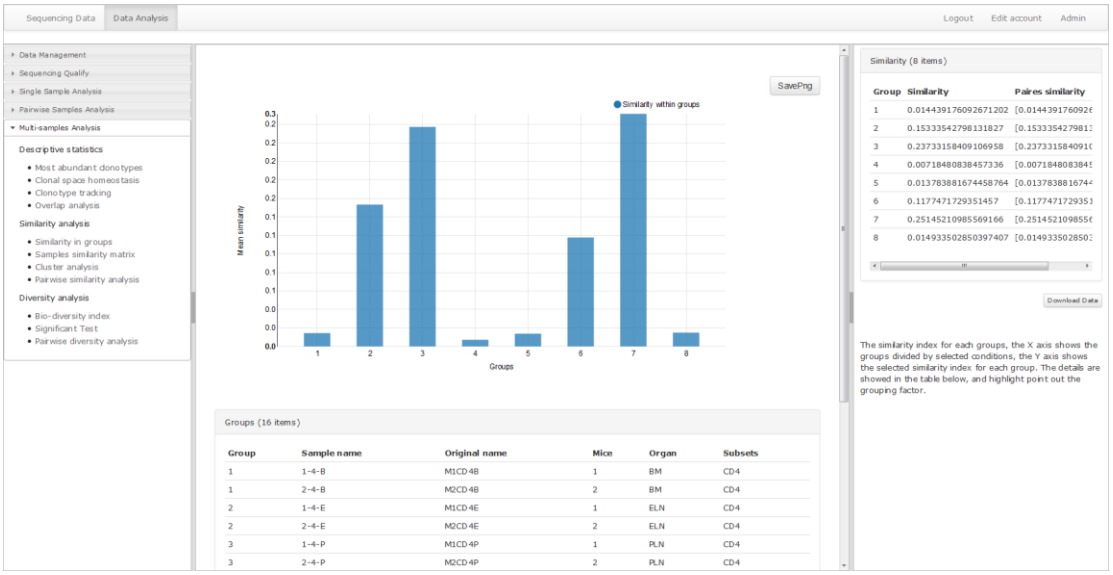


Figure 4.32 Screenshot showing the histogram representing the average similarity of all samples in a selected data set.

- **Sample similarity matrix**

The “Sample similarity matrix” shows similarities within each group, defined by the average similarity between all paired samples in two related data sets. To open the dialogue box, click the “Similarity in groups” button, and select the data type, method of analysis, and factors for grouping. The resultant data will be shown in the form of a heat map (Figure 4.33). Details of the corresponding rows and columns will be shown once individual cells are selected in the heat map.

The colors and size of the heat map can be adjusted by using the bars on top of the graph.

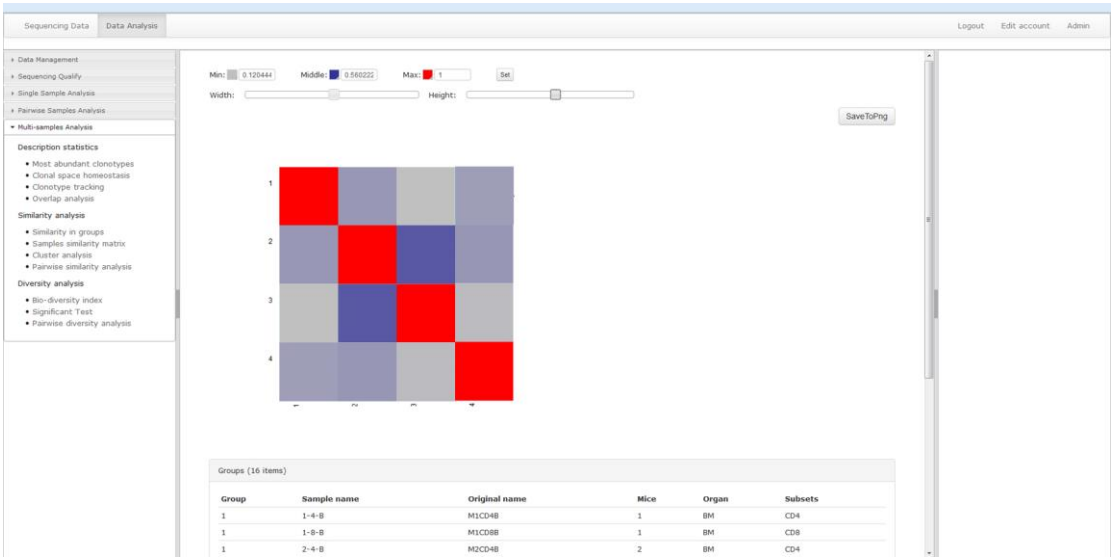


Figure 4.33 Screenshot showing the resultant heat map in the ‘samples similarity matrix’ tab.

- **Cluster analysis**

The “Cluster analysis” function shows the similarity matrix of all grouped data sets together, using several available methods for cluster analysis. The cluster analysis is performed with the similarity profiles between all grouped TCR sequencing datasets. The similarity profiles of one grouped TCR sequencing dataset is a vector that represented the similarity between the grouped TCR sequencing dataset to all of the grouped TCR sequencing datasets. To open the dialogue box, click the “Cluster analysis” button and select the data type, methods of analysis, and factors for grouping (Figure 4.34). The resultant data will be displayed in a heat map (Figure 4.35).

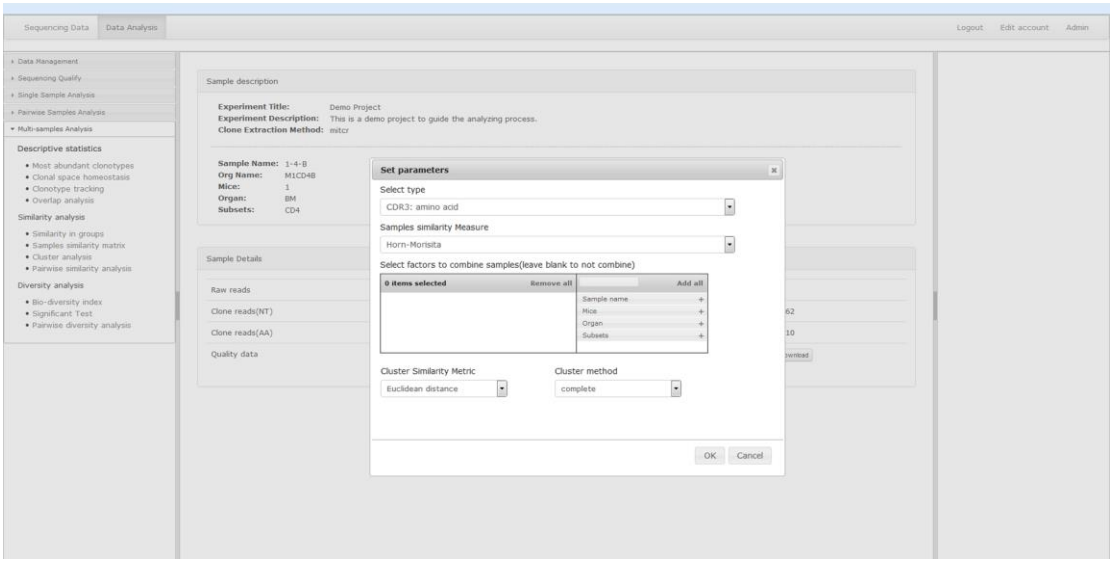


Figure 4.34 Screenshot showing the dialogue box in the “cluster analysis” section.

Figure 4.35 Screenshot showing cluster analysis of the similarity profiles between all grouped TCR sequencing datasets. The similarity profiles of one grouped TCR sequencing dataset is a vector that represented the similarity between the grouped TCR sequencing dataset to all of the grouped TCR sequencing datasets

- **Pairwise similarity analysis**

The “Pairwise similarity analysis” tab categorizes the similarity indices between selected sample groups, and performs statistical testing on these grouped samples. To open the dialogue box to set parameters, click on the “Pairwise similarity analysis” button, and select the data type, and methods for analysis (Figure 4.36). If the experimental samples are paired, a paired t-test can be performed if the user selects the paired option. In the dialogue box, users can designate whether their samples are single or grouped, and separate them in Group 1 or Group2. From this, the similarity between the two data sets will be calculated and statistical analyses will be performed. Please note that at least three samples need to be designated per group for accurate statistical analyses to be performed (Figure 4.37). The data will be presented in a box-and-whisker plot, and the standard error and significance report from the statistical testing will be provided. The individual groups will be shown on the X axis, and the values of the similarity index between the selected samples will be represented on the Y axis.

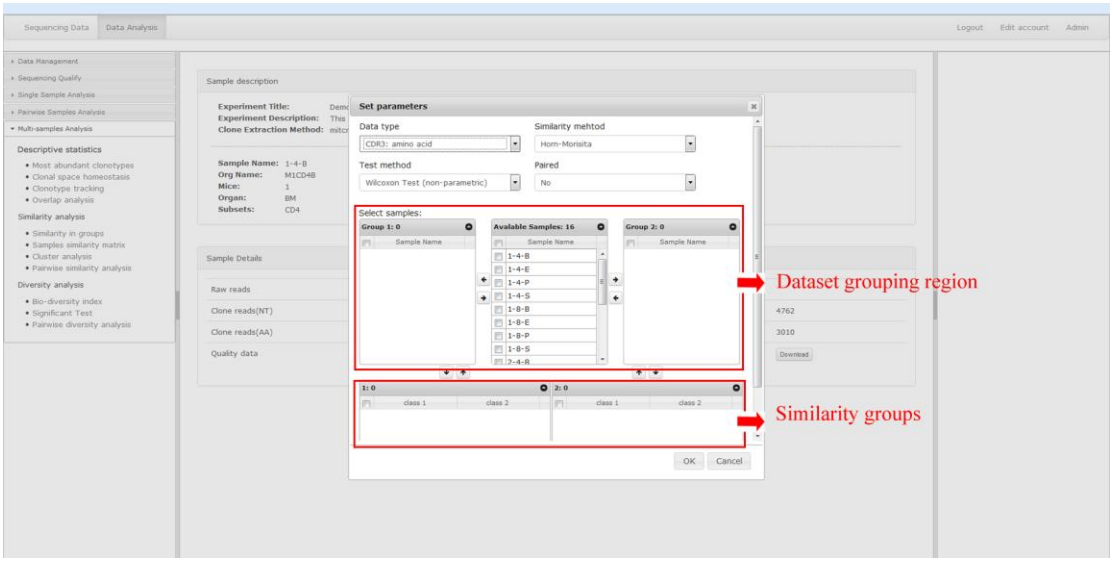


Figure 4.36 Screenshot of the dialog box in the 'pairwise similarity analysis' tab for parameter selection.

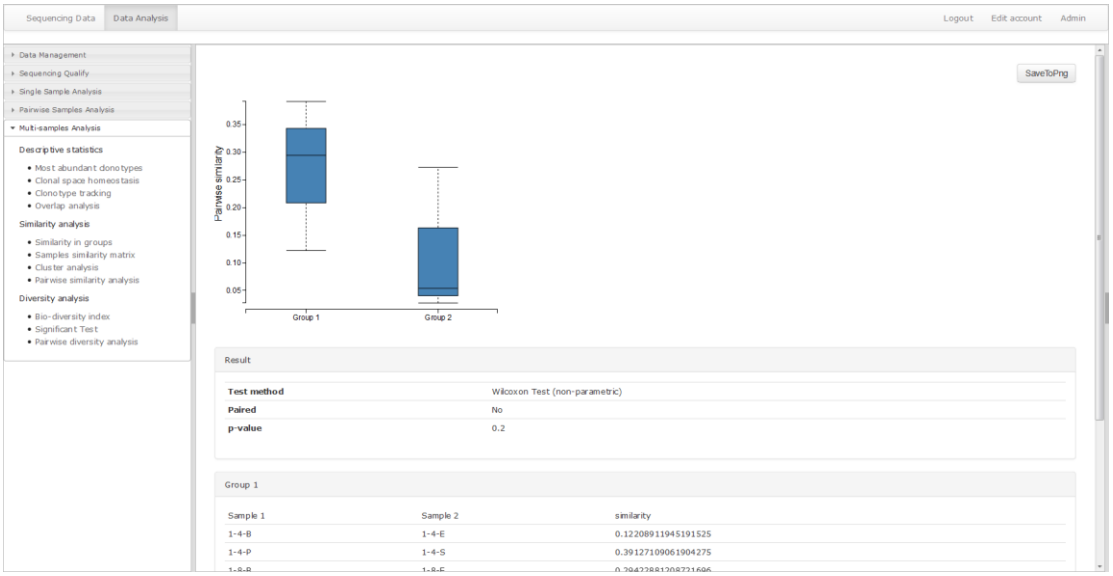


Figure 4.37 Screenshot showing the box and whisker plot and the accompanied standard error and significance report in the pairwise similarity analysis option.

4.5.3 Diversity analysis

- **Biodiversity index**

Samples are combined according to the factors listed in the experiment design file, and used to calculate the biodiversity index. Click the "Biodiversity index" button to open the dialogue box and select the data type, method for biodiversity analysis, and factor grouping (Figure 4.38). The biodiversity indices are displayed in the form of a histogram (Figure 4.39). For further information on diversity calculations, please refer to Section 5.2.

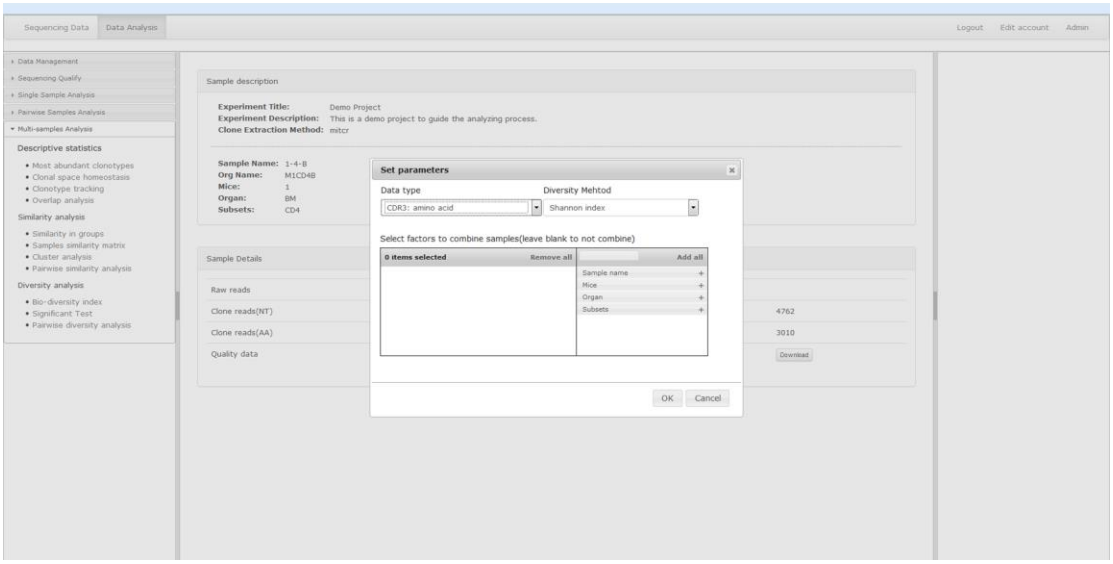


Figure 4.38 Screenshot of dialog box for setting biodiversity parameters and methods for analyzing biodiversity and grouping factors.

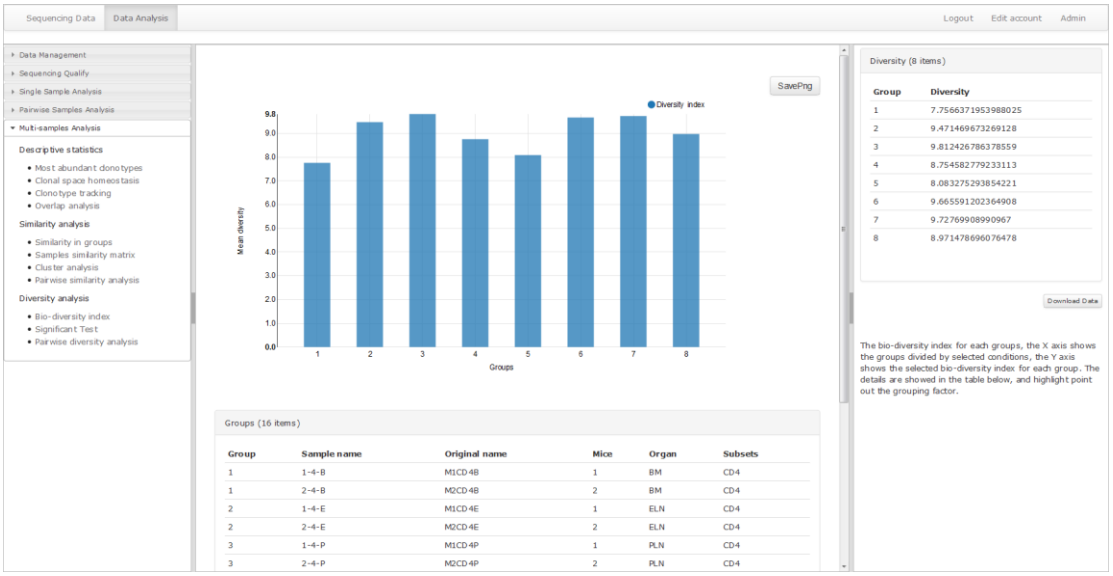


Figure 4.39 Screenshot showing the biodiversity index histogram.

- **Significance test**

The Significance Test option performs a one-way ANOVA of the biodiversity between different data sets. To open the dialogue box, click the “Significance test” button (Figure 4.40), and then select the diversity index and data grouping factors. The data output will be generated as shown in Figure 4.41.

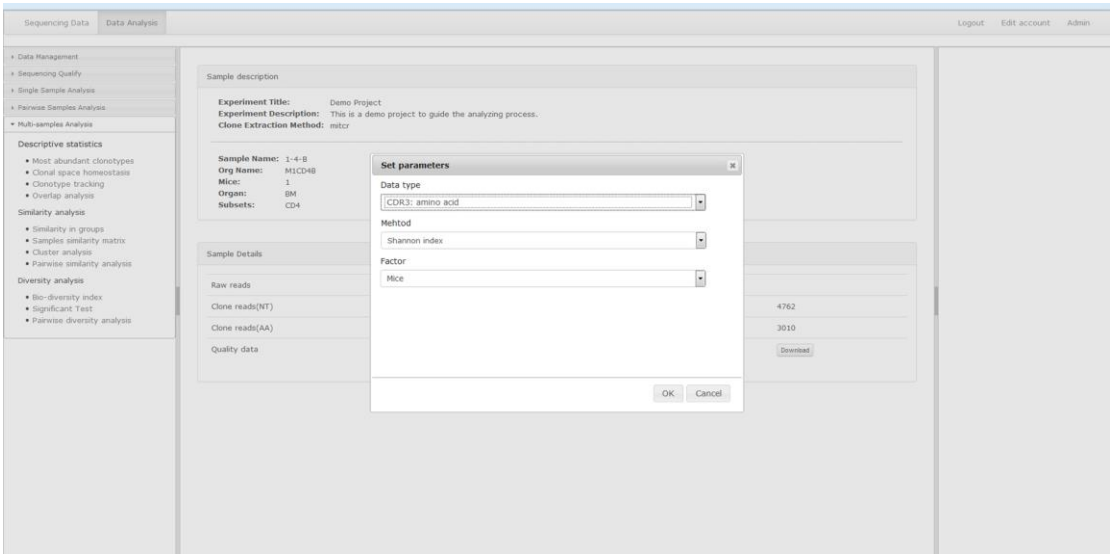


Figure 4.40 Screenshot of the dialog box to select parameters for diversity analysis and ways to group the data.

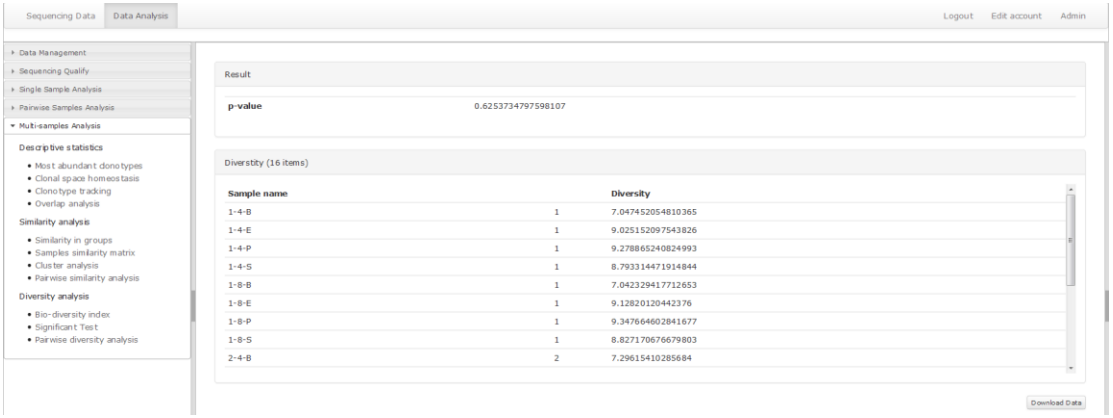


Figure 4.41 Screenshot showing the results of the one-way ANOVA biodiversity significance test on the selected data sets.

- **Pairwise diversity analysis**

To analyze pairwise diversity, two selected groups of samples are categorized together, and statistical hypothesis testing is performed. To open the dialogue box and select parameters, click the “Pairwise diversity analysis” button and select the data type, methods for analysis, and methods to test similarity (Figure 4.42). If the experimental samples are paired, select the ‘pair’ option, and a paired t-test will be performed. Individual or grouped samples can be designated into Group 1 and Group 2, but a minimum of three samples are required for statistically significant similarity index testing (Figure 4.43).

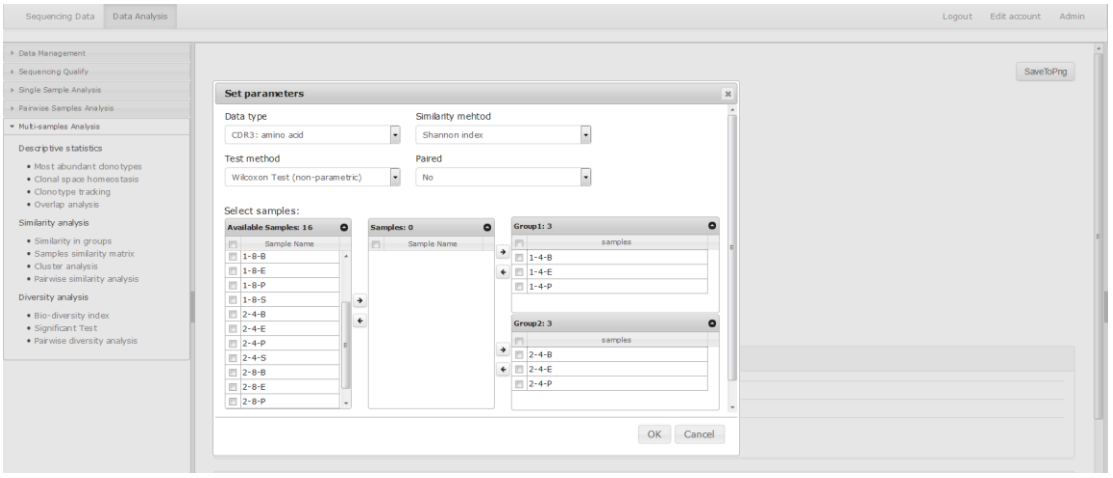


Figure 4.42 Screenshot showing the dialogue box to select groups for pairwise diversity analysis.

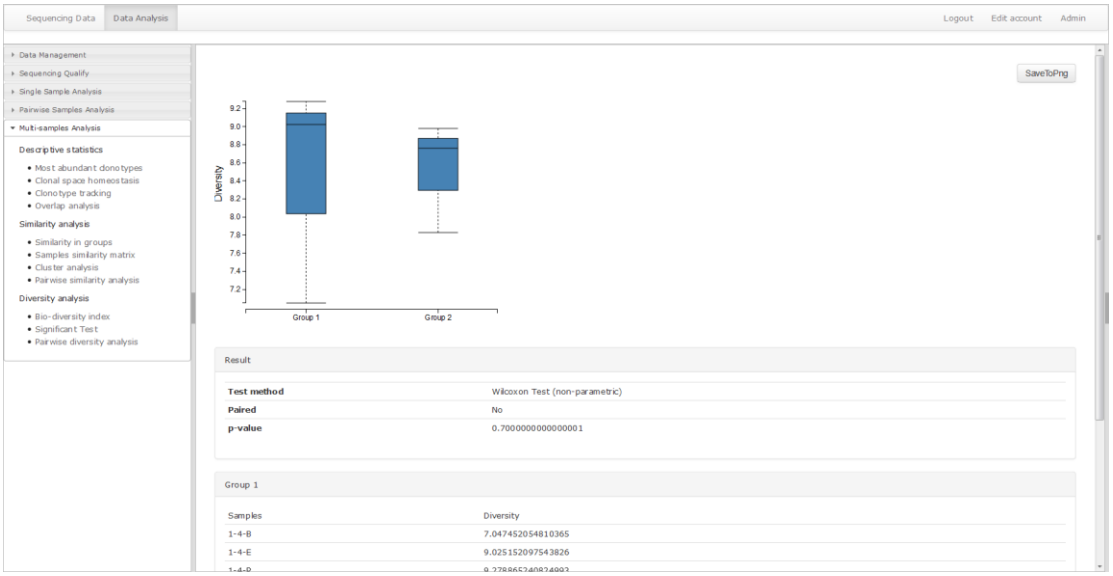


Figure 4.43 Screenshot showing the box-and-whisker plot output, including the standard error and significance report of pairwise diversity analysis.

5. Appendices

5.1 Analysis methods for similarity

The following similarity indexes are used to estimate the similarity between samples. For the following descriptions, let N be the total number of clonotypes present in sample x and y. x_i and y_i represent the total clonotype read number in sample x and y respectively:

Morisita Index[6]: This similarity index measures the dispersion of clonotypes in a sample. The formula is:

$$\mathrm{S(x,y)} = \frac{2\sum_{i=1}^{N} x_i y_i}{(\lambda_x + \lambda_y)\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}$$

where,

$$\lambda = \frac{\sum x_i(x_i - 1)}{\sum x_i \sum(x_i - 1)}$$

Horn-Morisita Index[6]: This similarity index is an improvement from the Morisita Index. It uses a different λ and is more sensitive to the clone sizes of the dominant clonotypes, where:

$$\lambda = \frac{\sum x_i^2}{(\sum x_i)^2}$$

Jaccard Index[7]: This similarity index is defined as the size of the intersection divided by the size of the union of the clonotypes in two samples. The formula is:

$$\mathrm{J(x,y)} = \frac{\sum_{i=1}^{N} \min(x_i, y_i)}{\sum_{i=1}^{N} \max(x_i, y_i)}$$

Bray-Curtis Index[8]: This index is used to quantify the compositional dissimilarity between clonotypes based on counts of each clonotype. The formula is:

$$\mathrm{BC(x,y)} = 1 - \frac{\sum_{i=1}^{N} |x_i - y_i|}{\sum_{i=1}^{N} x_i + y_i}$$

Kulczynski Index[9]: This index represents the average of the proportion of the common clones in two repertoires. The formula is:

$$\mathrm{K(x,y)} = \frac{1}{2}\left(\frac{\sum_{i=1}^{N} min(x_i, y_i)}{\sum_{i=1}^{N} x_i} + \frac{\sum_{i=1}^{N} min(x_i, y_i)}{\sum_{i=1}^{N} y_i}\right)$$

Binomial Index[10]: This function is derived from binomial deviance under the null hypothesis that the two compared repertoires are equal. It can handle variable sample sizes. The formula is:

$$\mathrm{B(x,y)} = 1 - \sum_{i=1}^{N} \frac{1}{s_i}(x_i(\ln x_i - \ln s_i) + y_i(\ln y_i - \ln s_i) + \ln 2)$$

where

$$s_i = x_i + y_i$$

Cao Index[11]: This index is a function suggested as a minimally biased index for high beta diversity and variable sampling intensity. The formula is:

$$\mathrm{C(x,y)} = 1 - \frac{1}{N}\sum_{i=1}^{N}(\ln\frac{s_i}{2} - \frac{1}{s_i}(x_i ln y_i + y_i ln x_i))$$

where

$$s_i = x_i + y_i$$

5.2 Diversity analysis methods

Shannon index[12]: This index quantifies the uncertainty (entropy) in TCR repertoires; the core idea of this index is that many different clonotypes have an equal proportional abundance, and thus a smaller repertoire. The formula is:

$$D = -\sum_{i=1}^{N} p_i \, ln \, p_i$$

where

$$p_i = x_i / \sum_{j=1}^{N} x_j$$

Simpson's index & Inverse Simpson's index[13]: This index represents the probability that two clones taken at random from the repertoire of interest represent the same clonotype. The formula is:

$$D = \sum_{i=1}^{N} p_i^2$$

The Inverse Simpson's index: This index represents the inverse of Simpson's index

$$D = \frac{1}{\sum_{i=1}^{N} p_i^2}$$

Gini–Simpson index[13]: This index is similar to the Simpson's index, but represents the probability that the two clones represent different types. The formula is:

$$D = 1 - \sum_{i=1}^{N} p_i^2$$

Berger-Parker index[14]: This index shows the maximum p_i value in the dataset. It is used to assess the dominance of a specific TCR clonotype in an entire repertoire.

$$\mathrm{D} = \max(p_i)$$

Renyl entropy[15]: This is a generalization of the Shannon entropy to other values of q than unity. The formula is:

$$\mathrm{D} = \frac{1}{1-q}\ln(\sum_{i=1}^{N} p_i^q)$$

6. Copyright

<VisTCR, Software for T Cell Repertoire High-throughput Sequencing Data Analysis>
Copyright (C) <2016><QS Ni, JY Zhang, Y Wan, TMMU China>

This program is free software and can be redistributed and/or modified under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the license, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but without any real or implied warranty of merchantability or fitness for a particular purpose. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program; if not, write to the Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301 USA. Include information on how to contact you by electronic and paper mail.

7. References

1. Bostock, M., Ogievetsky, V. & Heer, J. D(3): Data-Driven Documents. *IEEE Trans Vis Comput Graph* **17**, 2301-2309 (2011).
2. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175-185 (1998).
3. Bolotin, D.A. et al. MiTCR: software for T-cell receptor sequencing data analysis. *Nat Methods* **10**, 813-814 (2013).
4. Bolotin, D.A. et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* **12**, 380-381 (2015).
5. Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J. & Chain, B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* **29**, 542-550 (2013).
6. Anderson, M.J. & Millar, R.B. Spatial variation and effects of habitat on temperate reef fish assemblages in northeastern New Zealand. *J Exp Mar Biol Ecol* **305**, 191-221 (2004).
7. Anderson, M.J. & Thompson, A.A. Multivariate control charts for ecological and environmental monitoring. *Ecol Appl* **14**, 1921-1935 (2004).

8. Anderson, M.J., Ellingsen, K.E. & McArdle, B.H. Multivariate dispersion as a measure of beta diversity. *Ecol Lett* **9**, 683-693 (2006).
9. Chase, J.M., Kraft, N.J.B., Smith, K.G., Vellend, M. & Inouye, B.D. Using null models to disentangle variation in community dissimilarity from variation in alpha-diversity. *Ecosphere* **2** (2011).
10. Veech, J.A. A probabilistic model for analysing species co-occurrence. *Global Ecol Biogeogr* **22**, 252-260 (2013).
11. Chao, A., Chazdon, R.L., Colwell, R.K. & Shen, T.J. A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol Lett* **8**, 148-159 (2005).
12. Shannon A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379 (1948).
13. Simpson Mesurement of diversity. *Nature* **163** (1949).
14. Hill Diversity and evenness: a unifying natation and its consequences. *Ecology 54*, 427-432 (1973).
15. Jost, L. Entropy and diversity. *Oikos* **113**, 363-375 (2006).