

iRweb: Data Analysis Guide

Immune Repertoire NGS Data Analysis for iRepertoire Reagent Systems



iRepertoire, Inc.

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 GAllx™ 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.

Version Updates

Update Classification	Update Description	Version
Minor	Updated styling	V20200916
Minor	Updated spelling errors and layout	V20201009

Table of Content

Introduction	5
Data Structure & Design	6
Logging In & Accessing Data	7
Sample Analysis Menu	9
2D Map	11
3D Map	14
List CDR3 new	15
List CDR3 old	16
List CDRs	16
CDR3 algebra	17
Diversity 50 (D50)	19
Diversity Index (Di)	21
Entropy (Shannon)	22
Tree Map	22
Distribution Analyses (V-gene, J-gene, Nucleotide trimming, etc.)	22
Normalized Distribution Analyses (V-gene, J-gene, Nucleotide trimming, etc.)	23
Raw Data	24
Raw Data: #####_pep.csv	25
Raw Data: Whole Project Download	26
Frequently Asked Questions	27
Contact Information	28



iRepertoire, Inc.


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 TCR data to remove errors that may have occurred during the amplification and sequencing process. For BCR data, the filtering is less (paired-end filter) due to the added complexity of hypermutation and N-addition. 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.

Things To Remember:

- iRepertoire's pipeline is designed only for use with data created with our reagent systems and cannot be used with sequencing data created by other methods.
- Throughout this guide, key words or URLs will be listed in pink.
- The IMGT database was used as a reference for both the creation of the reagent systems and of the pipeline. Reference data for iRweb is pulled from the IMGT database.
- Only genes whose designation within the IMGT database is 'functional' were used for iRweb analysis.
- All portions of iRweb outputs can be downloaded for use. Images and graphics can be downloaded with right-click, Raw Data contains raw formats of all of the data on the site, and tables can be downloaded by clicking the  icon.
- The Raw Data and F.A.Q. sections of this guide are designed to help investigators maximize the amount of information they discover from their sequencing data.
- Additional bioinformatic analysis, beyond what is output through iRweb, may incur additional charges.

Demo Version

To access the demo version of iRweb, please go to <https://irweb.irepertoire.com/nir/>. All data in the demo is available for download and manipulation at no cost to investigators. If you have any questions after consulting this guide, please email Customer Service at info@irepertoire.com.

Username: demo

Password: 12345

If login was successful, you will be able to see a header like that in [Figure 2](#).

Data Structure & Design

iRepertoire's primer design is built around the paired-end reads available with the Illumina HiSeq and MiSeq platforms. The term "paired-end read" or PER refers to the reading of both the forward and reverse template strands of the same receptor sequence during sequencing. The overall read length of the sequence can be increased by using the sequence read from both strands (with some overlap between both reads to increase confidence in the paired-read). We call this process read stitching.

For all of our V-C primer systems, Read 1 begins within the first part of the C-region and moves towards the V-gene. During Read 1, the molecular barcode used for demultiplexing samples is also read. Read 2 begins within the V and moves towards the C-region. The software pipeline first demultiplexes sequencing data based on molecular barcode and then stitches Read 1 and Read 2 in order to extend the sequencing coverage of the receptor sequence. What follows is a breakdown of read stitching for our human long-read primers. The stitching process is similar on the short read systems; however, the insert is approximately 150 bp, not 380 bp.

For human, the average MiSeq amplicon length is **500 bp including adaptors**. Subtract approximately 120 bp for the adaptors, and **the insert is on average 380 bp**. Remember that this value varies depending upon receptor editing.

Figure 1. Read stitching and amplicon size, as explained with our human long-read primers.

It will be mentioned a couple more times in this guide, but when reads are stitched within the analysis pipeline, stitched information is converted to uppercase and single-read information is in lower case. An example pulled from the [joinedSeq](#) column of the downloaded [Raw Data](#) directory for sample data follows:

```
agactctcctgtacagcgtctggattcacctttagcagctatgccatgagctgggtccgccaggctccaggggaaggggctggagtgggtctcagctattagtggtagtgggtgtagcacatactacgcagactccgtgaAGGGCCGGTTCACCATCTCCAGAGACAATTCCAAGAACACGCtgtatctgcaaatgaacagcctgagagccgaggacacggccgtatattactgtgcgagaagtctgtggtgactgccccgaagactactggggccaggggaaccctgggtcaccgtctcctcagggagtgcacccccaacccttttccccctcgtct
```

Logging In & Accessing Data

Logging In

Please log in to your account first by going to <https://irweb.irepertoire.com/nir/> . You should be brought to a screen like this, following a successful login:

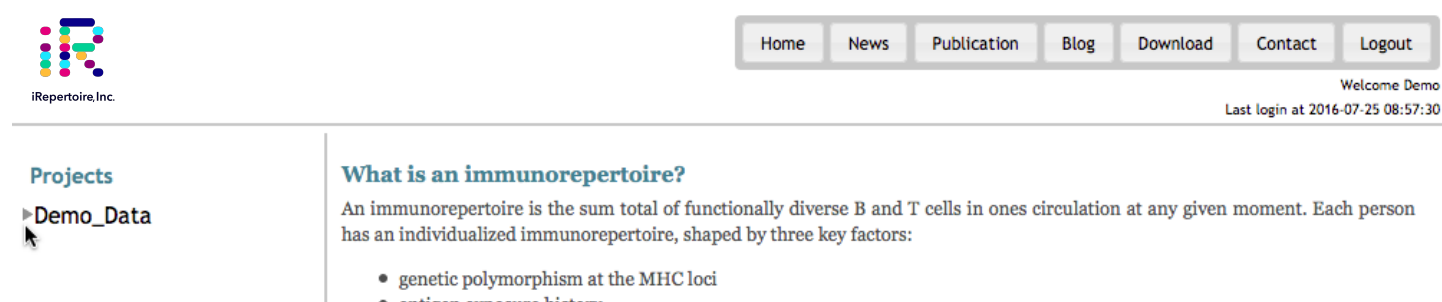


Figure 2. Main page after successful login; name and last login will be display on the right-hand side.

Accessing & Understanding Data

In order to access sample data on iRweb, it is necessary to use the left-hand menu. Clicking the **Project Name** will give users a summary chart of their data. The **Summary Table** (Figure 4) provides a generalized breakdown for each sample in the project. The more samples in the project, the larger this table becomes. Clicking the arrows to the left of the **Project**, **Study**, or **Sample** is what grants access to the analysis iRweb provides.

A note about naming (Figure 3): Unless otherwise specified, iRepertoire will create a standard Project Name that includes the Institution and the Principal Investigator's last name. Investigators have the ability to designate Study Names and Sample Names.

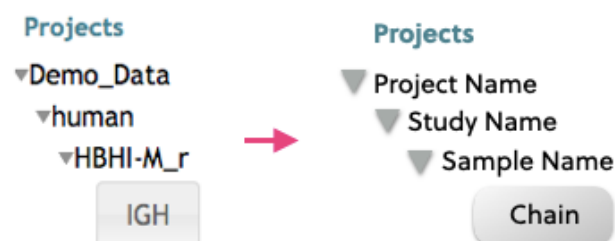


Figure 3. Break down of the left-hand menu and its levels.

Summary Chart Breakdown

A generalized summary of each sample is given by clicking the **Project Name** in the left-hand menu of iRweb. This the most upper-level way of looking at sample data. It is necessary to click the arrow in the left-hand menu to access sample data in its more complex forms.


Project 	Study	Sample	Sample id	Species	Chain	Reads	CDR3	Unique CDR3	D50	Entropy
Demo_Data	human	HBHI-M	21747	h	IGH	1420700	1403796	106292	28.2	13.0
Demo_Data	human	HBHI-M_r	21748	h	IGH	1502092	1485800	104070	28.4	13.0
Demo_Data	human	HBKLI-M	21744	h	IGL	130838	123272	3366	0.9	8.0
Demo_Data	human	HBKLI-M	21746	h	IGK	1268618	1204844	11493	0.5	8.5
Demo_Data	human	HBKLI-M_r	21743	h	IGL	54293	50444	3464	1.1	8.5
Demo_Data	human	HBKLI-M_r	21745	h	IGK	1142987	1083301	14869	0.5	8.9
Demo_Data	human	HTAlvc_01	21583	h	TRA	2249690	2048876	124357	17.7	12.5
Demo_Data	human	HTBI-M	21014	h	TRB	832285	827305	72352	0.3	9.2
Demo_Data	human	HTDI-M	21552	h	TRD	197342	184142	4631	3.0	9.6
Demo_Data	human	HTGI-M	21551	h	TRG	93685	57117	2221	1.1	8.0
Demo_Data	mouse	MBHI-M_SortedCell	21555	m	IGH	340136	330852	12195	4.4	10.7
Demo_Data	mouse	MBKLI-M-05	21462	m	IGL	624207	483743	971	0.1	1.2
Demo_Data	mouse	MBKLI-M-05	21464	m	IGK	19113	18215	1250	2.1	7.4
Demo_Data	mouse	MTBI-M	21013	m	TRB	1007946	964388	96716	32.3	13.1
Demo_Data	mouse	MTDI-M	25671	m	TRD	46231	35176	2399	11.1	10.2
Demo_Data	mouse	MTGI-M	25670	m	TRG	251950	158619	1755	0.6	6.1
Project	Study	Sample	Sample id	Species	Chain	Reads	CDR3	Unique CDR3	D50	Entropy

Figure 4. Summary Table.

- **Sample id.** Generated and assigned automatically by the pipeline as samples complete analysis. Can be used as a reference for the sample in addition to the Sample Name. Clickable for downstream sample analyses.
- **Species.** iRepertoire currently offers systems for two species - human (h) and mouse (m).
- **Chain.** There are seven chains for the six products currently available: TRB, TRA, TRG, TRD, IGH, IGK, IGL
- **Reads.** The number of sequencing reads for the library that passed demultiplexing and filters.
- **CDR3.** The number of CDR3s captured within the library.
- **Unique CDR3s.** The number of unique peptide CDR3s within the sample.
- **D50.** A proprietary diversity index. The closer the value is to 50, the more diverse the sample.
- **Entropy.** This value is the calculated Shannon entropy for the sample.



Sample Analysis Menu

Once a sample has been selected, iRweb should open up a new window in your browser, with an updated left-hand menu. This menu contains all of the data manipulations available for the sample data on the website as well as the **Raw Data**, which enables users to download data to perform more advanced functions.

Query CDR3 peptide:

If the peptide version of the CDR3 is known, it is possible to use this function to search data available on iRweb.

Summary

Project: Demo_Data
Study: human
Sample: HBHI-M
Description:
Chain: IGH
Reads: 1420700
total CDR3: 1403796
distinct CDR3: 106292
D50: 28.2

This data is a reiteration of the information seen in the **Summary Table**, but is specific to the sample selected.

Raw Data

This will lead to a download of a .zip file containing the raw files used to generate all of the information present on iRweb. This can be downloaded and parsed to be placed through other bioinformatic suites and analysis tools. More information about **Raw Data** can be found on page 24).
NOTE: This does not contain the raw sequencing data, nor demultiplexed data for the sequencing run. It is necessary to contact Customer Service to get this data.

Analysis

Show 2D map

2D heat map of V- & J-gene combinations in relation to frequency (page 11)

Show 3D map

3D chart of V- & J-gene combinations in relation to frequency (page 14)

List CDR3 new

List of CDR3s w/ hierarchical tree maps for V(D)J-C combinations (page 15)

List CDR3 old

Another iteration of V-J gene combinations by peptide (page 16)

List CDRs

A list of all called CDRs for a given peptide sequence (page 16)

CDR3 algebra

Shared CDR3 calculations across samples (page 17)

Compute D50

iRepertoire's own diversity calculation: D50 (page 19)

Tree Map

A graphical representation of the V-J combinations (page 20)



Distribution

- V usage
- J usage
- V trimming
- J trimming
- CDR3 length
- N addition

A percentage use of V-genes within the sample (page 20-21)

A percentage use of J-genes within the sample (page 20-21)

V-gene nucleotides trimmed through recombination events (page 20-21)

J-gene nucleotides trimmed through recombination events (page 20-21)

CDR3 lengths calculated for the sample (page 22)

Nucleotide additions within the sample based on references (page 22)



Normalized

- V usage
- J usage
- V trimming
- J trimming
- CDR3 length
- N addition

More information about **Normalized** data will be given on page 20. In short, each uCDR3-VDJ combination is treated as a quantity of 1 regardless of read count, and then analyzed for V usage, J usage, etc.

Analysis: Show 2D Map

Figure 5 is an example of a two-dimensional heat map from the T helper population of both a colon cancer patient and a normal control. The relative frequency of a consensus germline V-gene allele (as per alignment with the IMGT database) is plotted relative to the consensus 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. There are two points of interactivity within the 2D Map:

1. Along the top and right axes, specific genes can be selected and a 3D bar graph of the V-/J-gene combinations for that line will appear. This works for either direction. Figure 6.
2. 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 7. Many sequences may appear in this output list because it contains representative 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 (red), and identification of [CDR1, CDR2, and (long read)] CDR3 (underlined). In addition, the CDR3 sequence will also be listed in the FASTA-like header for the sequence.

Figure 5. Heat map of the T-helper population of a colon cancer patient (A) and normal patient (B).

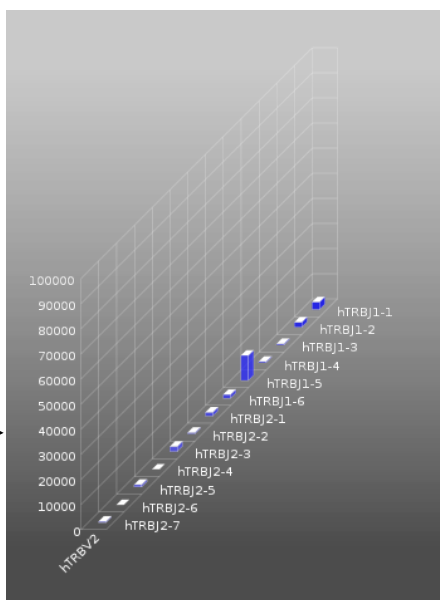
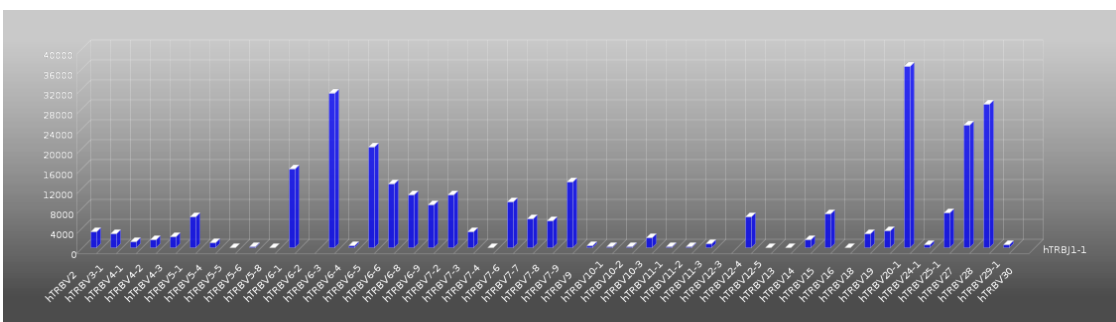
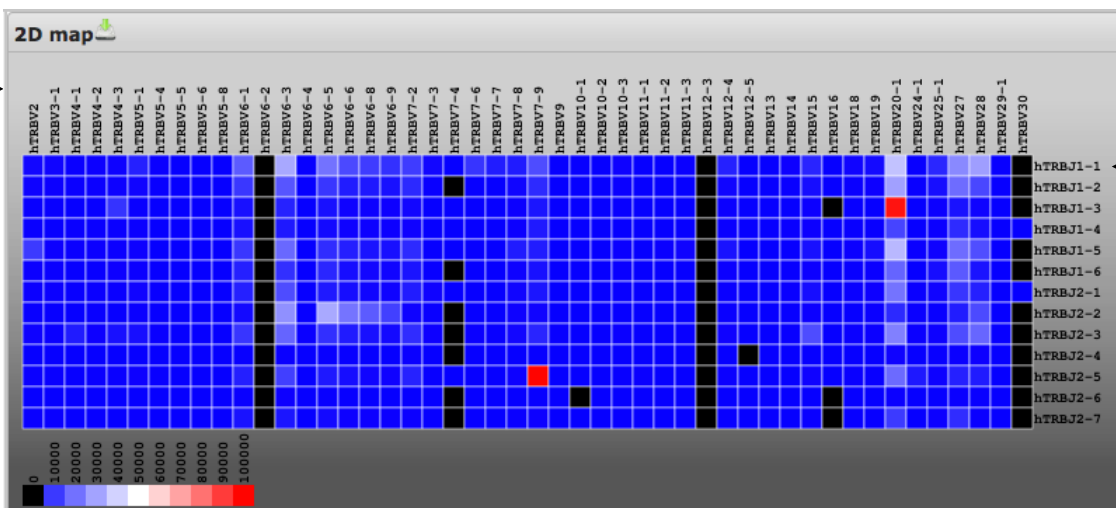


Figure 6. The selection of a gene within either axes will provide a 3D bar graph of the V-J combinations pertinent to the selected gene. It is possible to right-click and select “View Image” in most browsers to download the final image of either 3D representation. The 2D heat map can be downloaded by clicking the download icon above the map.



The identification of CDR1 through CDR3 will depend on the sequencing method utilized. There are two types of sequencing available, long-read 250/300 paired-end reads (PER) and short-read 100/150 paired-end reads (PER). The short-read primer systems capture about 150 base pairs of sequence data around the CDR3. The long-read primer systems allows for the sequencing of around 350 bp around the CDR3 and is better suited for BCR sequencing because it can provide sufficient information about the CDR1 to CDR3 region with hypermutation patterns with human samples. As demonstrated in Figure 6, information pertaining to CDR1, CDR2, and CDR3 is displayed for the human MiSeq 250-PER. For mouse systems, CDR2 and CDR3 are available with the long read primer sets. All reagent systems and sequencing platforms allow for the identification of unique CDR3s *at minimum*.

Figure 7. Partial alignment output when a square on the heat map is selected. The output from the short-read primers for TCR sequences covers approximately 150 base pairs surrounding the CDR3 (A). The output of the long-read primer systems (250/300-paired-end reads) 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. Upper case bases are those in the overlapping region which are identical at both forward and reverse reads. Lower case bases are otherwise. For single-end reads, all bases are upper case.

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 8](#). 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 ([Figure 6](#)). 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](#). There is an availability to set the upper frequency limit displayed in the 3D map. This allows for the close-up examination of V- and J- gene combinations that have lower levels of expression. The numbers present in red above each column are the levels of expression beyond the threshold ([Figure 9](#)).

Figure 8. Three-dimensional map of the T-helper population of a colon cancer patient (A) and a normal patient (B).

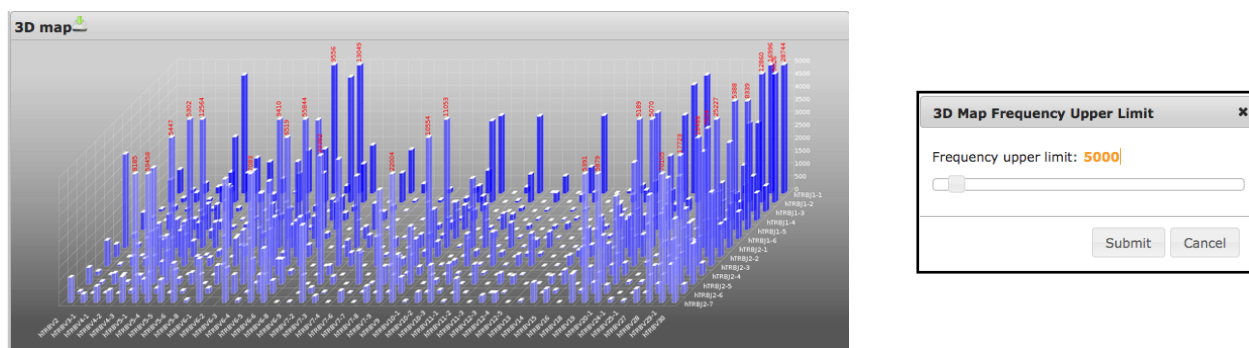


Figure 9. When the 'frequency upper limit' is set, it is possible to see more detail on V-J combinations that have lower levels of expression in the 3D map.

Analysis: List CDR3 new

“List CDR3 new” is the second-generation function of viewing CDR3 derivations. In addition to seeing the peptide sequence, it is possible to get access to a hierarchy of gene combinations that lead to the peptide sequence of the CDR3. At the end of each branch of the tree is a clickable red dot that will take you to the top ten alignments for each V(D)J-C branch for a given peptide sequence.

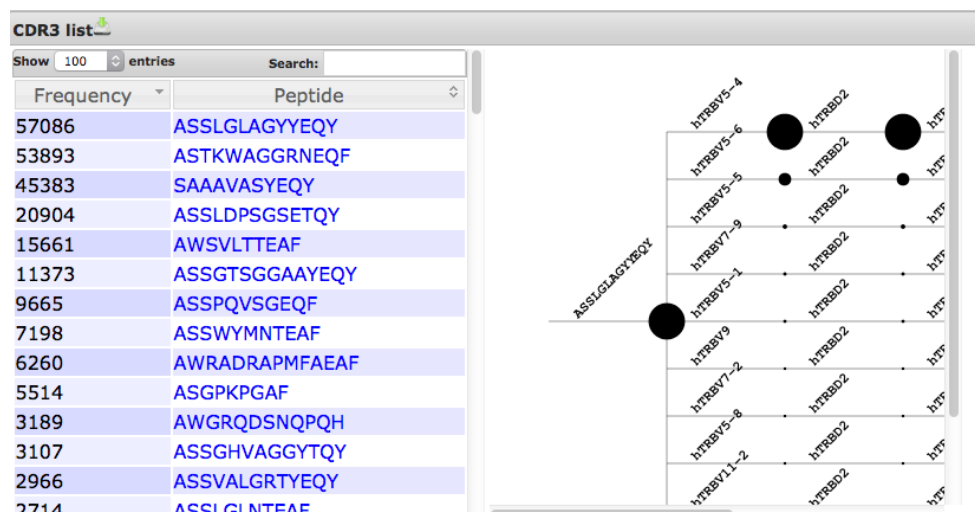


Figure 10. List CDR3 new provides the user with a breakdown of all of the gene combinations available in the data that produce a given peptide CDR3. An expanded version of the hierarchical tree map can be seen in [Figure 11](#).

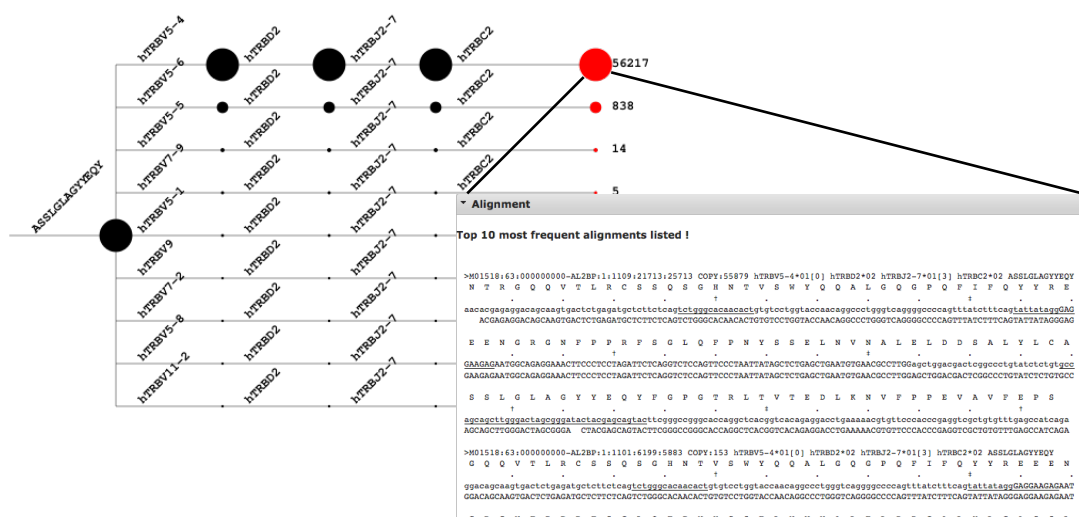
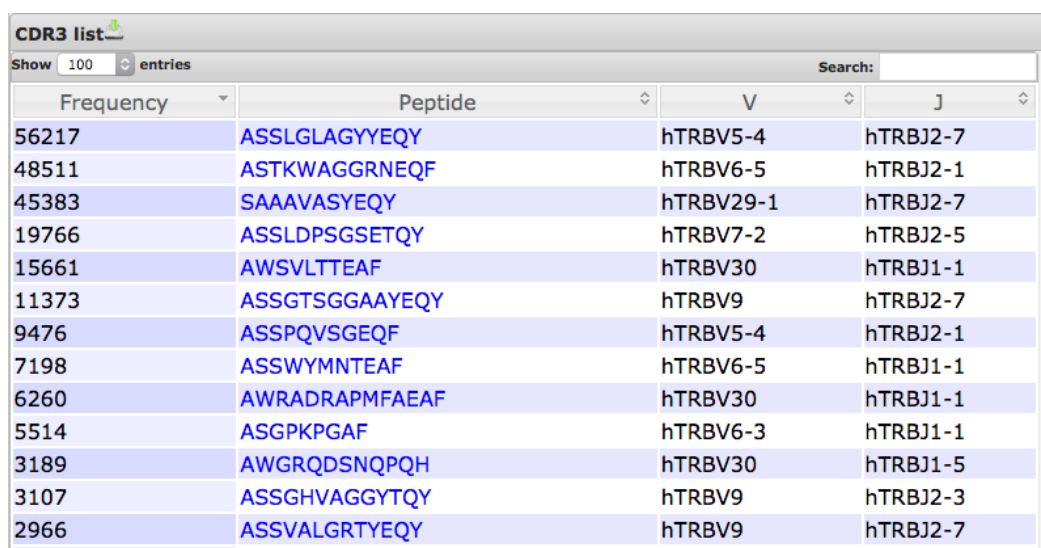


Figure 11. Selecting any of the red dots at the very end of a branch of the hierarchical tree map will provide a Top 10 most frequent alignments, as one would find upon clicking a particular gene combination within the 2D/3D maps.

Analysis: List CDR3 old

The List CDR3 old table is the first generation of looking at V- & J-gene combinations, and it displays the CDR3s ranked by frequency with the specific V- and J-gene used to create that CDR3. **Since different V- and J- combinations can result in the same CDR3, the same CDR3 can show up on the list multiple times (associated with a different V- and J gene combination).** However, the CDR3 peptide sequence on the list is clickable. When the CDR3 peptide sequence is selected, an alignment of sequences opens which contains representative sequences containing that CDR3. Therefore, in the alignment, the CDR3 *with various V- and J-alleles* used to create that CDR3 will be demonstrated (not just the CDR3 with that specific V-J combination - see new CDR3 list for this type of breakdown). The frequency on the table will likely not match the frequency of the alignment reads (found by clicking on the CDR3) because of this distinction between the List CDR3 old table and the alignment displayed upon clicking on the uCDR3 peptide (same uCDR3 with different V- and J- genes used to form that uCDR3).



Frequency	Peptide	V	J
56217	ASSLGLAGYYEQY	hTRBV5-4	hTRBJ2-7
48511	ASTKWAGGRNEQF	hTRBV6-5	hTRBJ2-1
45383	SAAAVASYEQY	hTRBV29-1	hTRBJ2-7
19766	ASSLDPSGSETQY	hTRBV7-2	hTRBJ2-5
15661	AWSVLTEAF	hTRBV30	hTRBJ1-1
11373	ASSGTSGGAAYEQY	hTRBV9	hTRBJ2-7
9476	ASSPQVSGEQF	hTRBV5-4	hTRBJ2-1
7198	ASSWYMNTEAF	hTRBV6-5	hTRBJ1-1
6260	AWRADRAPMFAEAF	hTRBV30	hTRBJ1-1
5514	ASGPKPGAF	hTRBV6-3	hTRBJ1-1
3189	AWGRQDSNQPH	hTRBV30	hTRBJ1-5
3107	ASSGHVAGGYTQY	hTRBV9	hTRBJ2-3
2966	ASSVALGRTYEQY	hTRBV9	hTRBJ2-7

Figure 12. List CDR3 old provides the main V-J gene combination that led to the amino acid peptide CDR3.

Analysis: List CDRs

This function is a simple column listing of all called CDR1s, CDR2s, and CDR3s, listed by peptide sequence and frequency.

Analysis: CDR3 Algebra

CDR3 algebra allows for the calculation of shared CDR3s across samples

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. This allows for a comparison amongst disease state samples and controls or for a comparison amongst time points during treatment. When you select CDR3 Algebra, a selection box will appear as shown in **Figure 13**. 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 to. The data can be filtered by the frequency of a CDR3 so that only shared CDR3 sequences with a pre-set frequency in the original data are displayed.

With CDR3 algebra, there is also an “exclusion of” function, which is useful for listing the CDR3s shared among patients, but not found in controls. This allows you to exclude the CDR3 found in a data set by selecting the data set from the right column or right box. For instance, if two samples were disease samples and one was a control, you could ask for the sharing between the two disease samples by clicking the left boxes for those two samples. However, you may not want to see the CDR3s if they are also in your control sample. Therefore, for the control sample, you would select the right box and click submit.

More recently, a version of sharing has been created that can be accessed by clicking “Extra” prior to clicking “Share.” This Extra function preserves the original top 100 unique CDR3s of a sample, whether they are shared or not. In the .CSV that is output by the function, one need only sort in descending order for a given sample to get the original top 100 unique CDR3s for a sample. To look at those uCDR3s that are only shared, simply click the “Share”, avoiding the “Extra” button.

*All CDR3 frequencies are now artificially scaled to 10 million reads to account for differences in read depth among samples, making comparisons between samples easier. Please see **Table 1** for more details. A downloadable .csv file is also produced which contains the shared CDR3 sequences.*

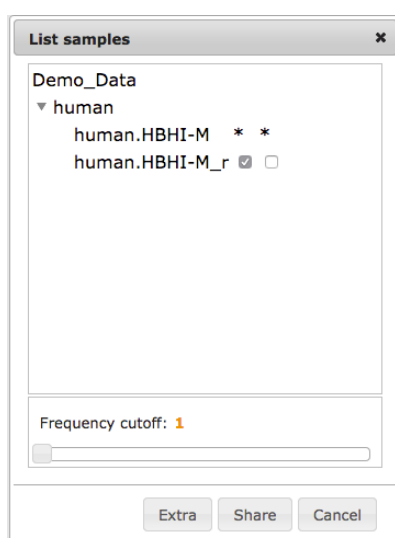


Figure 13. CDR3 algebra selection box. Selecting the left-hand boxes for samples will include all unique CDR3s that are shared. If the right-hand box is selected, CDR3s that are shared are excluded from the generated .csv. Clicking “Extra” will incorporate the top 100 unique CDR3s from each sample, shared or not.

Original Frequency of example uCDR3 in Sample	Original No. of Reads for the Entire Sample	Normalization Factor	Normalized Frequency of uCDR3 calculated through CDR3	Normalized No. of Reads
500	1,000,000	10	5,000	10,000,000
1,500	250,000	40	60,000	10,000,000
2,500	1,000,000	10	25,000	10,000,000

Table 1. Values within CDR3 algebra are not 1:1 with the original frequencies listed in the sample. For each sample, the reads are placed on a notionally common scale by scaling the total reads of a data set to 10 million and scaling the frequencies of uCDR3s within the sample accordingly. A CDR3 within a sample that contained 1 million reads, has its frequency artificially increased 10-fold. This is to account for read depth differences among samples and to facilitate easy comparison of uCDR3s across samples (on the same scale).

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. There are two algorithms in which the D50 is calculated: one with unique CDR3s below 10,000 and one with uCDR3s above 10,000. Example calculations are shown below.

Instances with uCDR3s above 10,000 (see Figure 14)

Unique CDR3s are arranged by rank dominance based upon frequency. The top 10,000 unique CDR3s are selected, and the number of reads from these uCDR3s is totaled. In this example, the distinct number of uCDR3s is 14,971. The sum of the number of reads associated with the top 10,000 uCDR3s is 2,909,346 (as counted from the List CDR3 new). 50% of this is 1,454,673 reads. Between 27 & 28 unique CDR3s are contained within those 1.45 million reads, so the D50 calculation is as follows:

$$(28 * 100) / 10,000 = 0.28 \text{ (or } 0.3)$$

$$(\text{No. of uCDR3s that make up 50\% of the reads of the top 10k uCDR3s} * 100) / 10,000 = \text{D50}$$

Instances with uCDR3s below 10,000 (see Figure 15)

Unique CDR3s are arranged by rank dominance based upon frequency. The number of unique CDR3s and total number of reads are used with this calculation (as counted from the List CDR3 new). 50% of the total reads for the sample is 176,430 reads. Between 22 & 23 unique CDR3s are contained within those ~176k reads, so D50 is calculated thusly:

$$(23 * 100) / 2,963 = 0.77 \text{ (or } 0.8)$$

$$(\text{No. of uCDR3s that make up 50\% of the total reads} * 100) / \text{No. of uCDR3s} = \text{D50}$$

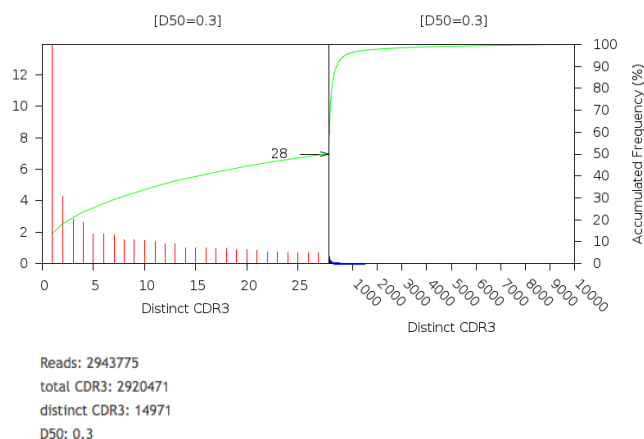


Figure 14. D50 calculation chart, as seen on iRweb for a sample containing more than 10,000 uCDR3s.

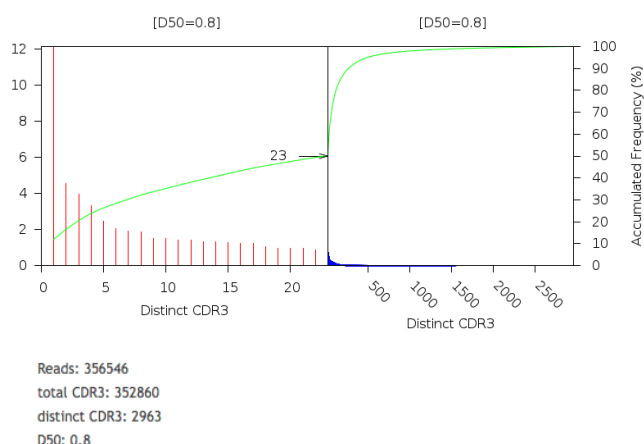


Figure 15. D50 calculation chart, as seen on iRweb for a sample containing less than 10,000 uCDR3s.

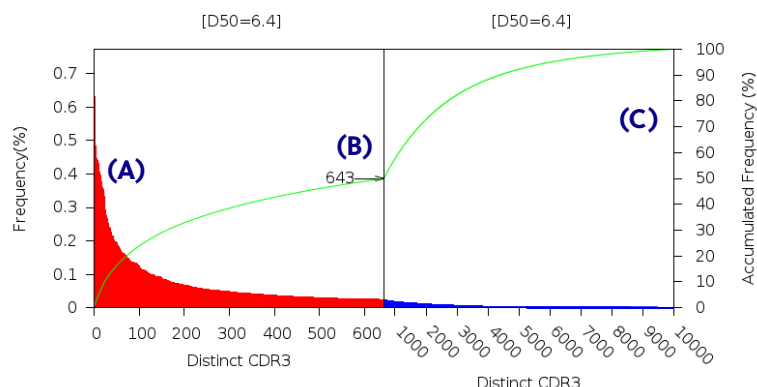


Figure 16. D50 graphic broken down - see descriptions below.

Analysis: D50 graphic

The two panels of a D50 graphic conceptualize the data in two different, yet related ways:

Left-hand panel

The red bars represent the % of the number of reads for the top uCDR3s for 50% of the top 10k uCDR3s that make up all of the top 10k uCDR3s. In the case of Figure 16, the peptide ARDPSSGWYGDDY (not pictured) is the most dominant clone with a frequency of 6,727. When this frequency is counted in relation to the number of reads making up all of the reads of all of the top 10k uCDR3s, it is 0.77%.

$$(6727/870505)*100 = 0.77\% \text{ (red values)}$$

Right-hand panel

The blue bars represent the % of the number of reads for each uCDR3 in relation to the uCDR3s that make up 50% of the top 10k uCDR3s. In the case of Figure 16, the peptide ARDPSSGWYGDDY (not pictured) has a frequency of 6,727. When this frequency is counted in relation to the number of reads making up the top 50% of the top 10k uCDR3s, it is 1.55%.

$$(6727/435388)*100 = 1.55\% \text{ (blue values)}$$

Analysis: Diversity Index

An explanation of Diversity Index charts

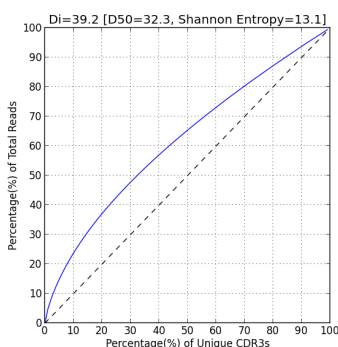
The latest feature of iRweb, the Diversity Index (Di), includes the Di value, D50, and Shannon entropy for a given sample. The Di is defined mathematically as

$$\text{Assume that } \underbrace{r_1 \geq r_2 \geq \dots \geq r_i \geq r_{i+1} \geq \dots \geq r_n}_n,$$

where r_i is the frequency of the i -th CDR3 and n is the total number of unique CDR3s

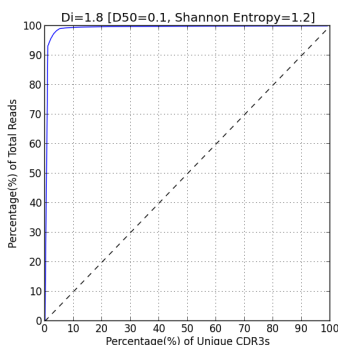
$$x_k = \frac{k}{n}, \quad y_k = \frac{\sum_{i=1}^k r_i}{\sum_{i=1}^n r_i}$$

The line assembles a curve that describes the overall diversity of the sample with “perfect” diversity being the black dashed line in the middle (each unique clonotype receives equivalent reads, i.e. no clonal expansion or dominant clone). The higher the diversity of the sample, the closer the blue solid line is to the dashed lines. The examples below are from a few of our Demo_Data samples available on iRweb upon request. We’ve provided the Di, as well as the top 10 CDR3s, their frequencies, and their total read counts to help elucidate.



CDR3	MTBI-M
ASSDGEQY	289
ASSPGTTNTEVF	276
ASSDSAETLY	270
ASRDNSGNTLY	240
ASSGTANTEVF	234
ASSANTEVF	233
ASSPGTANTEVF	231
ASSPGQNTTEVF	228
ASSDRNTEVF	225
ASSLGQANTEVF	219

Project: Demo_Data
Study: mouse
Sample: MTBI-M
Description:
Chain: TRB
Reads: 1007946
total CDR3: 964388
distinct CDR3: 96716
D50: 32.3



CDR3	MBKI-M
GVGDTIKEQFVYV	424467
ALWYSNHWV	7820
GVGDTIKEHYV	3199
GVGDTIKEQFVYV	3168
ALWYSNHFI	2605
GVGDTIKEQFDYV	2586
ALWYSNHLV	2044
GVGDTIKEQFYV	1863
GVGDTIKEQFVYV	1574
GVGDTIKGQFVYV	1537

Project: Demo_Data
Study: mouse
Sample: MBKI-M-05
Description:
Chain: IGL
Reads: 624207
total CDR3: 483743
distinct CDR3: 971
D50: 0.1

Analysis: Entropy

A calculation for Shannon entropy is given in the [Summary Table](#) for each [Project](#). The formula used in the calculation of the Shannon entropy:

$$- \sum_i^{10000} p_i \log_2 p_i$$

Note: Only the top 10000 CDR3 are included into this calculation, where P_i is the frequency of i th CDR3 within the top 10000 CDR3 (in other words, 10001th CDR3 and beyond are excluded in the calculation for P_i).

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-uCDR3, where the size of a spot denotes the relative frequency as demonstrated in [Figure 17](#). The entire plot area is divided into sub-areas according to V-usage, which is then subdivided according to J-usage, and then each uCDR3 within a given V-J- combination is subsequently represented by a rounded rectangle (sized by frequency). The unevenness of squares reflects areas of clonal expansion within the immune repertoire sampled.

Figure 17: 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 18](#)), J-usage, V-trimming ([Figure 19](#)), J-trimming, CDR3 length ([Figure 20](#)), and N-addition ([Figure 21](#)). 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.

Distribution Analyses: Normalized

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. In short, each uCDR3-VDJ combination is treated as a quantity of 1 regardless of read count, and then analyzed for V usage, J usage, etc. This allows for a view of the repertoire removing the skewing which may occur due to one or just a few highly dominant clones.

V-usage Example

CDR3 Length Distribution Example

Figure 18. 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 20. 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.

V-trimming Distribution Example

N-addition Distribution Example

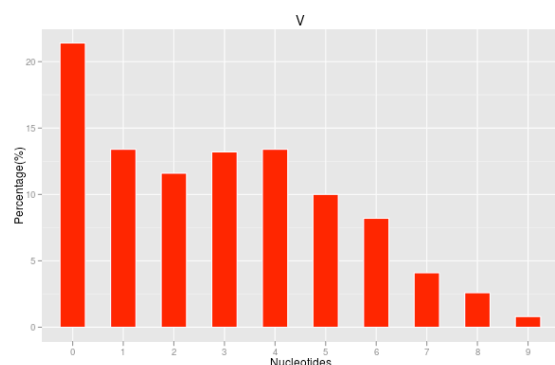


Figure 19. V-trimming distribution. The percentage of sequences with trimmed nucleotides on the V gene is displayed. For instance, the approximately 22% of sequences have no nucleotides trimmed from the V-gene, while about 12.5% have 1 nucleotide trimmed.

Figure 21. N-addition distribution. The plot demonstrates the distribution of nucleotides that are added in the process of N-addition.

Raw Data

Raw Data contains all of the documents necessary to replicate the tables and charts used in iRweb. It makes it possible for investigators to recreate their own graphics or tables in Excel or perform additional analyses with other software packages with the data received from iRepertoire. Raw Data is available only on a per sample basis, but it is possible to contact Customer Service and put in a request for our Bioinformatics staff to pull the Raw Data for the entire Study or Project. There are 16 files for each Sample ID. In the list below, ##### will be the Sample ID number.

Note: A "0" in the file name represents non-normalized data, while a "1" represents normalized (counting each V-J-uCDR3 combination as a frequency of 1 despite read depth).

#####_0_CDR3Length	A non-normalized version of the CDR3 lengths in the sample.
#####_1_CDR3Length	A normalized version of the CDR3 lengths in the sample.
#####_0_Naddition	A non-normalized version of the nucleotide addition seen in the sample.
#####_1_Naddition	A normalized version of the nucleotide addition in the sample.
#####_CDR3_list_1	The equivalent of List CDR3 new; only the CDR3 peptide and frequency are listed.
#####_CDR3_list_2	The equivalent of List CDR3 old; CDR3 peptide, V-gene, J-gene, and frequency are listed.
#####_CDRs	A .csv containing the relative frequency of unique peptide CDR3s with their associated CDR1 and CDR2 peptide sequences.
#####_J_0_trim	A non-normalized version of the J-gene trimming chart.
#####_J_0_usage	A non-normalized version of the J-gene usage of the sample.
#####_J_1_trim	A normalized version of the trimming for J-genes in the sample.
#####_J_1_usage	A normalized version of the J-gene usage of the sample.
#####_V_0_trim	A non-normalized version of the V-gene trimming chart.
#####_V_0_usage	A non-normalized version of the V-gene usage of the sample.
#####_V_1_trim	A normalized version of the trimming for V-genes in the sample.
#####_V_1_usage	A normalized version of the V-gene usage of the sample.
#####_pep	This file contains the most information of all of the downloadable data, including reference positions, gene calls, the full and stitched read, as well as copy numbers. This file can be used for the re-creation of alignments, and due to the stitched read, is one of the most useful for downstream analysis purposes.

Table 2: A very general description of the contents downloaded from Raw Data, a series of files which should allow users to perform advanced analyses, as well as replicate some of the charts on iRweb.

Raw Data: #####_pep.csv

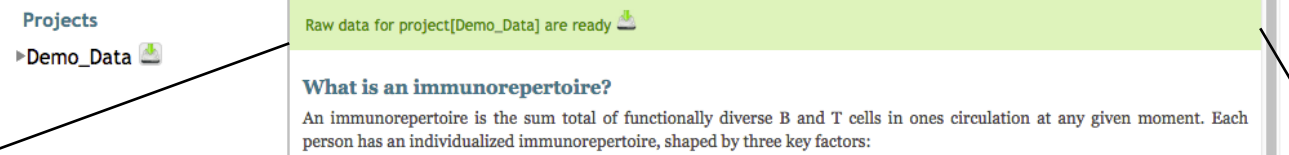
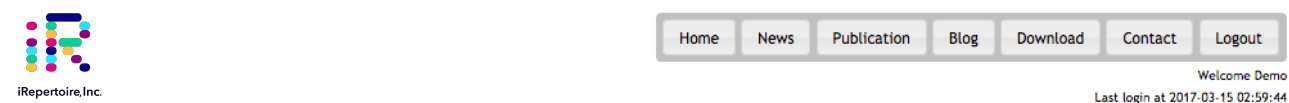
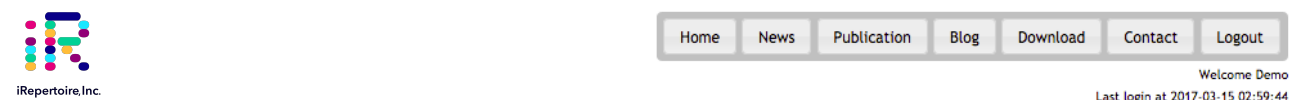
The most valuable tool, in terms of building on top of the analysis available from iRepertoire, is the [pep.csv](#) file available under the **Raw Data** download for each sample. Below is an overview of the information available within the pep.csv for any given sample, per column.

Column Header	Description
CDR3 (pep)	The CDR3 peptide sequence; Any * represents a STOP codon.
V	The V-gene the sequence aligns to
VRefBegin	Position of the beginning of the sequence alignment, in relation to the reference sequence.
VRefEnd	Position of the end of the sequence alignment, in relation to the reference sequence.
VReadBegin	Where the V-gene alignment begins within the read from sample data
VReadEnd	Where the V-gene alignment ends within the read from sample data
D	The D-gene the sequence aligns to, if uncalled -0.
DRefBegin	Position of the beginning of the sequence alignment, in relation to the reference sequence.
DRefEnd	Position of the end of the sequence alignment, in relation to the reference sequence.
DReadBegin	Where the D-gene alignment begins within the read from sample data, if uncalled -0.
DReadEnd	Where the D-gene alignment ends within the read from sample data, if uncalled -0.
J	The J-gene the sequence aligns to
JRefBegin	Position of the beginning of the sequence alignment, in relation to the reference sequence.
JRefEnd	Position of the end of the sequence alignment, in relation to the reference sequence.
JReadBegin	Where the J-gene alignment begins within the read from sample data
JReadEnd	Where the J-gene alignment ends within the read from sample data
C	The C-gene the sequence aligns to
CRefBegin	Position of the beginning of the sequence alignment, in relation to the reference sequence.
CRefEnd	Position of the end of the sequence alignment, in relation to the reference sequence.
CReadBegin	Where the C-gene alignment begins within the read from sample data
CReadEnd	Where the C-gene alignment ends within the read from sample data
joinedSeq	The full stitched read, post-filtering. All uppercase letters are 100% matches from the stitched and overlapped reads.
CDR3 (nuc)	The nucleotide sequenced of the CDR3, pulled from the joinedSeq
copy	During the analysis process, identical reads are collapsed and counted. Analysis is performed on the collapsed read to reduce processing time, and the number of copies of the stitched reads is kept as the copy number. This copy number is for the unique joinedSeq nucleotide sequence. This number may not be 1:1 with list CDR3 new or old values as these rely on peptide and V- & J-gene usage.



Raw Data: Whole Project Download

For much larger projects, it is now possible to download all of the Raw Data for the project in one large .zip file. Now, Projects will appear with a green download button beside their name on the left-hand side. Once clicked, this will begin the process of collecting all of the Raw Data files for all the samples in the study. A blue banner should appear on the right-hand side. Once the files have been collected and the .zip file is ready, the page will refresh and the banner will be green. The download icon in the green banner can be used to download the .zip. Please note that for much more extensive projects, this feature may take some time to work - with the page refreshing a couple times.



Frequently Asked Questions (F.A.Q.)

How do I get the raw sequencing data for my study?

Access to raw sequencing data is dependent upon the pooling strategy for your study. If an entire flow cell or lane was purchased with which to pool your study, it is possible for Customer Service to provide access to this information - as well as a lab report that details the molecular IDs used in your study for each sample. If an entire flow cell or lane was not purchased and your samples were pooled with R&D or other customers, it is possible for you to receive the raw, demultiplexed data. There are two forms of demultiplexed data: (1) The stitched reads, without quality data or (2) the demultiplexed R1 & R2 with quality data intact.

I do not see a gene I am interested in. Why is it not here?

iRepertoire's primer systems were developed to cover genes with designation 'Functional' on the IMGT (<http://www.imgt.org/>) Gene Tables. Please be sure to consult these tables for the appropriate species and chain to be able to assess which genes are or are not covered by iRepertoire's primers. Genes with a designation of ORF (Open Reading Frame) or P (Pseudogene) are not covered.

Is subclass information available?

Class information is available for all Ig samples. Subclass information and allelic information, though available in the Raw Data download, need to be grouped by Class only, not subclass for downstream analysis.

How do I get data formatted for IMGT/HighV-QUEST?

Please contact Customer Service to put in a request for your data to be in IMGT/HighV-QUEST format. Otherwise, if programming skills are available, it is possible to parse information from the "pep.csv" file from the Raw Data download of each sample, as the column **JoinedSeq** contains the entirety of the stitched read.

How do I get data to iRepertoire to analyze?

Please go online to our website (www.irepertoire.com) and complete a **Data Submission Form** (<https://irepertoire.com/data-submissions/>). If this form is unavailable in your country or otherwise inaccessible, please contact Customer Service about getting an Excel version of the form. Always complete forms, whether online or in Excel format, to the fullest extent possible, as this allows us to provide the quickest turnaround for the analysis of your data. Errors, typos, missing or incorrect file names and Study Names will delay return of your analyzed data. Data can be uploaded via SFTP (please contact Customer Service for a login), shared via Google Drive, DropBox, or even a physical HDD can be shipped to iRepertoire's offices (ATTN: Data Management). Please be sure, if you ship a HDD, to include return postage and/or packaging.



iRepertoire, Inc.

D-gene information doesn't seem to be available for my sample

For samples in which D-gene information is relevant, please look for it in a graphical format under the List CDR3 new tree map. Otherwise, this information is available in the pep.csv from Raw Download for the sample

What nomenclature system is used by iRepertoire?

iRepertoire uses the system listed on IMGT. Correspondence between nomenclatures can be found by searching www.imgt.org.

Contact Information

Please be sure to include your name, organization, and a detailed description of the request and/or error you are encountering. If requesting raw, demultiplexed reads, please provide a specific Project and Study Name including the date submitted.

Office Hours: Monday - Friday, 8:00 AM to 5:00 PM CST (Central Standard Time)

Telephone/Fax: 1 (256) 327-0948

Email: info@irepertoire.com or datasupport@irepertoire.com