iRepertoire

We Find Health in Your Diversity.

# iRweb: Data Analysis Guide

Immunorepertoire Amplification and Next-Gen Sequencing

For Research Use Only, Not to be Used for Clinical Diagnostics.

v092013

iRepertoire

We Find Health in Your Diversity.

iRepertoire® is a registered trademark of iRepertoire, Inc. The iR logo is a trademark of iRepertoire, Inc. Illumina®, HiSeq®,and MiSeq®, are registered trademarks of Illumina, Inc. HiSeq2000™ and GAIIx™ are trademarks of Illumina, Inc. 454®, 454 Sequencing®, GS FLX Titanium®, and GS Junior® are registered trademarks of Roche Diagnostics GmbH. Ion Torrent® is a registered trademark of Life Technologies Corporation, Inc.

iRepertoire, Inc. does not assume any liability, whether direct or indirect, arising out of the application or use of any products, component parts, or software described herein or from any information contained in this guide. Furthermore, sale of iRepertoire, Inc. products does not constitute a license to any patent, trademark, copyright, or common-law rights of iRepertoire or the similar rights of others. iRepertoire, Inc. reserves the right to make any changes to any processes, products, or parts thereof, described herein without notice. While every effort has been made to make this manual as complete and accurate as possible as of the publication date, iRepertoire assumes no responsibility that the goods described herein will be fit for any particular purpose for which you may be buying these goods.

# Table of Contents

# Introduction

High throughput sequencing produces a massive amount of detailed TCR or BCR sequence information for each library sequenced, which must be processed in order to extract meaningful information. To facilitate data analysis, we have implemented an automated software pipeline. This pipeline applies stringent filters to remove errors that may have occurred during the amplification and sequencing process. Once the data is filtered, several types of analyses are performed.

### Recommended Browser

For best viewing results, please use the Mozilla Firefox or Google Chrome web browsers.

### Logging In

Please log in to your account first by going to https://irweb.irepertoire.com/nir/ . If you want to preview the software, you can log in to the demo account with username: demo and password: 12345. Once you log in, you will see a screen similar to that demonstrated in Figure 1.

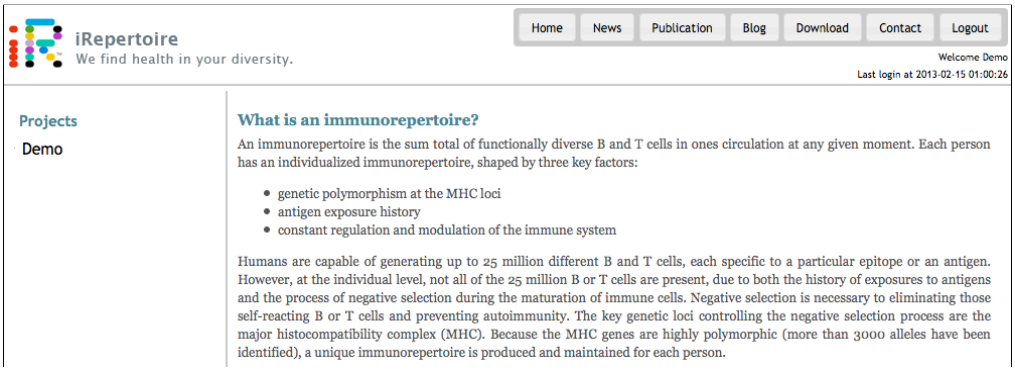


**Figure 1**: Initial page after logging into the demo account.

To access your data, select demo or your sample name from the left panel. On the demo account, there are three demo panels that appear: Demo 1, Demo 2, and Demo 3. Click on one of these demo modules. For instance if Demo 1 is selected, an option to preview IGH (Immunoglobulin Heavy Chain), IGK (Immunoglobulin Kappa chain), IGL (Immunoglobulin Lambda Chain), TRA (TCR-alpha), TRB (TCR-beta), TRD (TCR-delta), and TRG-(TCR gamma) will be accessible. If you select, for instance, IGH, a new page will be launched, and several analyses will be available on the left panel including “Show 2D Map,” “Show 3D Map,” “List CDR3,” “CDR3

algebra," "Compute D50," and "Tree Map." Also available are several distribution analyses including "V usage," "J usage," "V trimming," "J trimming," "CDR3 length," and "N-addition." The normalized distributions of these options are also available. Above theses analyses buttons, there is also a summary of the statistics for that particular data set.

# Analysis: Show 2D Map

## Show 2D Map: Heat Map

Figure 2 is an example of a 2 dimensional heat map from the T helper population of both a colon cancer patient and a normal control. The relative frequency of a germline V-gene allele (as per alignment with the IMGT database) is plotted relative to the germline J-gene allele. Therefore, it is immediately evident which V-J combination is used either frequently or infrequently by the color of the map. The map is interactive. Once a specific box is clicked, the sequence alignment for representative sequences in the library containing that specific V-J combination appears, as demonstrated in Figure 3. Many sequences may appear in this output list because it contains all sequences in the library with a particular V-J combination. The list provides an abundance of detailed sequence information including the translated protein sequence, the DNA sequence of the read, its alignment with the IMGT database, any differences with the germline allele sequence, and identification of CDR1, CDR2, and CDR3. In addition, the CDR3 sequence will also be listed in the fasta-like header for the sequence.

The identification of CDR1 through CDR3 will depend on the sequencing method utilized. There are three types of sequencing available, Illumina HiSeq 100 and 150-paired-end reads (PER), Illumina MiSeq (100, 150, or 250-PER), and Roche 454. The Illumina HiSeq platform and MiSeq (100-PER or 150-PER) provides about 150 base pairs of sequence data around the CDR3. The MiSeq 250-PER and Roche 454 platforms allow for the sequencing of around 450 bp around the CDR3 and are better suited for BCR sequencing because they can provide sufficient information about the CDR1 to CDR3 region with hypermutation patterns. As demonstrated in Figure 3, information pertaining to CDR1, CDR2, and CDR3 is displayed for the MiSeq 250-PER. All sequencing platforms allow for the identification of unique CDR3s.

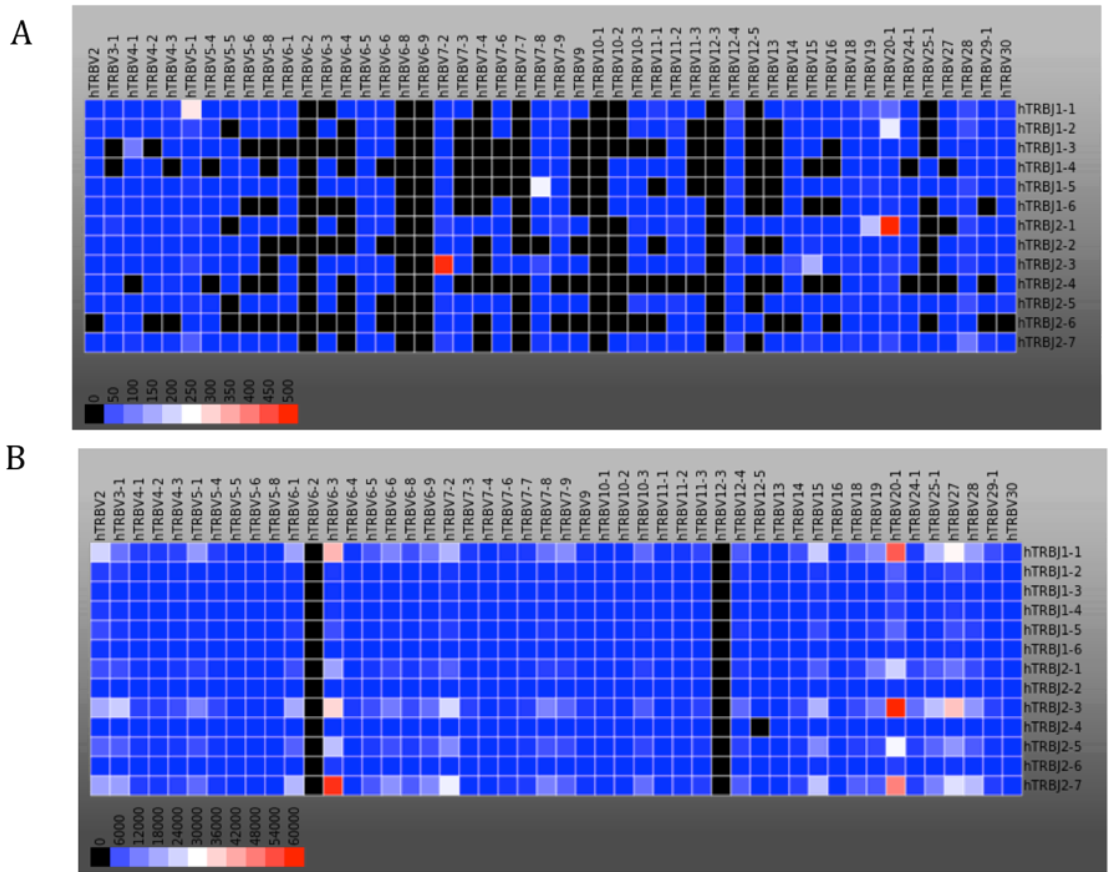


**Figure 2:** Heat map of the T-helper population of a colon cancer patient (A) and normal patient (B).

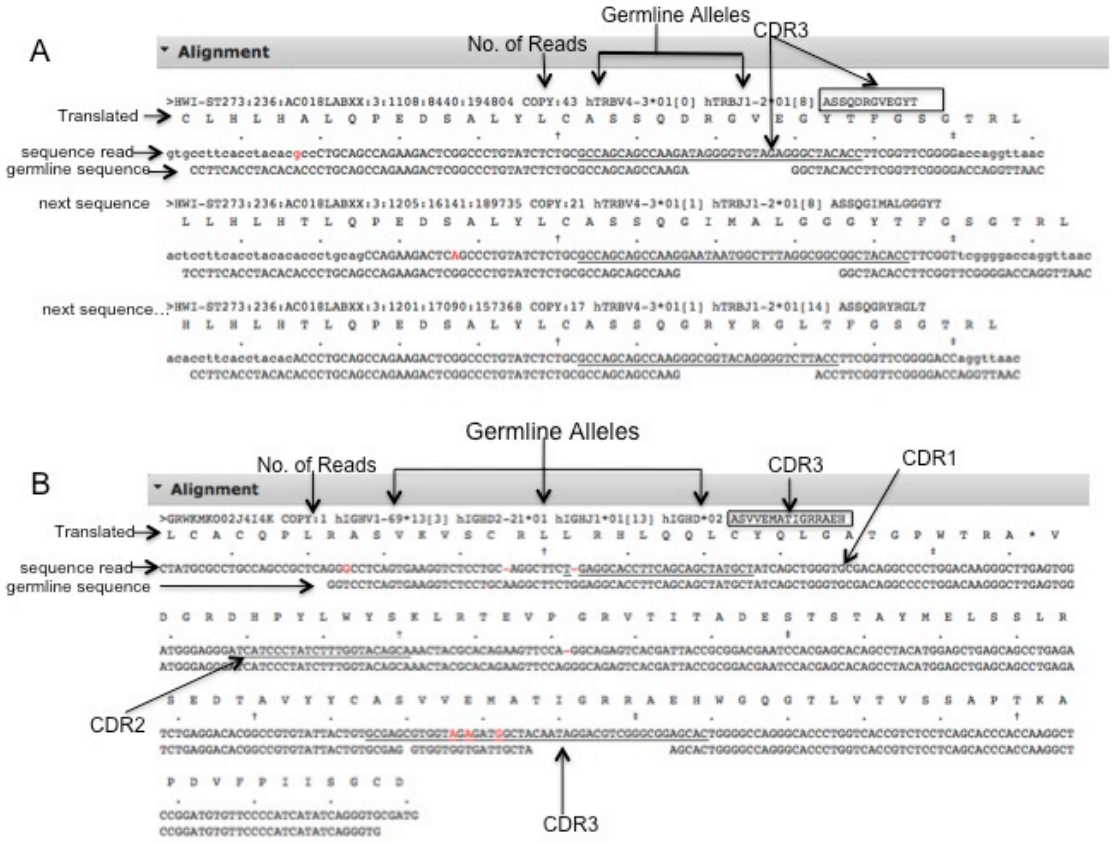


**Figure 3:** Partial alignment output when a square on the heat map is selected. The output from the Illumina HiSeq for TCR sequences covers approximately 150 base pairs surrounding the CDR3 (A). The output of the Illumina MiSeq (250-paired-end reads) or Roche 454 for BCR sequences covers CDR1, CDR2, CDR3, and the beginning of the C-region (B). Nucleotides highlighted in red are differences with the germline allele. In addition, the nucleic acid sequences associated with CDR1-3 are underlined. Every 10 nucleotides a "·" is placed above the nucleotide. Every 50 nucleotides a "†" symbol is placed, and every 100 nucleotides a "‡" is placed above the nucleotide.

# Analysis: Show 3D Map

Besides plotting the information as a heat map, there is an option of viewing the V-J frequencies in a three-dimensional plot. The construct is similar to the heat map; however, the frequencies are plotted as a bar graph with the read count of a particular sequence serving as the z-axis as shown in Figure 4. The V-J combination with the number of reads beyond the z limit has the read count in red above that particular bar. In order to observe only one specific V allele with the J alleles as a 3D map, or vice-versa, return to the heat map and select a particular V-allele column or J-allele column. A much smaller three-dimensional map will be generated showing the frequency for the selected V-allele with respect to the J-alleles as shown in Figure 5.

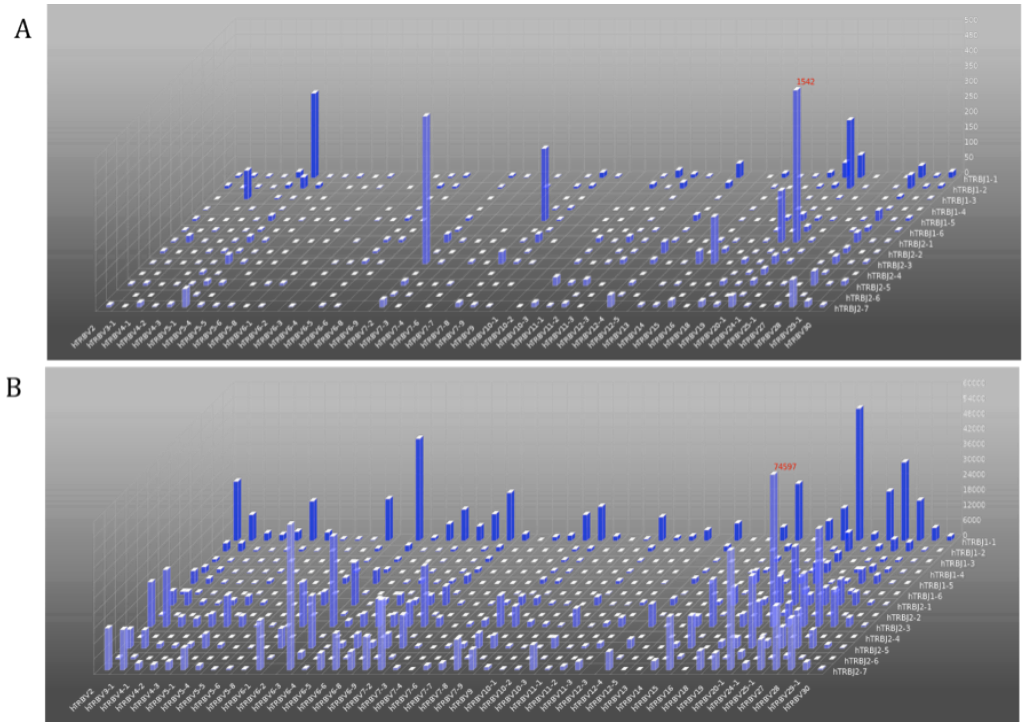


**Figure 4:** Three-dimensional map of the T-helper population of a colon cancer patient (A) and a normal patient (B).

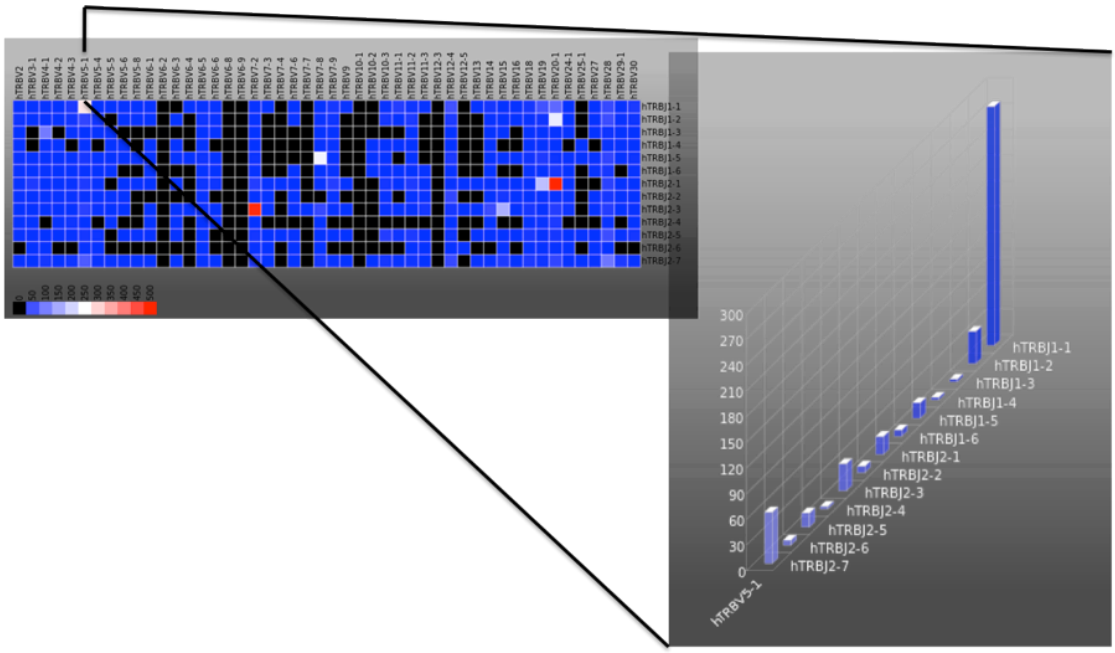


**Figure 5:** When a particular V-allele is selected on the heat map, a three-dimensional map of just the V-allele with respect to the J-alleles will be plotted.

# Analysis: CDR3 List and CDR3 Algebra

## CDR3 list

The CDR3 region is of particular interest to most researchers as the antigen-specificity is highly correlated with this region of the TCR or BCR [1]. Therefore, for a given library, a list of the CDR3s is provided as a sortable list as demonstrated in Figure 6. Once a particular CDR3 is selected, a detailed sequence list will appear, similar to the list provided by the heat map, showing representative sequences containing that particular CDR3.

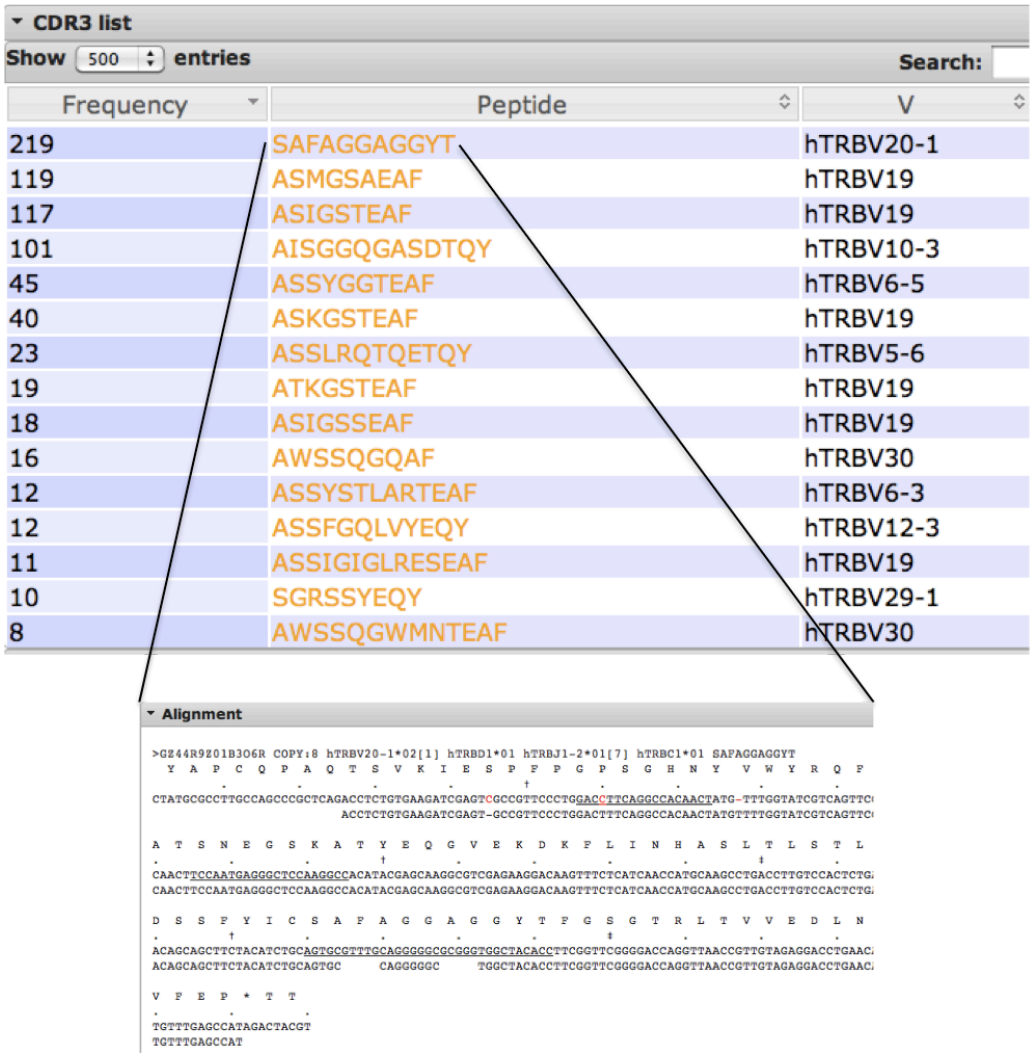


**Figure 6:** Sorted list of CDR3 sequences. Once a sequence is selected, a sequence alignment list (similar to the previous alignment in Figure 2) is displayed containing only sequences with that specific CDR3.

## CDR3 Algebra

A very convenient feature of the software is CDR3 Algebra, which allows the comparison of the CDR3 sequences from one data set to other data sets in order to identify shared CDR3s. When you select CDR3 algebra, a selection box will appear as shown in Figure 7. Sometimes you may need to scroll over to the right so that the selection boxes are visible.Select the data sets by clicking the boxes in the left column that you would like the current data set to be compared against. The data can be filtered by the frequency of a CDR3 so that only shared CDR3 sequences with a pre-set frequency are displayed. A sample output is demonstrated in Figure 7. A downloadable csv file is also produced which contains the shared CDR3 sequences. In addition, you can also exclude the CDR3 in a data set by selecting the data set from the right column. For instance, this is useful if you want to list the CDR3 shared among patients, but not found in healthy controls.

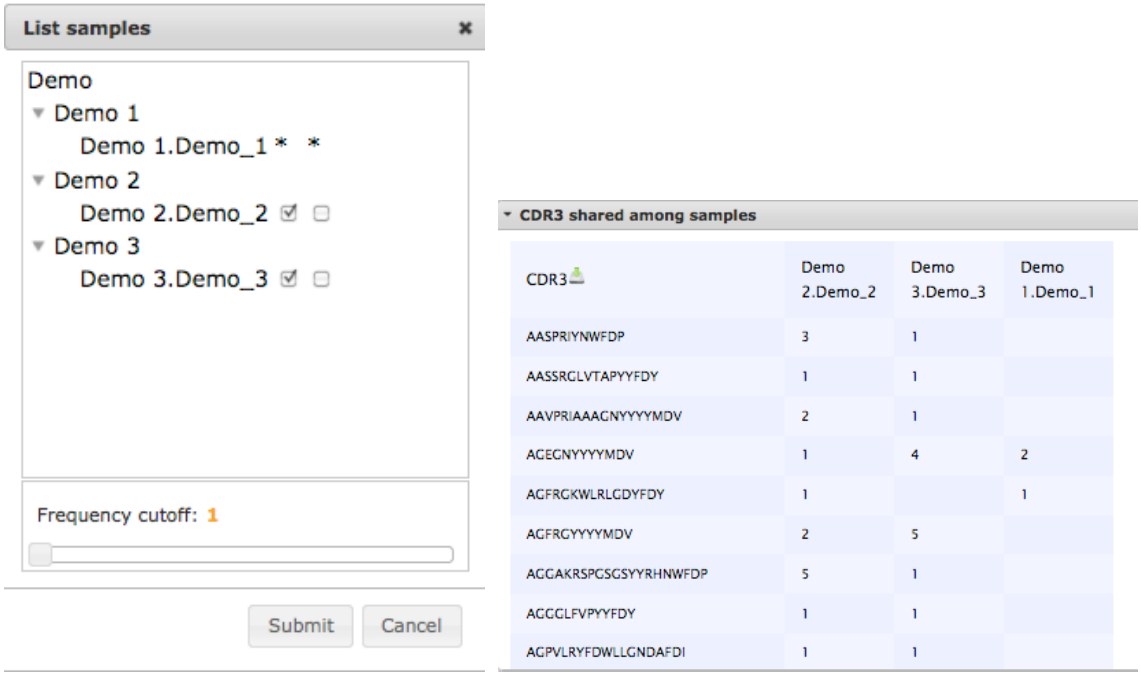


**Figure 7:** CDR3 algebra selection box and output. The left panel shows the selection box for data set comparison, while the right panel shows a sample output when shared CDR3s for three data sets are compared.

# Analysis: D50

In order to describe and compare the relative diversity of libraries, we have developed a proprietary analysis, termed D50, which assigns a single value that defines the diversity of a library. The D50 is a quantitative measure of the degree of diversity of T cells or B cells within a sample. The D50 is the percent of dominant and unique T or B cell clones that account for the cumulative 50% of the total CDR3s counted in the sample. The more diverse a library, the closer the value will be to 50. Low diversity values are associated with decreased diversity. In addition to the D50 value, a graphical display of the calculation is displayed in Figure 8.

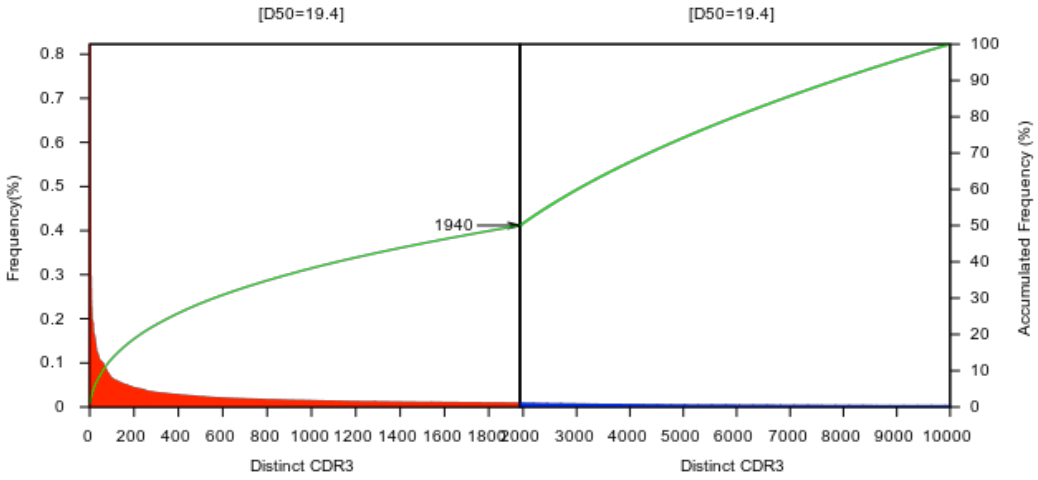


**Figure 8:** A graphical display of the D50 output.

# Analysis: Tree map

Tree map is another illustrative approach to show diversity. In a tree map, each rounded rectangle represents a unique entry: V-J-CDR3, where the size of a spot denotes the relative frequency as demonstrated in Figure 9. The entire plot area is divided into sub-areas according to V-usage, which is subdivided according to J-usage and then CDR3 frequency, subsequently. The unevenness of squares reflects the intrinsic bias of the underlying immunorepertoire. We typically do not use these maps for scientific representation as they are very difficult to read; however, they can be used for making interesting scientific repertoire art.

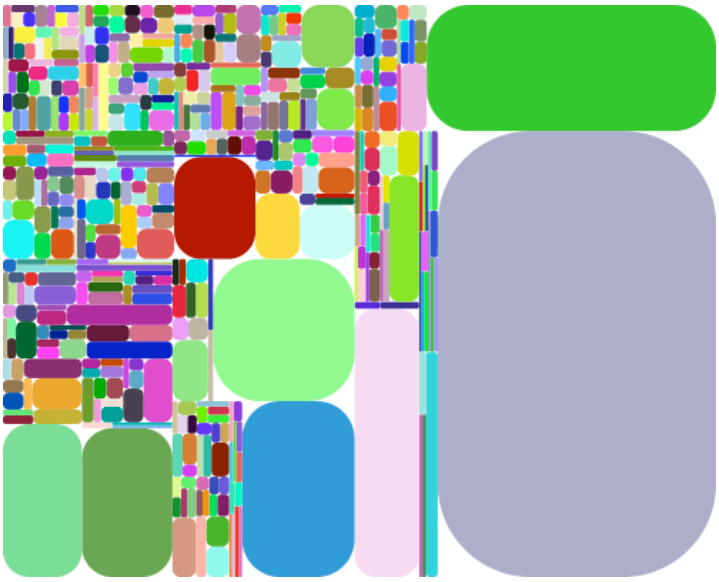

**Figure 9:** A sample output tree map of a T-helper population from a colon cancer patient.

# Distribution Analyses

The software also provides several types of distribution analysis including V-usage (Figure 10), J-usage, V-trimming (Figure 11), J-trimming, CDR3 length (Figure 12), and N-addition (Figure 13). The same analyses are also provided as normalized distributions. The difference between the regular distribution and normalized distribution is how the data are counted. The regular distribution is based on the number directly observed from the read count data. The normalized distribution counts the value (for V, J, N-addition, CDR3 length, etc.) of each distinct CDR3 as one, no matter how many of the particular CDR3s are observed.

### V-usage Example

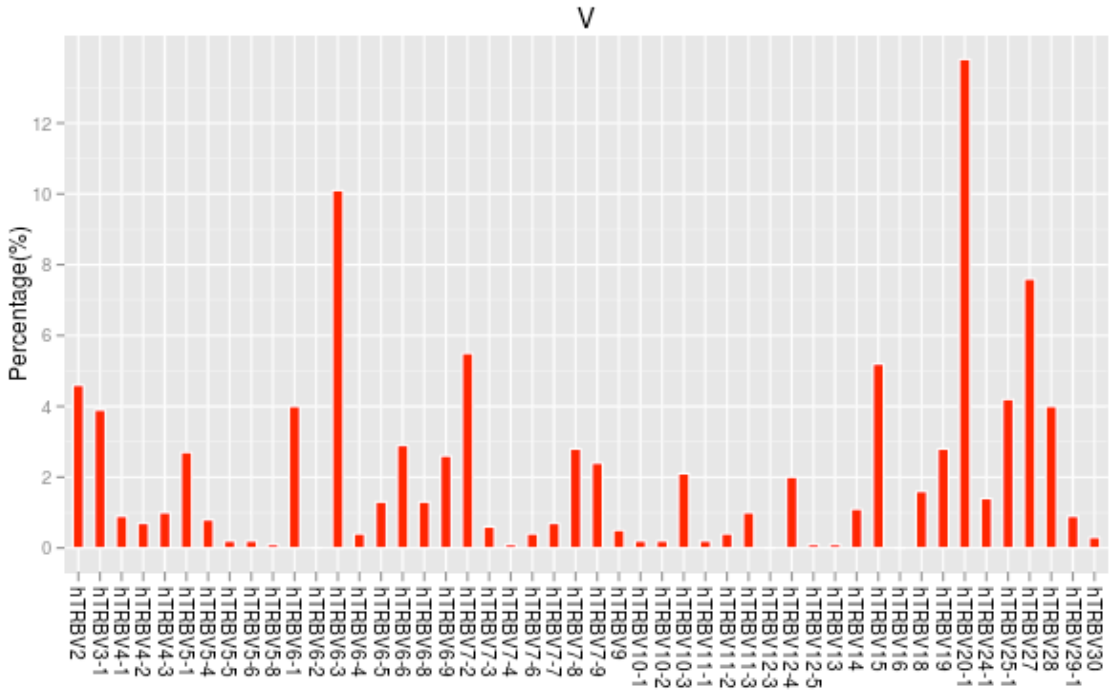


**Figure 10:** V-usage distribution. The percentage of reads containing the germline V-alleles are plotted so that it is simple to discern which V-alleles are used either frequently or infrequently.

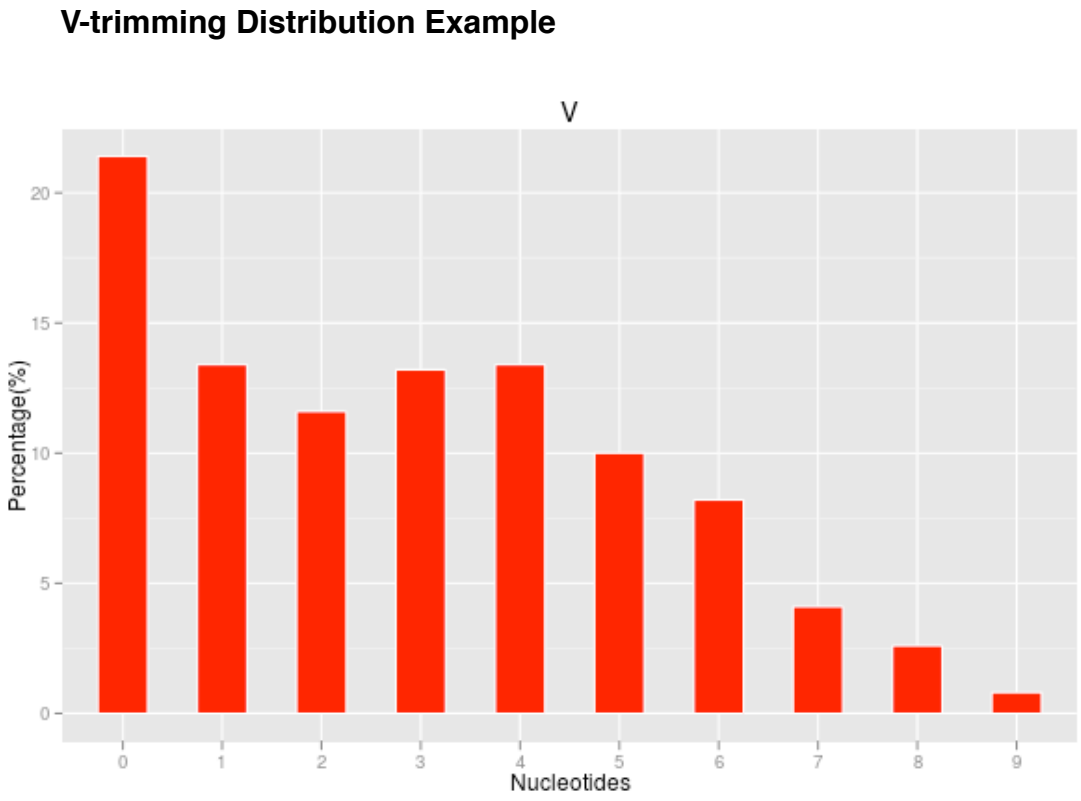


**Figure 11:** V-trimming distribution. The percentage of sequences with trimmed nucleotides on the V gene is displayed. For instance, the approximately 22% of sequences have no nucleotides trimmed from the V-gene, while about 12.5% have 1 nucleotide trimmed.

## CDR3 Length Distribution Example

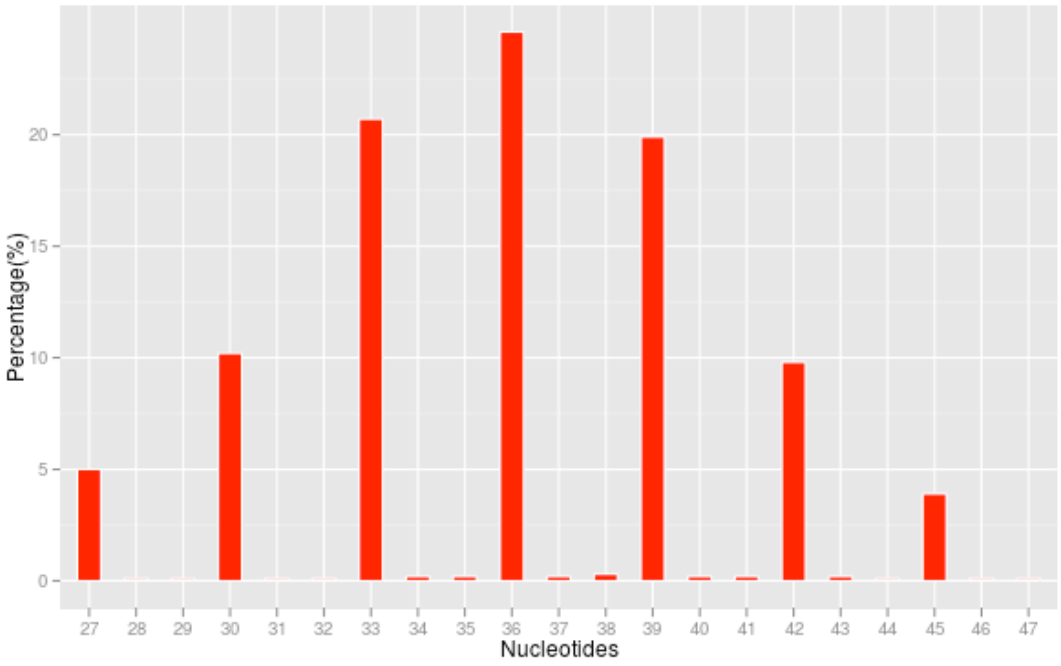


**Figure 12:** CDR3 length distribution. The plot demonstrates the distribution of nucleotides that comprise the CDR3 region. For instance, approximately 25% of CDR3 sequences are comprised of 36 nucleotides.

## N-addition Distribution Example

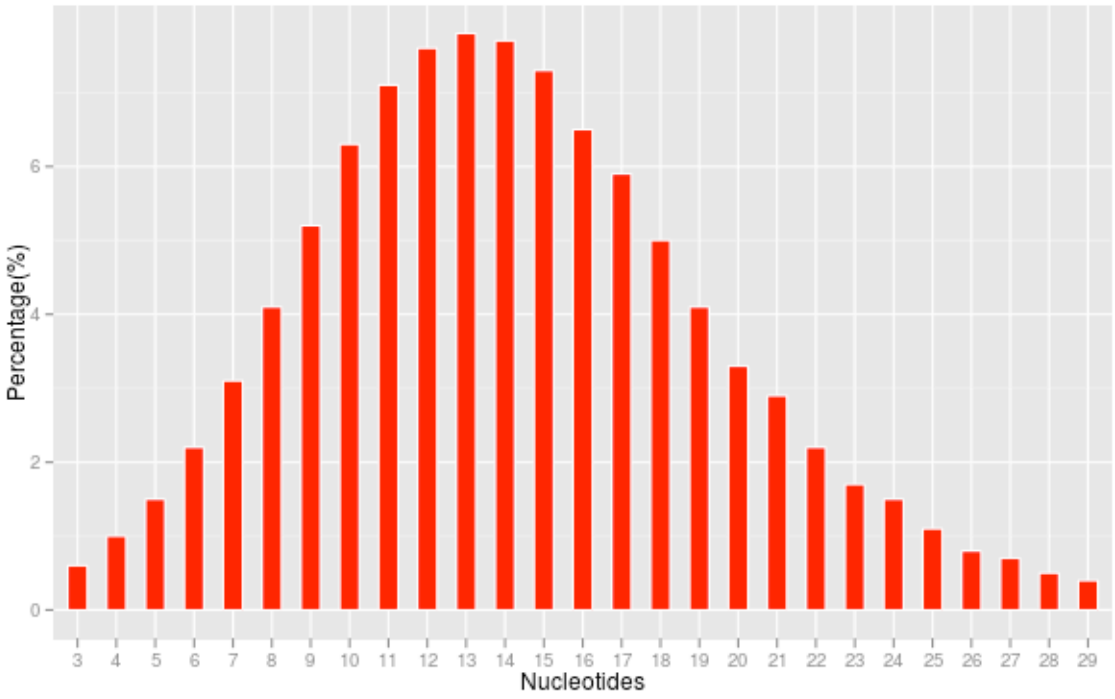


**Figure 13:** N-addition distribution**.** The plot demonstrates the distribution of nucleotides that are added in the process of N-addition.

# References

1. Xu JL, Davis MM: Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* 2000, 13(1):37-45.

We Find Health in Your Diversity.

WWW.IREPERTOIRE.COM